A multi-channel speech enhancement framework for robust NMF-based speech recognition for speech-impaired users

Gert Dekkers[1,2,4], *Toon van Waterschoot*[1,2], *Bart Vanrumste*[1,2,4], *Bert Van Den Broeck*[1,2,4], *Jort F. Gemmeke*[3], *Hugo Van hamme*[3], *Peter Karsmakers*[1,2,4]

[1]ESAT-ETC/AdvISe, KU Leuven TC Geel, Kleinhoefstraat 4, 2440, Geel, Belgium
[2]ESAT-STADIUS, KU Leuven, Kasteelpark Arenberg, 3001, Leuven, Belgium
[3]ESAT-PSI, KU Leuven, Kasteelpark Arenberg, 3001, Leuven, Belgium
[4]iMinds, Medical IT, Kasteelpark Arenberg, 3001, Leuven, Belgium
gert.dekkers@kuleuven.be

Abstract

In this paper a multi-channel speech enhancement framework for distant speech acquisition in noisy and reverberant environments for Non-negative Matrix Factorization (NMF)-based Automatic Speech Recognition (ASR) is proposed. The system is evaluated for its use in an assistive vocal interface for physically impaired and speech-impaired users. The framework utilises the Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filter (SP-SDW-MWF) in combination with a postfilter to reduce noise and reverberation. Additionally, the estimation uncertainty of the speech enhancement framework is propagated through the Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction to allow for feature compensation in a later stage. Results indicate that a) using a trade-off parameter between noise reduction and speech distortion has a positive effect on the recognition performance with respect to the well-known GSC and MWF and b) the addition of a postfilter and the feature compensation increases performance with respect to several baselines for a non-pathological and pathological speaker.

Index Terms: multi-channel speech enhancement, speech recognition, uncertainty of estimation, dysarthric speech

1. Introduction

A Vocal User Interface can make a significant difference for people with a physical disability for whom controlling domestic devices would require a substantial amount of physical effort [1]. Current state-of-the-art Automatic Speech Recognition (ASR) is not sufficiently robust to dialectic or dysarthric speech, which is often encountered with disabled users. Therefore, this study is conducted in the context of a speaker-dependent recognizer that learns from user interactions [2]. Secondly, and most relevant for this work, is the fact that for most users it is not convenient or comfortable to wear a close-talk microphone, creating the need for a robust far-talk speech acquisition system. Speech signals recorded with a distant microphone contain noise and reverberation which degrade the performance of ASR systems. Whereas speech enhancement aims to improve speech quality or intelligibility, robust ASR aims to reduce the mismatch between the noisy and reverberant speech features and the trained acoustic model. Despite this fact it is reasonable to assume that speech enhancement algorithms are useful to achieve robust ASR.

Multi-channel algorithms may obtain a significant gain over single-channel algorithms since the former exploit spatial diversity. Regarding adaptive multi-channel speech enhancement algorithms we can distinguish adaptive beamforming and Multi-channel Wiener Filtering (MWF) [3]. A common implementation of adaptive beamforming is the Generalized Sidelobe Canceller (GSC) [4]. While the GSC relies on the assumption that the microphone signals are delayed versions of each other and needs an estimate of the angle of arrival, the MWF utilises no a-priori information on the signal model. In general, the goal of MWF is to estimate the desired speech in a Minimum Mean Squared Error (MMSE) sense using second-order statistics. In [3] the Speech Distortion Weighted-MWF (SDW-MWF) was introduced providing a trade-off between noise reduction and speech distortion. Here the SDW-MWF was integrated in the GSC obtaining the Spatially Pre-processed SDW-MWF (SP-SDW-MWF) providing more robustness against signal model errors. In case of diffuse noise fields, when both desired and undesired acoustic sources are in the same direction or due to signal model mismatch, the previous algorithms may not provide sufficient noise reduction. To reduce the residual noise and reverberation some researchers have proposed a scheme with multi-channel processing followed by a single-channel postfilter [5, 6].

In the last decade observation uncertainty techniques have been introduced in the context of robust ASR [7, 8, 9, 10]. In traditional feature compensation techniques the cleaned up speech feature vector is assumed to be a deterministic estimate. Observation uncertainty techniques describe each feature as a probabilistic density function to include the uncertainty. Many speech enhancement techniques however operate in the Short Time Fourier Transform (STFT) domain whereas the speech recognition features are, in our case, Mel-Frequency Cepstral Coefficients (MFCCs). In [9] a method was introduced to estimate the statistics in the STFT domain and propagate these through the feature extraction process to obtain a probabilistic MFCC feature description. An estimate of the uncertain variance was obtained empirically using the measure of change between input and output. In [11, 12] these were extended based on a Gaussian model of uncertainty.

The state-of-the-art speaker-dependent ASR for speech-impaired users introduced in [2] has not yet been evaluated in adverse acoustic environments. The main contributions of this paper are a) a robust distant speech enhancement framework based on the generic SP-SDW-MWF scheme with postfilter and observation uncertainty techniques, b) the evaluation of the pro-

posed framework for its use in the speaker-dependent ASR in adverse acoustic environments for both a pathological and a non-pathological speaker.

2. Proposed framework

This Section introduces the proposed framework which consists out of a multi-channel speech enhancement stage followed by a postfilter where instead of a deterministic estimate of the clean speech, a probabilistic representation is estimated which is incorporated into the ASR to allow for feature compensation.

2.1. Multi-channel speech enhancement

Consider a model where each acoustic path between a speaker and microphone m is represented as a Room Impulse Response (RIR) $h_m[n]$. Using this representation each microphone signal can be modelled as follows:

$$u_m[n] = s[n] * h_m[n] + v_m[n], m = 1 \dots M, \qquad (1)$$

with $*$ denoting the convolution operator, $v_m[n]$ the additive noise component and $u_m[n]$ the acquired input signal, at time index n. Our objective is to estimate the desired speech signal $s[n]$. The SP-SDW-MWF depicted in Fig. 1 is a general scheme that encompasses multiple multi-channel speech enhancement algorithms. It consists out of a Delay-and-Sum Beamformer (DSB), blocking matrix and an adaptive stage. The DSB is a fixed beamformer that, ideally, steers a beam in the direction of the desired speech source by time aligning and adding the M different microphone signals $u_m[n]$ $(m = 0, \dots, M-1)$ to obtain the so-called speech reference $y_0[n]$. The blocking matrix subtracts the different time-aligned microphone channels in a pair-wise manner to obtain $M-1$ noise references $y_m[n]$. The goal of the adaptive stage is to estimate the noisy component in the speech reference $y_0^v[n]$. Let K be the number of reference channels (for now, $K = M$) to the adaptive stage and L the filter length. Consider the L-dimensional vector $\mathbf{y}_i[n]$ defined as $[y_i[n] \quad y_i[n-1] \quad \dots \quad y_i[n-L+1]]^T$ and the KL-dimensional stacked vector $\mathbf{y}[n]$ defined as $[\mathbf{y}_{M-K}^T[n] \quad \mathbf{y}_{M-K+1}^T[n] \quad \dots \quad \mathbf{y}_{M-1}^T[n]]^T$. Let $\mathbf{y}[n]$ be the reference input to the adaptive stage and $\mathbf{w}[n]$ a vector of Finite Impulse Response (FIR) coefficients which is defined similarly as $\mathbf{y}[n]$. The speech reference signal $y_0[k]$ is delayed by Δ samples to obtain a non-causal filter. The estimated noise is subtracted from the speech reference to obtain the output signal $z[n]$:

$$z[n] = y_0^v[n-\Delta] - \mathbf{w}^T[n]\mathbf{y}[n], \qquad (2)$$

with $(.)^T$ denoting transpose operation. The SP-SDW-MWF aims to minimize a weighted sum, depending on parameter μ, of the residual noise energy $e^v[n]$ and speech distortion energy $e^s[n]$:

$$\begin{aligned} J(\mathbf{w}[n]) = E\{ \overbrace{\|y_0^v[n-\Delta] - \mathbf{w}^T[n]\mathbf{y}^v[n]\|^2}^{e^v[n]} \} \\ + \frac{1}{\mu} E\{ \underbrace{\|\mathbf{w}^T[n]\mathbf{y}^s[n]\|^2}_{e^s[n]} \}, \end{aligned} \qquad (3)$$

with $E(.)$ denoting the expectation parameter and upperscript s or v denoting the speech or noise component respectively. If $\mu > 1$ emphasis is put on noise reduction. When $\mu < 1$ emphasis is put on reducing speech distortion. In case $\mu = 1$ the MMSE estimate is obtained. The minimizer for the cost function is obtained in (4).

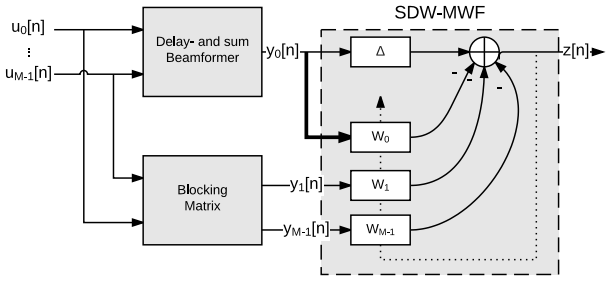

Figure 1: *SP-SDW-MWF scheme consisting of a DSB, blocking matrix and a SDW-MWF in the adaptive stage.*

$$\begin{aligned} \mathbf{w}[n] =& [E\{\mathbf{y}^v[n]\mathbf{y}^{v,T}[n]\} \\ &+ \frac{1}{\mu} E\{\mathbf{y}^s[n]\mathbf{y}^{s,T}[n]\}]^{-1} E\{\mathbf{y}^v[n] y_0^v[n-\Delta]\}. \end{aligned} \qquad (4)$$

Due to the assumption that speech and noise are uncorrelated $\mathbf{y}^s[n]\mathbf{y}^{s,T}[n]$ can be estimated by subtracting $\mathbf{y}^v[n]\mathbf{y}^{v,T}[n]$ from $\mathbf{y}[n]\mathbf{y}^T[n]$ which are estimated in periods of only noise, and noisy speech respectively. This introduces the need for a Voice Activity Detection (VAD). Depending on the usage of the spatial pre-processor, of inclusion of the FIR filter $\mathbf{w}_0$ and the choice of the trade-off parameter μ, several algorithms are obtained. Without $\mathbf{w}_0$ ($K = M-1$) and with $\mu = +\infty$ the algorithm behaves as a GSC. For any other values of μ speech distortion is regularized depending on μ and the amount of speech leakage. This algorithm is formally known as the Speech Distortion Regularized-GSC (SDR-GSC) [3]. Consider the L-dimensional vector $\mathbf{u}_i[n]$ defined as $[u_i[n] \quad u_i[n-1] \quad \dots \quad u_i[n-L+1]]^T$ and the ML-dimensional stacked vector $\mathbf{u}[n]$ defined as $[\mathbf{u}_0^T[n] \quad \mathbf{u}_1^T[n] \quad \dots \quad \mathbf{u}_{M-1}^T[n]]^T$. In case no spatial preprocessing is used, by changing $\mathbf{y}^{s,v}$ and y_0^v in Equation 4 to $\mathbf{u}^{s,v}$ and u_1^v respectively the SDW-MWF is formed. For $\mu = 1$ this resembles a MMSE estimate which is denoted as the MWF. In this paper an adaptive frequency-domain implementation (Recursive Least Squares-type) is used. More information regarding this implementation can be found in [13].

2.2. Single-channel postfilter

The single-channel postfilter is directly applied to the output $z[n]$ of the multi-channel speech enhancement scheme. $z[n]$ can be expressed in the STFT domain as:

$$\begin{aligned} Z_{l,k} &= S_{l,k} + D_{l,k} + R_{l,k} + N_{l,k} \\ &= X_{l,k} + R_{l,k} + N_{l,k}, \end{aligned} \qquad (5)$$

with k denoting the discrete frequency index and l the frame index. This signal is assumed to contain the desired speech signal $S_{l,k}$, residual early $R_{l,k}$ and late reverberation $D_{l,k}$ along with residual noise $N_{l,k}$. Due to the fact that $D_{l,k}$ does not have a negative effect on the recognition performance [14] the objective is estimating $X_{l,k}$ by applying a spectral gain factor $G_{l,k}$ as follows,

$$\hat{X}_{l,k} = G_{l,k} Z_{l,k}. \qquad (6)$$

In [15] a MMSE Log-Spectral Amplitude (LSA) estimator was proposed. This estimator uses the logarithm of the Short-Time Spectral Amplitude rather than STFT in its optimization criterion and is defined as:

$$\hat{A}_{l,k} = \arg\min_{\hat{A}'_{l,k}} E\{|\log A_{l,k} - \log \hat{A}'_{l,k}|^2\}, \qquad (7)$$

where $A_{l,k}$ is the amplitude of the clean Fourier coefficient after marginalizing out the phase in $X_{l,k}$. Assuming a Gaussian prior distribution on $A_{l,k}$, $R_{l,k}$ and $N_{l,k}$ with zero mean and variances $\lambda_{X_{l,k}}$, $\lambda_{R_{l,k}}$ and $\lambda_{N_{l,k}}$ respectively and a likelihood $p(Z_{l,k}|A_{l,k})$ with mean $A_{l,k}$ and variance $\lambda_{R_{l,k}} + \lambda_{N_{l,k}}$ the criterion in (7) corresponds to the expectation of the posterior distribution [15]:

$$\hat{A}_{l,k} = \exp(E\langle \log A_{l,k}|Z_{l,k}\rangle), \tag{8}$$

which leads to the following optimal spectral gain [15]:

$$G_{l,k} = q_{l,k} \exp\left(\frac{1}{2}\int_{q_{l,k}\gamma_{l,k}}^{\infty} \frac{\exp -t}{t} dt\right),\ q_{l,k} = \frac{\zeta_{l,k}}{1+\zeta_{l,k}}, \tag{9}$$

with $q_{l,k}$ the Wiener gain, $\zeta_{l,k}$ being the a-priori Signal-to-Noise Ratio (SNR) and $\gamma_{l,k}$ the a-posteriori SNR defined as:

$$\zeta_{l,k} = \frac{\lambda_{X_{l,k}}}{\lambda_{R_{l,k}} + \lambda_{N_{l,k}}},\ \gamma_{l,k} = \frac{|Z_{l,k}|^2}{\lambda_{R_{l,k}} + \lambda_{N_{l,k}}}. \tag{10}$$

Regarding equation 9 the reader is advised to read the literature in [15]. $\zeta_{l,k}$ is estimated using the Decision-Directed (DD) algorithm [15]. Both the DD algorithm and $\gamma_{l,k}$ need an estimate of $\lambda_{R_{l,k}}$ and $\lambda_{N_{l,k}}$. $\lambda_{N_{l,k}}$ is estimated using an algorithm called Improved Minima Controlled Recursive Averaging (IMCRA) [16] which estimates the speech presence probability to determine the recursive smoothing parameter for each frequency bin. In [17] it was shown that the late reverberant variance $\lambda_{R_{l,k}}$ can be estimated directly from the variance of the speech signal $\lambda_{X_{l,k}}$ based on a statistical RIR model:

$$\lambda_{R_{l,k}} = e^{-2\eta t_m} \lambda_{X_{l',k}}, \qquad l' = l - \frac{t_m}{t_{shift}}, \tag{11}$$

where t_{shift} resembles the amount of shift between consecutive frames, t_m the boundary between early and late reverberation (typically 50ms) and η the decay rate which is inversely proportional to the reverberation time T60.

2.3. Integration of front-end and back-end

Training and recognition are both based on Non-negative Matrix Factorization (NMF) which relies on modelling utterances as a linear combination of acoustic units [2]. Unlike other ASR, the NMF-based ASR can deal with weak supervision. It is designed to learn from a spoken command accompanied by an action on some device's user interface (e.g. "Turn on the light" when pressing a certain button). The NMF-based ASR needs a set of posterior probabilities $p(c_i|\mathbf{o}_l)$ where a codebook represents multiple codewords c_i and a MFCC feature vector $\mathbf{o}_l$ is given for a particular frame l. In the original setup a deterministic estimate of the desired speech was used to extract frame-based MFCC features. To obtain a codebook, the available training data is clustered in an unsupervised manner, using a k-means iterative process [2], to represent the data as a set of codewords c_i each containing a Gaussian. Given the codebook and a MFCC feature vector $\mathbf{o}_l$, a vector containing a posterior probability $p(c_i|\mathbf{o}_l)$ for each codeword is obtained using soft Vector Quantization (soft VQ):

$$p(c_i|\mathbf{o}_l) = \frac{p(\mathbf{o}_l|c_i)p(c_i)}{p(\mathbf{o}_l)} \approx \frac{p(\mathbf{o}_l|c_i)}{\sum_i p(\mathbf{o}_l|c_i)}. \tag{12}$$

More information regarding the training and recognition process can be found in [2].

Instead of using a deterministic estimate of $X^l(k)$, combining this with an uncertainty of the estimate provides more information for the ASR. The integration between front-end and back-end basically consists of 3 steps: 1) estimation of posterior distribution $p(X_{l,k}|Z_{l,k})$ for each frame l and frequency index k, 2) propagation of the posterior distribution through the MFCC feature extraction and 3) usage of this uncertainty for codebook-based feature compensation. To obtain an estimate of the uncertainty of observation in the STFT domain a complex Gaussian model of uncertainty was used [12]:

$$p(X_{l,k}|Z_{l,k}) = \mathcal{N}_{\mathbb{C}}(\hat{X}_{l,k}, \lambda_{l,k}). \tag{13}$$

However, for the posterior distribution of Equation 8 no closed-form solution is available. It was shown that the uncertainty $\lambda_{l,k}$ of the MMSE-LSA can be approximated by using the Bayesian MSE of the single-channel Wiener filter [12]:

$$\lambda_{l,k} \approx \frac{\zeta_{l,k}}{1+\zeta_{l,k}}(\lambda_{R_{l,k}} + \lambda_{N_{l,k}}). \tag{14}$$

Once the posterior distribution of the enhanced speech is obtained it is transformed through the MFCC feature extraction using the technique in [11, 12], known as Uncertainty Propagation (UP). In short, the different steps in the MFCC extraction process are treated separately containing both linear and non-linear transformations. Finally, each MFCC vector is characterized by a likelihood $p(\mathbf{o}_l|\mathbf{Z}_l)$ where $\mathbf{Z}_l$ represents the STFT vector of $z[n]$ for a particular frame l. Obtaining the posterior distribution $p(c_i|\mathbf{o}_l)$ in (12) is achieved by using Modified Imputation (MI) [12] which will allow for a codebook-based feature compensation of $\mathbf{o}_l$. The imputed value of the MFCC vector $\mathbf{o}_{i,l}^{MI}$, assuming it is generated by codeword i, with prior $p(\mathbf{o}_l'|c_l)$ taking the uncertainty model $p(\mathbf{o}_l'|\mathbf{Z}_l)$ into account is given as:

$$\mathbf{o}_{i,l}^{MI} = \arg\max_{\mathbf{o'}_l} \{p(\mathbf{o'}_l|\mathbf{Z}_l)p(\mathbf{o'}_l|c_l)\}. \tag{15}$$

3. Experimental setup

The results presented in this work have been obtained using the DOMOTICA-2 database acquired during the ALADIN project [18] which contains recordings of dysarthric and impaired speakers controlling a home automation system [2, 19]. For the evaluation, a pathological (PS) and non-pathological (NPS) speaker were selected. For these speakers, Speech Intelligibility (SI) scores were obtained using an automated tool [20], leading to a score of 64,2% and 93,4% respectively which are the lowest and highest SI in the DOMOTICA-2 corpus. Each utterance can be decomposed into one or more so-called slots values (e.g. "Kitchen light" and "on"). The amount of slots values determines the complexity of the classification problem. The corpus of the PS contained 18 slots values while the NPS has 21. In this experiment 4 different realizations of 28 different commands were used. For the purpose of evaluating these algorithms on robust distant speech recognition a simulation environment was defined. RIRs were simulated using the Image Source Method [21] in a room with dimensions 5x5x3m and a T60 time of 0.4s. A linear microphone array containing three microphones with an inter-microphone distance of 6.8 cm located at [0.01m;2.5m;2m] was used. The position of the desired source and noise were randomly chosen for each utterance to minimize position-related bias. Each command was convolved with a different RIR and each added to a different noise realization using stationary (white Gaussian) and non-stationary noise

	stationary noise				non-stationary noise				stationary noise				non-stationary noise			
SNR(dB)	-3	6	15	AI	-3	6	15	AI	-3	6	15	AI	-3	6	15	AI
Clean	98.9	98.9	98.9		98.9	98.9	98.9		93.6	93.6	93.6	0.0	93.6	93.6	93.6	
Unpre-processed	54.6	77.3	79.5	0.0	33.2	56.3	68.4	0,0	34.9	53.5	61.2	0.0	24.0	37.7	47.9	0.0
DSB	61.6	77.8	80.8	2.1	37.4	54.2	65.0	1.3	38.7	53.3	62.2	1.7	28.2	45.7	51.7	6.5
GSC	75.5	85.5	86.4	11.0	66.5	78.5	84.8	24.1	57.7	63.7	67.1	12.1	46.0	57.7	66.2	20.5
MWF	58.8	81.7	82.3	3.6	31.2	61.0	71.7	2.1	44.7	56.7	63.1	5.3	21.9	26.8	48.2	-1.0
SDR-GSC	75.7	86.3	88.4	11.3	63.4	81.5	85.0	24.7	56.8	64.7	68.8	13.0	46.3	61.5	68.5	22.5
SDW-MWF	71.9	84.4	84.4	8.4	62.6	80.9	86.1	24.7	52.3	58.0	61.2	6.3	40.0	51.1	59.9	14.1
SP-SDW-MWF	73.4	82.6	83.7	7.7	57.3	79.2	84.6	22.0	49.1	60.6	62.2	7.0	36.8	52.0	62.4	14.9
SDR-GSC-PF	81.0	92.0	92.7	17.1	67.8	77.7	88.3	25.1	60.2	73.9	75.5	21.6	**49.2**	60.3	70.7	23.8
SDR-GSC-PF-U	**84.6**	**95.0**	**93.0**	**19.9**	**68.4**	**81.5**	**89.6**	**27.6**	**65.9**	**77.0**	**78.0**	**24.7**	47.3	**63.7**	**74.5**	**25.0**

Table 1: *Results for the non-pathological speaker (left) and pathological speaker (right) for stationary and non-stationary noise in function of SNR (dB). Overall, the SDR-GSC-PF-U outperforms the alternatives with an AI ranging from 19.9 to 27.6%.*

(Chime corpus [22]). For evaluation purposes a five-fold cross-validation was performed. Commands were grouped in groups of nearly equal sizes by minimizing the Jensen-Shannon divergence between the slot value distributions of each block [2]. Due the fact that the ASR is speaker-dependent, in a practical situation, the ASR is trained based on pre-processed recorded data containing residual noise and reverberation. Each command for each fold in the trainingsset was duplicated and combined with two different SNR values (1.5dB and 10.5dB), convolved with a different RIR and added to a different noise realization. This data set was also pre-processed for each speech enhancement algorithm to obtain the training data set.

For the implementation of the SP-SDW-MWF scheme presented in [13] a filter length L of 400 samples was used at a sampling rate of 16kHz. For the (SP-)SDW-MWF a trade-off parameter μ of 100, and for the SDR-GSC a value $\mu = 1000$ was chosen which turned out to be good values based on the training data. Other parameters were set as described in [13]. The filter length of the single-channel postfilter was set to 400 samples with an overlap of 240 which is similar to the MFCC framing. For the boundary between early and late reverberation a value t_m of 50ms was used and the T60 was assumed to be known a priori. Regarding the estimation of the noise variance with IMCRA we refer to [16] for the extensive list of parameters. For the propagation of the uncertainty a full covariance was used [12].

4. Results

In Table 1 the results are shown for the experimental setup described in Section 3. For the entire corpus we have the golden standard description available, which will be used to evaluate the system based on the slot F-score $(\%)$ which considers both the slot precision and slot recall using a harmonic mean. The proposed framework is compared with an unpre-processed microphone signal, clean speech, DSB, GSC and MWF. Results for -3, 6 and 15 dB are shown along with the average improvement (AI) with respect to the unpre-processed baseline. The AI is based on SNR values of -3 to 15 dB in steps of 3dB. Compared to the clean baseline, results drop considerably for the unpre-processed microphone signal. Although the results for clean speech are close, the drop in performance for the PS with respect to a decreasing SNR is much higher. Several variations of the SP-SDW-MWF scheme are evaluated. The SDR-GSC outperforms the alternatives and therefore only this algorithm is shown with the addition of a postfilter (denoted PF) and usage of uncertainty for feature compensation (denoted as U). Due the fact that the ASR is trained on the enhanced noisy speech, speech distortion has a smaller impact on the recognition performance compared to clean speech training. However, it is not beneficial to put all emphasis on noise reduction. Although a small difference, the SDR-GSC ($\mu = 1000$) performs better than the GSC ($\mu = +\infty$). Also, the MWF is outperformed by the SDW-MWF where in this case more emphasis is put on noise reduction compared to the MWF. This shows that using a proper value for μ has a beneficial effect on recognition performance. As shown in [3] the SP-SDW-MWF outperforms the SDR-GSC in terms of SNR enhancement but introduces additional speech distortion which could explain the fact that the SDR-GSC outperforms the SP-SDW-MWF. The addition of a postfilter does increase the performance especially for the stationary noise type in case of low SNRs which is expected because of the lower noise variance estimation errors during speech presence compared to the non-stationary noise type. For the entire experiment it is clear that the SDR-GSC-PF-U outperforms the other alternatives with an AI 19.9-27.6% for the non-pathological speaker and 24.7-25.0% for the dysarthric speaker.

5. Conclusions

Several multi-channel speech enhancement algorithms based on GSC and MWF were compared for robust ASR in two controlled environments using data acquired from a pathological and non-pathological speaker. It was shown that the usage of the trade-off parameter between noise reduction and speech distortion has a positive effect on the recognition performance. Overall, the SDR-GSC performed significantly better than the alternative multi-channel speech enhancement algorithms. The additional postfilter has a positive effect on all algorithms especially in the case of stationary noise at low SNRs. Combining this with an estimate of the estimation uncertainty also further improved the performance. Overall, the SDR-GSC-PF-U outperformed the other algorithms for both speakers in both environments with an average improvement of 22.3%. Further research will focus on a) making the usage of the parameter μ more practical by adaptively estimating its optimal value and b) evaluating the framework for a larger set of speakers.

6. Acknowledgements

We thank IWT-SBO projects ALADIN [18] (100049), SINS [23] (130006) and IC1303 COST Action AAPELE.

7. References

[1] J. Noyes and C. Frankish, "Speech recognition technology for individuals with disabilities," *Augmentative and Alternative Communication*, vol. 8, no. 2, pp. 297–303, 1992.

[2] B. Ons, N. Tessema, J. van de Loo, J. F. Gemmeke, G. De Pauw, W. Daelemans, and H. Van hamme, "A self learning vocal interface for speech-impaired users," in *Proc. of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*. Grenoble, France: Association for Computational Linguistics, August 2013, pp. 73–81.

[3] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[4] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transaction on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, January 1982.

[5] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[6] B. Cauchi, I. Krodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc. 2014 REVERB Workshop*, Florence, Italy, May 2014.

[7] R. F. Astudillo, S. Braun, and E. A. P. Habets, "A multichannel feature compensation approach for robust ASR in noisy and reverberant environments," in *Proc. 2014 REVERB Workshop*, Florence, Italy, May 2014.

[8] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments," *Computer Speech Language*, vol. 27, no. 3, pp. 837–850, May 2013.

[9] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2005, pp. 82–85.

[10] T. T. Kristjansson and B. J. Frey, "Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition," University of Toronto, Tech. Rep., 2013.

[11] R. F. Astudillo, A. Abad, and J. P. da Silva Neto, "Integration of beamforming and automatic speech recognition through propagation of the Wiener posterior," in *Proc. ICASSP*, Kyoto, March 2012, pp. 4909–4912.

[12] R. F. Astudillo, "Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition," Ph.D. dissertation, Technical University Berlin, Germany, 2010.

[13] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.

[14] A. Sehr, E. A. P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 433–445, 1985.

[16] 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, 2003.

[17] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, pp. 770–773, 2009.

[18] ALADIN. Adaptation and Learning for Assistive Domestic vocal INterfaces. [Online]. Available: http://www.aladinspeech.be/

[19] J. F. Gemmeke, B. Ons, N. Tessema, H. Van hamme, J. van de Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the ALADIN project," in *INTERSPEECH*, Lyon, France, 2013, pp. 2039–2043.

[20] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, Belgium, 2012.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[22] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," 2010, pp. 1918–1921.

[23] SINS. Sound INterfacing through the Swarm. [Online]. Available: http://www.esat.kuleuven.be/sins/