3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)

Proceedings of Engineering & Technology (PET)
pp. 689-693
Copyright IPCO-2016

An instrumental evaluation framework for Post-Filter in multichannel Speech Enhancement systems

Adel Hidri[1], Hamid Amiri[2]

LR-SITI-ENIT, University Tunis El Manar
BP 37, LE BELVEDERE 1002 TUNIS
[1]hidri_adel@yahoo.fr
[2]hamid.amiri@gmail.com

Abstract— **Developing a multichannel speech enhancement algorithm, there are two main concerns that should be measured: speech distortion and residual noise, which mostly lead to a double-edged algorithm optimization. Hence, an algorithm performance assessment becomes an important task to have a better understanding of the behavior of an algorithm, and later on for its parameterization. This paper presents a new measurement test methodology based on the decomposition of the enhanced speech signal into its components speech and residual noise. Both of them give some impressions about speech distortion and residual noise, by computing the speech and noise components of an enhanced speech signal, respectively. Tow multichannel speech enhancement algorithms selected from the state-of-the-art are instrumentally evaluated. Tow instrumental measures are addressed in this paper: the perceptual evaluation of speech quality mean opinion score (PESQ-MOS) and the signal-to-noise ratio improvement (ΔSNR). These tow instrumental measures refine the algorithm performance evaluation in terms of speech distortion and residual noise.**

***Keywords*— Multichannel Speech enhancement, speech distortion, residual noise, objective evaluation, performance assessment.**

I. Introduction

In a noisy environment, speech quality may be significantly reduced dependent on the level of the background noise. Given a noisy speech signal, the aim of a speech enhancement algorithm in many application is to obtain an estimate of the desired clean speech signal with sufficiently strong noise attenuation, a naturally sounding residual noise and a low speech distortion. Hence, two quantities are usually measured to evaluate the performance of any speech enhancement algorithm, that is to say the amount of speech distortion and residual noise. The sturdiness of such method is that one achieves two separate signals: the filtered speech component and the filtered noise component, which represent the distorted talker's speech signal and the residual noise signal respectively. Focusing on speech enhancement such as speech distortion and noise attenuation can the comfortably be measured or additively assessed. However, this is a highly intrusive approach, which not only requires access to the input and the output signal of the algorithm, but also to the internal processing of the speech enhancement system. Firstly, we present a signal separation technique that allows for a detailed analysis of unknown speech enhancement systems. This method separates the speech and the residual noise of the speech enhancement system in the sending direction. This makes it possible to independently judge the speech distortion and the noise attenuation. While state of the art tests always try to judge the sending direction signal mixture, our technique allows a more reliable analysis in shorter time. Secondly and in order to instrumentally evaluate the performance of different post-filters for noise attenuation, an intrusive instrumental evaluation methodology will be introduced and applied. This framework provides us the possibility for evaluating the noise attenuation performance and the quality of the isolated speech component by reporting on the signal to noise ratio improvement (ΔSNR) and the perceptual evaluation of speech quality mean opinion score for the speech component (PESQ-MOS) respectively.

This paper is organized as follows: In the next section, we will present the proposed signal separation technique. In section 3, we will recapitulate some relevant Beamformer algorithms a long with the formulation of the baseline Zelinski post-filter and McCowan post-filter. In section 4, our choice of objective measures is discussed. Finally, section 5 presents our findings by a comparison of two exemplary speech enhancement systems using the above objective quality measures in the framework of the proposed test methodology.

II. Signal separation technique

The separation of the enhanced speech signal into the different signal components is generally accomplished through this process: While carrying out the enactment of the tested speech enhancement algorithm, we should make sure that all operational impact on the noisy speech signal should must be stocked apart and then implemented separately to both the noise signals and clean speech. The aforementioned structure is presented in Fig. 1. Both noise signals and the clean speech are independently liable to the resulting spectral weighting tenant to create the speech and noise constituents, correspondingly. The latter phenomenon is solely achieved in a frequency domain speech enhancement. Excepting the speech enhancement system's input and output signals, this method entails an open path to the system's internal processing.

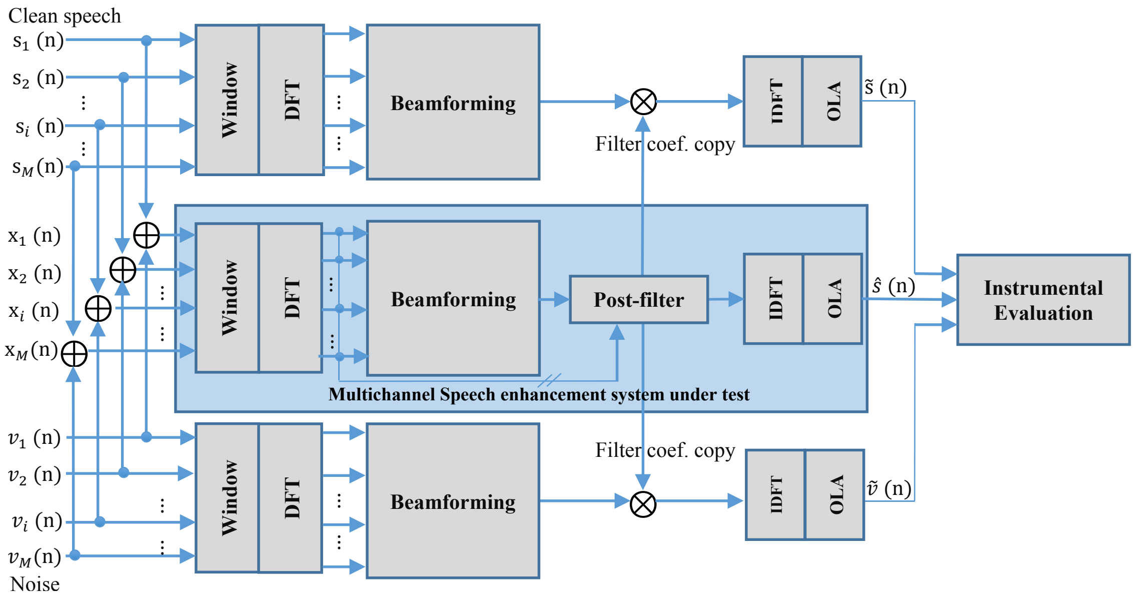

Fig. 1: Blok diagram of the framework for instrumental evaluation setup

This method is basically driven by the inconsistent availability of a certain system's internal processing in a practicable context. Actually, when it comes to the hands-free systems' hardware actualization, it frequently obstruct the access to the system's internal processing. As a matter of fact, any access preconditions to the system internal processing have become useless in this new method. The latter will be eventually approached through the input and output signals of the speech enhancement system. While presenting the internal processing as a pattern for the multiple-valued spectral weighting regulations. Those rules are finally predicted and later implemented in order to achieve the signal separation.

After having finishing the signal separation we can safely argue that the system performance might be assessed by both the noise constituents and the resulting speech. Given the fact that the clean speech signal is used as a unique regulator, then an alteration of the speech constituent to the clean speech signal shows a probable speech distortion of the system. The effectiveness of the system to constrict the noise signal is well shown through the resulting noise constituent. As long as the noise enfeeblement is more solid this will automatically lead to less remaining noise. Yet, this will be at the expense of greater amount of speech distortion, the opposite is valid too. Consequently, gauging both quantities coincidentally via instrumental measurement experiments and/or biased listening experiments will show efficient an algorithm can diminish these distortions and accomplish a balanced accommodation between the two camps. This will be constructive within the operation of algorithm parameterization and improvement.

III. Beamforming and Post-Filter

Post-filtering methods for multichannel speech enhancement algorithms have recently attracted an increased interest. It is well known that beamforming methods yield a significant improvement in speech quality. However, when the noise field is spatially incoherent or diffuse, the noise reduction is insufficient and additional post-filtering is normally required. In general, the post-filter is applied as a post processing at the output of the Beamforming in the preprocessing stage. The Beamforming technique used in this paper was presented in my previous work [1]. In this paper, we will instrumentally evaluate two post-filters, Zelinski post-filter and McCowan post-filter, with a new methodology. In the next, we briefly describe these two selected post-filter.

Consider a microphone array with M channels in a noisy environment. After applying the short time Fourier transform of length K, the vector of the outputs microphone signals can then be formulated with frame index l and frequency bin k as :

$$X(l,k) = S(l,k) + V(l,k) \qquad (1)$$

where $X(l,k) = \left(X_1(l,k)X_2(l,k) \dots X_M(l,k)\right)^T$ is the noisy signal, $V(l,k) = \left(V_1(l,k)V_2(l,k) \dots V_M(l,k)\right)^T$ is the additive noise and $S(l,k) = S_d(l,k)\,D(k)$ is the speech signal. The term $S_d(l,k)$ denotes the desired source signal and $(.)^T$ the vector transpose. $D(k)$is the propagation vector modeling the delays each channel for the desired source signal based on the reference microphone depending on the microphone array geometry: $D(k) = (e^{-\frac{j2\pi k\tau_1}{c}} \dots e^{-\frac{j2\pi k\tau_M}{c}})^T$ where c being the speech of the sound.

Following this signal model the multichannel signals $X(l,k)$will be preprocessed by the well-studied MVDR Beamformer:

$$W_{MVDR}(l,k) = \frac{\Phi_{VV}^{-1}(l,k)D(k)}{D^H(k)\Phi_{VV}^{-1}(l,k)D(k)} \qquad (2)$$

where $W_{MVDR}(l,k)$being the beamforming coefficients vector, $\Phi_{VV}(l,k)$ being the $M \times M$ normalized cross-power spectral density matrix of the noise and $(.)^H$denoting the Hermitian operator, respectively.

The single channel Beamformer output is then given by:

$$S_{BF}(l,k) = W_{MVDR}^{H}(l,k).X(l,k) \qquad (3)$$

Meanwhile employing a Beamformer alone is insufficient to substantially reducing the level of the noise, a post-filter has to be applied to provide further noise reduction. Hence, the post-filter is often utilized to improve the limited performance in terms of the noise attenuation. The output of the post-filter in the frequency domain is given by:

$$\hat{S}(l,k) = H_{PF}(l,k).S_{BF}(l,k) \qquad (4)$$

The coefficients of the post-filter are defined as:

$$W_{PF}(l,k) = \frac{\Phi_{SS}(l,k)}{\Phi_{SS}(l,k) + \Phi_{VV}(l,k)} \qquad (5)$$

where $\Phi_{SS}(l,k)$ and $\Phi_{VV}(l,k)$ being the clean speech signal and noise auto-power spectral densities after beamforming respectively.

With * and ** the output of the post-filter in the frequency domain is given by:

$$\hat{S}(l,k) = H_{PF}(l,k).W_{MVDR}^{H}(l,k).X(l,k) \qquad (6)$$

The most commonly used post-filter in multichannel speech enhancement structure is depicted in Fig.1.

The post-filter proposed by Zelinski [ ] is given by:

$$H_{ZE}(l,k) = \frac{\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\mathrm{Re}\{\hat{\Phi}_{X_iX_i}(l,k)\}}{\frac{1}{M}\sum_{i=1}^{M}\hat{\Phi}_{X_iX_i}(l,k)} \qquad (7)$$

and the post-filter proposed by McCowan [ ] is given by:

$$\hat{H}_{PF}(l,k) = \frac{\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\hat{\Phi}_{SS}^{(i,j)}(l,k)}{\frac{1}{M}\sum_{i=1}^{M}\hat{\Phi}_{X_iX_i}(l,k)} \qquad (8)$$

IV. Instrumental evaluation Methodology

In this paper, a new intrusive instrumental evaluation methodology is used to compare two post-filters performance in multichannel speech enhancement system. As shown in Fig.1, the clean speech signal $s_i(n)$and noise signal $v_i(n)$ with $i = 1 \ldots M$are the inputs to the multichannel speech enhancement system consisting of beamforming and post-filter. The post-filter coefficients will be computed based on the synthetically generated noisy signals $x_i(n) = s_i(n) + v_i(n)$ where $v_i(n)$ has been achieved by scaling the multichannel noise with a factor γ yielding predefined values of the input signal to noise ratio $\mathrm{SNR_{in}}$ measured by the active speech level tool according to ITU-T Recommendation P.56 [5]. The enhanced speech signal can be expressed by its components:

$$\hat{s}(n) = \tilde{s}(n) + \tilde{v}(n) \qquad (9)$$

By separate processing of the speech components $s_i(n)$and of the noise components, $v_i(n)$ we get the speech component of the output signal $\tilde{s}(n)$ and of the attenuated noise components $\tilde{v}(n)$ respectively. Using this intrusive evaluation approach, it is possible to calculate the output signal to noise ratio$\mathrm{SNR_{out}}$. Therefore, to evaluate the noise attenuation performance and the quality of the speech, the signal to noise ratio improvement ΔSNR , which is the difference between the $\mathrm{SNR_{out}}$ and $\mathrm{SNR_{in}}$, and the perceptual evaluation of speech quality mean opinion score (PESQ-MOS) are used. These two criteria quality measures can be calculated as:

$$\Delta SNR = \mathrm{SNR_{out}} - \mathrm{SNR_{in}} \qquad (10)$$

$$PESQ - MOS = f(\tilde{s}(n), s(n)) \qquad (11)$$

with reference signal s(n) being chosen as the best clean speech signal from $s_i(n)$ and the function $f(\tilde{s}(n), s(n))$is computed according to ITU-T Recommendation P.862.2 [6] Instrumentally estimating the wideband PESQ-MOS of $\tilde{s}(n)$ against $s(n)$.

V. Experiments

To demonstrate the application of the signal separation method, we evaluate two speech enhancement systems consisting of post-filter sub-systems. System A comprises Zelinski post-filter [3]. In contrast, system B consists of McCowan post-filter [4]. The performance of both speech enhancement systems are evaluated instrumentally under the "Multichannel In-Car Speech and Noise, Database" [7].

In the following, the experimental setup for validating the proposed framework are described in section A. The experimental results and discussion are presented in section B.

A. *Experimental setup*

In our experiments, the speech and noise signals are selected from the multichannel in-car speech and noise database [7].The applied microphone array in this work consists of four microphones with 3.6 cm distances located on the left side of the radio and navigation system display. Multiple recordings are made synchronously for each channel with clean speech signals being spoken from the driver position. Background noises have been separately recorded also in synchronous manner. In this work, two Background noise conditions are investigated. The first one is where car engine in resting state, window closed and air condition being set at 50%. The second condition is where the car driven on an expressway with a speed of 50 km/h, window closed and air condition being 50%. Each noisy signal is prepared for five$\mathrm{SNR_{in}}$ conditions of -5, 0, 5, 10 and 15dB. To performed noisy mixture, car noise signal at different level is added for each clean speech. The noisy mixtures were simulated recording separately the noise and speech. The noisy mixture scenario is shown in Fig.1. The sampling frequency $f_s = 16kh$ is used. The signal is windowed by a Hann window of length 512 samples (32 ms), followed by an FFT with length 512 and a frame shift of 50%.

B. *Experimental results*

As a first measure, the amount of speech distortion is evaluated by means of PESQ-based MOS [2] of the filtered speech component $\tilde{s}(n)$ relative to the speech signal $s(n)$. PESQ scores are averaged over all test signals. Secondly, the signal-to-noise ratio improvement (ΔSNR) is to be computed.

In Fig. 2 and 3 we present simulation results for the approaches proposed by Zelinski [3] and McCowan [4] with two different post-filters. Fig.2 shows the results for the first background noise condition. As expected, Zelinski's approach hardly provides any signal-to-noise-ratio improvement.

Employing the a priori noise field coherence, McCowan's approach shows a well-improved noise attenuation performance against Zelinski scheme.

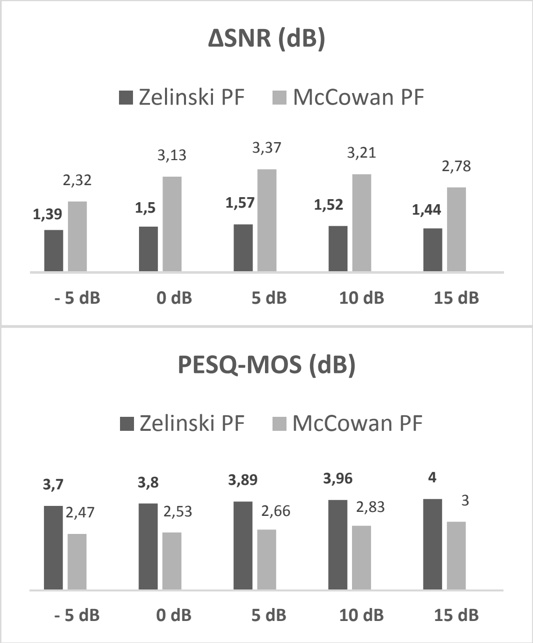

Fig. 2: ΔSNR and PESQ-MOS for the test condition with background noise from a car being idl (0km/h), window closed, 50% level air conditioning

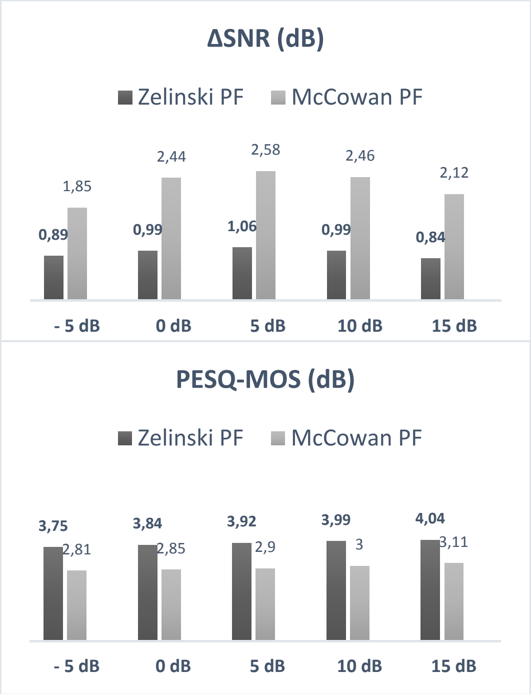

Fig. 3 : ΔSNR and PESQ-MOS for the test condition with background noise from a car driven with (50km/h), window closed, 50% level air conditioning

Although, Zelinski's approach delivers the best PESQ-MOS scores, its noise attenuation performance is too poor. Fig.3 shows the results for the second background noise condition. Similar conclusion can be drawn as in the first condition. However, in all approaches noise attenuation performance has decreased to some extent. This is due to the problem of the mismatch between the moving car noise field and the diffuse noise field model. Yet the PESQ-MOS score of all approaches have improved accordingly, which shows a trade-off noise attenuation performance and the preservation of the quality of the speech component.

VI. Conclusions

An instrumental evaluation framework for post-filter in multichannel speech enhancement systems has been presented. In addition to the noisy speech signal with the clean speech signal and the noise signal, the addressed instrumental evaluation framework can separate the enhanced speech signal into the filtered speech signal and the filtered noise signal. This framework provides us the possibility for evaluating the noise attenuation performance and the quality of the isolated speech component by reporting on the signal to noise ratio improvement (ΔSNR) and the perceptual evaluation of speech quality mean opinion score for the speech component (PESQ-MOS) respectively. This methodology offers researchers with a powerful means to measures the performance of algorithm simulation.

References

[1] A. Hidri and H. Amiri. "*A multichannel Beamforming-based Framework for Speech Extraction*". International Journal of Intelligent Engineering Informatics, Vol. 3, No 2/3, pp. 273-291, 2015.

[2] *Perceptual Evaluation of Speech Quality (PESQ)*, ITU-T P.862, Feb. 2001.

[3] R. Zelinski, "*A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms*". In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New York, NY, USA, Apr., pp. 2578–2581, 1988.

[4] I. McCowan, and H. Bourlard, "*Microphone Array Post-Filter Based on Noise Field Coherence*". IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 709–716, 2003.

[5] ITU-T Recommendation P.56, Objective Measurement of Active Speech Level, ITU-T 1993.

[6] ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, ITU-T 2005.

[7] H. Yu, "*Post-Filter Optimization for Multichannel Automotive Speech Enhancement*". PhD thesis, Technique University of Braunschweig. Germany, 2013.