In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension

Rachel E. Bouserhal, Tiago H. Falk, and Jérémie Voix

Citation: The Journal of the Acoustical Society of America **141**, 1321 (2017);
View online: https://doi.org/10.1121/1.4976051
View Table of Contents: http://asa.scitation.org/toc/jas/141/3
Published by the Acoustical Society of America

In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension

Rachel E. Bouserhal
Department of Mechanical Engineering, École de technologie supérieure, Montréal, Quebec H3C 1K3 Canada

Tiago H. Falk
Centre Énergie, Matériaux, Télécommunications, Institut National de la Recherche Scientifique, Montréal, Quebec H5A 1K6, Canada

Jérémie Voix[a)]
Department of Mechanical Engineering, École de technologie supérieure, Montréal, Quebec H3C 1K3 Canada

(Received 19 April 2016; revised 13 January 2017; accepted 14 January 2017; published online 1 March 2017)

Bone and tissue conducted speech has been used in noisy environments to provide a relatively high signal-to-noise ratio signal. However, the limited bandwidth of bone and tissue conducted speech degrades the quality of the speech signal. Moreover in very noisy conditions, bandwidth extension of the bone and tissue conducted speech becomes problematic. In this paper, speech generated from bone and tissue conduction captured using an in-ear microphone is enhanced using adaptive filtering and a non-linear bandwidth extension method. Objective and subjective tests are used to evaluate the performance of the proposed techniques. Both evaluations show a statistically significant quality enhancement of the noisy in-ear microphone speech with $\rho < 0.0001$ after denoising and $\rho < 0.01$ after bandwidth extension. © *2017 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4976051]

[JFL] Pages: 1321–1331

I. INTRODUCTION

Traditionally, communication headsets use a boom microphone, placed in front of the mouth, to capture speech in noisy settings. Although directional, these microphones often suffer from a low signal-to-noise ratio (SNR) in excessively noisy environments and require active noise control for enhancement (Gan *et al.*, 2005). Alternatively, speech captured through bone and tissue vibrations has been used to provide a signal with a higher SNR (Casali and Berger, 1996). Bone conduction speech can be captured either by microphones placed inside an occluded ear (Bou Serhal *et al.*, 2013; Kondo *et al.*, 2006) or through bone conduction sensors placed somewhere on the cranium (Zheng *et al.*, 2003). Although speech generated from bone and tissue conduction can have a relatively high SNR, it suffers from a limited bandwidth (less than 2 kHz), thus reducing signal quality and intelligibility (Turan and Erzin, 2013). For applications in which quality and intelligibility are important (e.g., command and control), bone and tissue conduction speech can be limiting factors. Therefore, to this day, communicating in noise is a difficult task to achieve as the communication signal either suffers from noise interference, in case of airborne speech, or from limited bandwidth, in case of bone conducted (BC) and tissue conducted speech.

Moreover, in excessively noisy environments where workers are exposed to noise levels greater than 90 dB(A) for 8 h, the Occupational Safety and Health Administration enforces the use of Hearing Protection Devices (HPDs) (OSHA, 1983). When worn correctly, HPDs can be very effective in preventing noise induced hearing loss (Berger, 2003). However, limited communication remains the number one complaint of workers equipped with HPDs (Murphy *et al.*, 2005).

Communication headsets are a great way of combining good hearing protection and communication. Most commonly, headsets made up of circumaural HPDs equipped with a directional boom microphone placed in front of the mouth are used. Circumaural HPDs can generally provide better attenuation than intra-aural HPDs, because they are easier to wear properly (Berger, 2003). The disadvantage of these types of communication headsets is twofold. First, the boom microphone is exposed to the background noise and can still capture unwanted noise that can mask the speech signal. Second, circumaural HPDs with boom microphones are not compatible with most other personal protection equipment. The use of other personal protection equipment alongside HPDs is common in noisy environments. For example, the use of helmets is required for construction workers as are gas masks for fire-fighters. Using bone and tissue conduction microphones to capture speech is a convenient way to eliminate both of those problems. Bone conduction sensors can be placed in various locations and can provide a relatively high SNR speech signal (McBride *et al.*, 2011). As mentioned previously, however, the elevated SNR comes at a price of very limited bandwidth, typically less than 2 kHz (Shin *et al.*, 2012). As a consequence, the enhancement of bone and tissue conducted speech is a topic of great interest. Many different techniques have been developed for the bandwidth extension (BWE) of BC speech (Turan and Erzin, 2013; Li *et al.*, 2014; Dekens and Verhelst, 2013; Rahman and Shimamura, 2011). Even

[a)]Electronic mail: jeremie.voix@etsmtl.ca

though theses techniques can enhance the quality of bone and tissue conducted speech, they are either computationally complex or require a substantial amount of training from the user (Shin *et al.*, 2012), thus limiting their widespread use in practical settings.

An effective compromise between the two extremes of noisy air conducted speech and bandlimited BC speech captured by bone conduction sensors is speech captured from inside an occluded ear using an in-ear microphone (IEM). Occluding the ear canal with an HPD causes BC vibrations originating from speech to resonate inside the ear canal leading the wearer to hear an amplified version of their voice; this is called the occlusion effect (Brummund *et al.*, 2014). By way of the occlusion effect, as a consequence of wearing an earplug, a speech signal is available inside the ear and can be captured using an IEM. Therefore, occluding the ear canal with a good acoustic seal via an earplug equipped with an IEM allows for the capturing of a speech signal that is not greatly affected by the background noise because of the passive attenuation of the earplug. Another advantage of using an IEM instead of a bone conduction microphone is that the speech is still captured acoustically and can share a significant amount of information with clean speech captured in front of the mouth in the 0 to 2 kHz range (Bouserhal *et al.*, 2015). However, in extremely noisy situations, some residual noise can exist inside the ear and capturing speech through air-conduction can result in a reduced SNR. Additionally, the speech captured inside the ear depends on resonance from bone and tissue and thus also suffers from a limited bandwidth. Because of the shared mutual information between the IEM speech signal and the air-conducted speech signal captured in front of the mouth, extending the bandwidth of IEM speech, in quiet conditions, is possible. A BWE technique that utilizes non-linear characteristics should extend the bandwidth of the IEM signal and add the high frequency harmonics (Iser and Schmidt, 2008).

In noisy conditions, however, extending the bandwidth of the bandlimited IEM speech becomes a difficult task because depending on the spectrum of the noise, simple BWE techniques may actually amplify the noise in the signal and decrease the SNR. Bandwidth extension techniques for noisy speech are rare and are typically computationally complex (Li *et al.*, 2014; Seltzer *et al.*, 2005). Since the SNR of the IEM speech is relatively high, denoising the speech signal becomes an easier task if the noise information inside the ear is known. In such extremely noisy conditions that the IEM signal becomes noisy, speech captured through air-conduction outside the ear has a very low SNR and is almost completely masked by the noise. Here, we propose the use of a microphone placed outside of the ear, on the outside of the earplug, such that the relationship between the sound outside the ear and inside the ear (i.e., the transfer function of the earplug) is known. This provides insight about the "in-ear" noise and enables denoising through adaptive filtering. Once the IEM speech signal is denoised, BWE can then be performed to further improve quality. Using combined techniques as such requires little training from the user and is computationally simple. Experimental objective and subjective results of an off-line simulation show that the proposed solution significantly improves the quality of the IEM speech. Increases of 44 points on the 100-point MUSHRA (MUlti Stimulus Test with Hidden Reference and Anchor) scale and 1.2 points on the 4.5-point POLQA (Perceptual Objective Listening Quality Assessment) scale were observed.

The remainder of this paper is organized as follows. Section II describes the methods and material used to perform and evaluate the proposed enhancement technique. The results are presented in Sec. III followed by a discussion and conclusion in Secs. IV and V, respectively.

II. METHODS AND MATERIALS

A. Speech corpus

We propose the use of the Auditory Research Platform (ARP) shown in Fig. 1, as a communication headset. The ARP uses an intra-aural custom molded earpiece for passive attenuation of ambient noise. Within the earpiece there is an IEM and a miniature loudspeaker. Located flush on the outer face of the earpiece is an Outer-Ear Microphone (OEM). It is of interest to see if in noisy conditions, with a configuration such as that of the ARP, a communication signal similar to that captured in front of the mouth in quiet conditions can be reached. Therefore, a speech corpus was recorded in an audiometric booth with the ARP as well as with a digital audio recorder (Zoom® H4n) placed in front of the speaker's mouth, later referred to as REF signal. A female speaker read out the first ten lists, totaling 100 sentences, of the Harvard phonetically balanced sentences (Rothauser *et al.*, 1969) and speech was recorded at an 8 kHz sampling rate and 16-bit resolution across the three microphones, simultaneously. The recordings were made with an 8 kHz sampling rate to stay true to realistic conditions with radio communications. A noisy speech corpus was then created from the clean corpus. Noise was injected to the OEM signals post recording to avoid any uncontrolled deviations in the speech between different recordings. To remain as close as possible

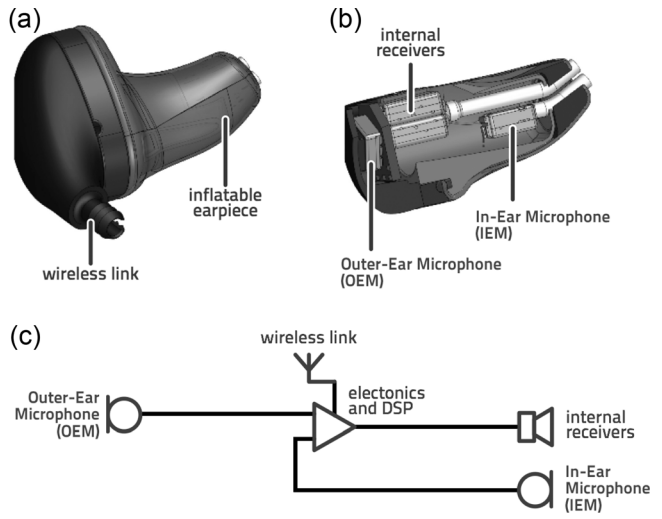

FIG. 1. ARP (a), its electroacoustic components (b), and equivalent schematic (c). Placed inside the ear the ARP captures speech produced by a talker using either the IEM or the OEM and transmits communication to other users through a wireless link.

to realistic conditions, the noise inside the ear, the IEM noise, was simulated using the OEM noise and the transfer function of the earplug. This was achieved by playing white noise over loudspeakers in the audiometric booth while the speaker was still equipped with the ARP (Nadon *et al.*, 2015). The noise signals collected by the IEM and OEM were then used to calculate the transfer function between the two microphones, i.e., the transfer function of the earpiece. Factory noise from the NOISEX-92 database (Varga and Steeneken, 1993) was then added to the OEM signal at an SNR of −5 dB. The noise was then filtered using the previously calculated transfer function of the earpiece and added to the IEM speech signal using MATLAB™ (MathWorks, Natick, MA). The REF signal was kept clean in order to provide an upper bound on the achievable performance. An SNR of −5 dB was chosen to simulate a typical industrial factory workplace setting. At this level, the signal captured by the OEM contains inaudible speech information, as it is buried in the noise (Bouserhal *et al.*, 2015).

B. Predicted quality

As shown in previous work, clear spectral differences between the IEM, OEM, and REF captured speech can be observed (Bouserhal *et al.*, 2015; Bou Serhal *et al.*, 2013). The IEM signal has a boost in the low frequency range but has a high frequency roll-off at about 1.8 kHz. The OEM and REF signals share the same bandwidth but have slight shifts in the formants (Bou Serhal *et al.*, 2013). These formant shifts are minimal and should not affect the quality of the clean OEM signal. To illustrate these spectral differences once more, the linear predictive coding (LPC) spectral envelope of the phoneme /i/ recorded with the REF, OEM, and IEM simultaneously is presented in Fig. 2.

1. Predicted quality in quiet conditions

Considering the shared mutual information between the OEM and REF signals (Bouserhal *et al.*, 2015) as well as their spectral differences (see Fig. 2), it is expected that the OEM speech signal is perceptually very similar to that of the REF speech in quiet conditions. Moreover, the "boomy" effect of the IEM, its limited bandwidth, and reduced shared mutual information with the REF signal should reduce its perceptual quality considerably compared to both the OEM and REF signals. To validate these predictions, as an objective quality measure, the International Telecommunication Union ITU-T Standard P.863, Perceptual Objective Listening Quality Assessment, POLQA (ITU-T 2011) with the REF signals as the reference was calculated for both the OEM and IEM signals. POLQA is the current recommendation for benchmarking and is used to evaluate speech for new and upcoming networks. The results of these measurements are shown in Sec. III.

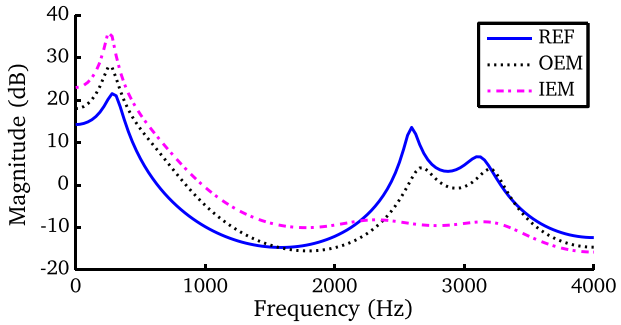

FIG. 2. (Color online) The LPC spectral envelope of the phoneme /i/ recorded with the REF, the OEM, and the IEM simultaneously.

2. Predicted quality in noisy conditions

In noisy conditions, because of the passive attenuation of the earplug, the quality of the IEM signal should not be greatly degraded by the presence of noise. The quality of the signal picked up by the exposed OEM, however, should be substantially reduced as the OEM speech is masked by the high level of noise. This prediction is supported by the maintenance of the amount of mutual information between the IEM and the REF signals in noisy conditions and the significant degradation observed between the OEM and the REF in noisy conditions (Bouserhal *et al.*, 2015). To validate the predicted changes in quality, POLQA was calculated for the noisy condition (SNR = −5 dB) of the IEM and the OEM speech as compared to the clean REF speech. The results of these measurements are shown in Sec. III.

C. IEM noise reduction

1. NLMS filtering

Once the noise level is high enough that the OEM speech is almost completely masked (SNR < −5 dB), the IEM speech signal can be denoised using normalized least mean squared (NLMS) adaptive filtering (Sayed, 2003). The choice of adaptive filtering comes from a need to create an algorithm that assumes no properties about the noise and is, thus, robust to various types of noise. Therefore, using adaptive filtering is beneficial for the user by enhancing the received communication signal.

To properly denoise the IEM speech signal produced by the user without affecting the speech content, the adaptation process must be frozen (OFF) when the user is speaking and active (ON) when the user is not speaking. This ensures that the adaptive filter cancels only the noise and does not interfere with any speech produced by the user. The two states of the adaptive filter are shown in Fig. 3. When the adaptation is ON the structure of the proposed adaptive filter follows the well-known structure commonly described in the literature (e.g., Manolakis *et al.*, 2005); the only exception being that the signal of interest is the error signal $e(n)$. Here, $H(z)$ is the transfer function of the earplug expressed as the ratio between the output signal captured by the IEM over the input signal captured by the OEM. $\hat{H}(z)$ is the estimated earplug transfer function. The method of estimating $\hat{H}(z)$ is further explained in Sec. II C 2. When the adaptation is ON, the user is not speaking. The OEM captures the noise outside the ear $n_o(n)$, while the IEM captures the residual noise inside the ear $n_r(n)$, colored by $H(z)$. The signal captured by the IEM is defined as the desired signal $d(n)$. The input $x(n)$ to the adaptive filter is the signal captured by the OEM filtered with the adaptive filter which is initialized by the estimated transfer

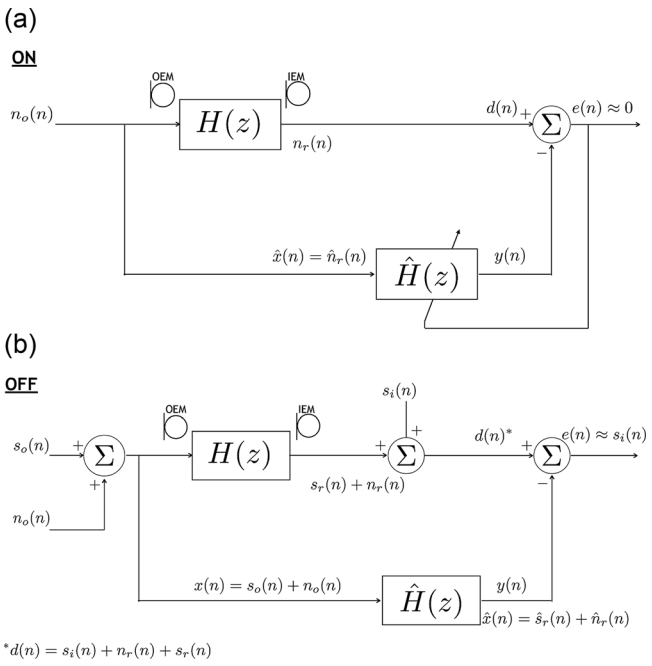

FIG. 3. Block diagram representing the NLMS filtering stage: (a) when the adaptation is ON and (b) when it is OFF.

function of the earplug $\hat{H}(z)$. The output of the adaptive filter $y(n)$ is thus a close estimate of the residual noise inside the ear and the difference between $d(n)$ and $y(n)$ should approach 0. A finite impulse response filter of order 160 is used. The order of 160 is chosen as it is the smallest order that can accurately reproduce the transfer function of the earplug with a delay of 20 ms, which is an undetectable delay for speech as shown by Lezzoum *et al.* (2016). The adaptive filter of order N is defined as follows:

$$\begin{aligned} y(n) &= \mathbf{w}^T(n-1)\mathbf{x}(n), \\ e(n) &= d(n) - y(n), \\ w(n) &= \mathbf{w}(n-1) + \frac{\mu e(n)\mathbf{x}(n)}{\epsilon + \mathbf{x}(n)^T\mathbf{x}(n)}, \end{aligned} \tag{1}$$

where n is the current time index, μ is the adaptation step size, $\mathbf{w}(n)$ is the $1 \times N$ vector of filter weights at time index n, $\mathbf{x}(n)$ is the $N \times 1$ vector frame of the OEM signal at the time index n, and ϵ is a very small number to avoid division by zero.

As presented in Fig. 3, when the adaptation is OFF, let $s_o(n)$ and $n_o(n)$ be the speech signal produced by the user and noise signal outside the ear, respectively. Therefore, the OEM picks up the sum of these two signals $x(n)$. Meanwhile, the IEM picks up the residual noise signal after the attenuation of the earplug $n_r(n)$ and the residual speech signal $s_r(n)$. The speech signal originating from bone and tissue conduction $s_i(n)$ is also picked up by the IEM. The sum of all the three signals picked up by the IEM is the desired signal $d(n)$. The signal $x(n)$ picked up by the OEM is then filtered using $\hat{H}(z)$ and the output $\hat{x}(n)$ is fed to the input of the NLMS adaptive filter. The output of the adaptive filter $y(n)$ is then subtracted from $d(n)$. The adaptive filter brings the difference between the residual noise $n_r(n)$ and the estimated residual noise $\hat{n}_r(n)$ to zero. Since the OEM speech is almost entirely masked by the noise, the effect of $s_r(n)$ and $\hat{s}_r(n)$ is negligible. Therefore, the resulting difference between the output of the adaptive filter and the signal captured by the IEM is the speech signal originating from bone and tissue conduction $s_i(n)$ with minimal effects of noise.

To properly denoise the IEM speech signal produced by the user without affecting the speech content, adaptation must only be performed when the user's speech is not present inside the ear. This ensures that the adaptive filter cancels only the noise and does not interfere with any speech produced by the user. Therefore, the adaptive filtering algorithm must also include a robust speech detection procedure that switches the adaptation process ON and OFF as a function of the speech inside the ear. This adaptation process, including the speech detection method, is described in Sec. II D.

2. Offline transfer function identification

First, the earplug transfer function must be estimated, as it varies from user to user. This is done in an offline identification stage. The field microphone-in-real-ear technique was used to obtain the transfer function of the earplug as it is not susceptible to disturbances caused by physiological noise as was shown by Voix and Laville (2009). As shown in Fig. 4, the ARP is worn and the user is exposed to white noise at 85 dB (SPL) using a loudspeaker outside the ear for at least 2 s. The OEM and IEM simultaneously capture the signals outside and inside the ear, respectively. After the OEM and IEM signals are collected the transfer function of the earplug $H(z)$ is estimated as $\hat{H}(z)$.

D. The adaptation process

To achieve denoising without affecting the speech content, the adaptation process is a function of whether or not the user is speaking. To denoise the user's speech, the adaptive filter must only adapt when the user is not speaking. This ensures that the filter is adapting to the earplug transfer function and thus the noise and only the noise is subtracted from the signal and not any relevant speech information. To

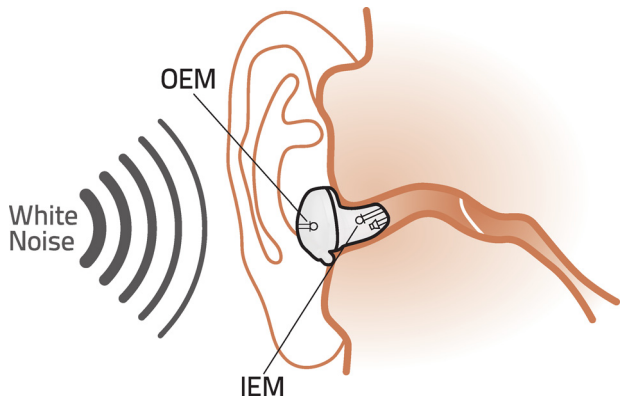

FIG. 4. (Color online) Offline identification stage of the earplug transfer function in the user's ear. White noise is played on a loudspeaker outside of the ear and recorded using both the IEM and the OEM. The transfer function of the earplug is calculated by assessing the noise outside the ear, recorded by the OEM, and the residual noise inside the ear, recorded by the IEM.

guarantee robustness of the speech detection process, voice activity detection inside the ear is achieved in the current project by monitoring the value of the coefficients of the adaptive filter. After completion of the 2 s identification stage, the vector of filter weights over the entire index of time w is used to detect if the user is speaking. To decide what criteria can be used to detect speech inside the ear using filter weights, test signals were developed using the first ten lists of the recorded Harvard phonetically balanced sentences discussed in Sec. II A, for both the OEM and the IEM. The test signals always started with at least 2 s of noise followed by 8 to 10 s of speech either by the user or by an external competing speaker. Exterior speech was added to simulate a case where the user is not speaking but an external speaker is loud enough that some residual speech exists after the passive attenuation of the plug. The residual speech should not trigger the speech activity of the adaptation process. For the IEM signal, the residual speech was simulated by passing the speech through $\hat{H}(z)$. The location of the user's speech and the residual speech was randomized to avoid any trends in the adaptation process.

Through analysis of the changes in the filter weights for the test signals recorded by the female speaker as described in Sec. II A, it was concluded that the maximum valued filter weight can be chosen as a good triggering criteria. Once the maximum filter weight increases more than a triggering threshold T_g from one time-index to the other, it is predicted that the user is speaking. Therefore once

$$\frac{\|w(n)\|_{\max}}{\|w(n-1)\|_{\max}} \geq T_g, \tag{2}$$

speech by the user is detected and the adaptation is turned OFF. The choice of value for T_g had to be done in a way that was not particular to one female speaker. Recorded conversation using the IEM and the OEM from four different speakers (two female, two male) was used to analyze the effect of using different triggering thresholds. Noise was inserted using the same procedure as discussed in Sec. II A. A sweep of the voice activity detection triggering threshold T_g from 1.01 to 1.2 was performed during the adaptation process. The upper limit of 1.2 was chosen empirically based on results up to 1.5 revealed a local maximum at $T_g = 1.06$. The bandwidth of the denoised signals for the four speakers resulting from the sweep was extended using the BWE process described in Sec. II E. The quality of these signals was measured before and after the BWE to see the effect of the different values for the triggering criteria. The choice of T_g was made as the triggering percentage value that produced the optimal objective quality over the four speakers as is shown in Sec. III.

The change in filter weights is triggered at the onset of speech but not the end. To ensure that the adaptive process starts back once speech inside the ear is no longer present, the overall change in energy Δ_ϵ at the onset of speech is also measured and monitored, per sample, i.e., $\Delta_\epsilon(n)$. Once triggered by the user's speech, the adaptation is disabled for at least 1 s and as long as Δ_ϵ is maintained. When the adaptation is OFF the filter weights of the adaptive filter are updated with those from the previous second $w(n - f_s)$, where f_s is the sampling rate. This is to ensure that the filter weights are those from when no speech is produced by the user. Once the change in energy is less than the onset change, $\Delta_\epsilon(n) < \Delta_\epsilon$, the adaptation starts again. The process of monitoring the change in Δ_ϵ gives a non *ad hoc* way to turn ON the adaptation once the user is no longer speaking. The adaptation process is demonstrated by the flow chart in Fig. 5.

E. IEM BWE

The adaptive filtering denoises the IEM signal by utilizing the information about the noise captured by the OEM. Once the IEM is denoised, its quality can be enhanced by extending its bandwidth in the high frequencies. Artificially extending the bandwidth of a clean bandlimited signal has been thoroughly studied (Jax and Vary, 2003; Kornagel, 2006; Bauer and Fingscheidt, 2008). Since the IEM signal shares mutual information with the REF signal between 0 and 2 kHz (Bouserhal *et al.*, 2015), it is only necessary to extend the bandwidth in the high frequency range, 2–4 kHz. As described by Iser and Schmidt (2008), a simple yet effective way of extending the bandwidth is through the application of the signal's nonlinear characteristics (Iser and Schmidt, 2008). A block diagram of the BWE process is shown in Fig. 6. First, the signal is up-sampled by a factor of 2 to provoke spectral folding. The excitation signal is extracted using a whitening filter and then cubed (Iser and Schmidt, 2008). To reach an excitation signal similar to that extracted from a wideband speech signal, the up-sampled signal is filtered by the whitening filter using the coefficients of an LPC analysis (Valin and Lefebvre, 2000). The whitening filter is a finite infinite response filter whose coefficients are those of an 18th order LPC filter at that time frame. Cubing the excitation reproduces the odd harmonics along the entire bandwidth including the high band, in this scenario from 1.8 to 4 kHz. Since the high frequencies are the only region of interest and to eliminate any overlap, the excitation signal is high passed at 1.8 kHz with a third order filter. Meanwhile, the up-sampled IEM signal is low passed at

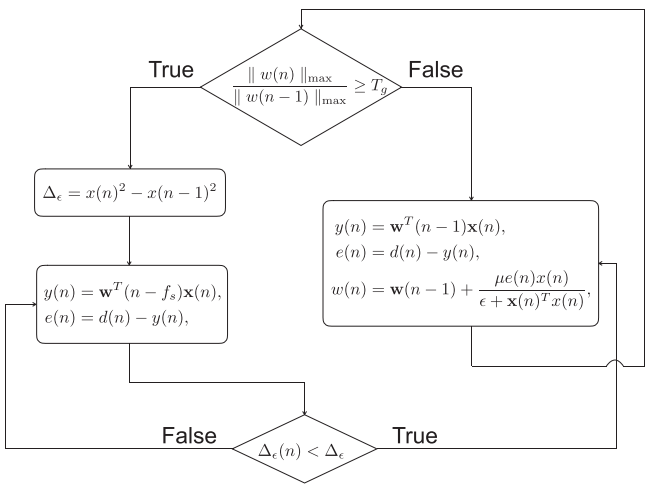

FIG. 5. Flow chart representing the adaptation process.

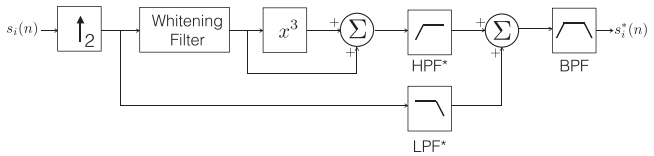

FIG. 6. Block diagram illustrating the BWE process. The HPF and the LPF are power complementary and designed for perfect reconstruction. The bandpass filter is a fourth order Linkwitz-Riley filter designed by cascading a second order low pass Butterworth filter and a second order high pass Butterworth filter.

1.8 kHz with a third order filter because it contains no relevant frequency information above 1.8 kHz (see Fig. 2). As described in Fig. 6, the high pass filter (HPF) used for the excitation signal and the low pass filter (LPF) used for the up-sampled IEM are designed to be power complementary for perfect reconstruction. The sum of the two filtered signals is then bandpassed with a fourth order Linkwitz-Riley filter at 160 Hz and 3.5 kHz by cascading a second order low pass Butterworth filter and a second order high pass Butterworth filter. This is done to eliminate the boomy effect coming from the bone and tissue conduction as well as any ringing caused by the odd harmonics of the cubed excitation signal. The overall output is then down-sampled by a factor of 2 to go back to an 8 kHz sampling frequency. It is important to note that this BWE technique adds missing harmonics in the high frequencies. However, missing formants and frication noise are not recovered.

F. Performance evaluation

The performance of the denoising and BWE processes were evaluated using both objective and subjective measures. Objectively, the quality of the signals was measured using POLQA. To confirm the results from the objective measures, the quality of the denoised and bandwidth extended speech was also measured subjectively. The MUSHRA (ITU-R, 2001) was used for this latter test. As part of the MUSHRA evaluation, four test signals were compared to the reference: the clean IEM signal, the noisy IEM signal, the denoised IEM signal, and the bandwidth extended denoised IEM signal. The reference signals chosen are those recorded in front of the mouth (REF). The noisy IEM served as the anchor, since it is a bandlimited noisy version of the reference signal. Ten randomly selected REF speech signals and their corresponding test signals were chosen. The test was performed online and participants were invited to take part of the test through an email that was approved by the internal review board at École de technologie suprieure. A description of the nature of the test, as well as detailed instructions, were described in the email invitation. Instructions emphasized that a comfortable volume should be set at the beginning of the test and should not be changed throughout. No assumptions were made on the participants' hearing abilities and no repeatability checks on participant performance were performed. A total of 44 participants took part of the test that should have taken less than 30 min. After review of the results, two subjects were rejected from the pool as it appeared they did not understand the nature of the test (low scores for hidden reference). To measure the statistical significance of improvements in the quality both objectively and subjectively, the Analysis of Variance (ANOVA) was tested on the gathered data.

III. RESULTS

A. Pre-enhancement objective quality assessment

1. Quiet condition

The POLQA MOS-LQO (mean opinion score-listening quality objective) results comparing the IEM and OEM signals to the REF signals are shown in Fig. 7. It can be seen that the quality of the OEM signal in quiet is high and in some cases was measured to have the maximum POLQA score of 4.5, thus indistinguishable from the REF signal. The inferior quality of the IEM can also be seen. The great variability in the POLQA scores for the IEM signals could be attributed to the fact that POLQA was not designed to measure the quality of speech originating from bone and tissue conduction.

2. Noisy condition

To show the decrease in the POLQA MOS-LQO results between the clean and the noisy condition (SNR $= -5$ dB), the MOS-LQO of the noisy IEM and the noisy OEM is shown in Fig. 8. Again the large variability in the POLQA scores for the IEM signals could be due to the fact that POLQA was not intended for use with bone and tissue conducted speech. Descriptive statistics are used to evaluate the degradation caused by noise. The cumulative distribution of the difference between the clean and the noisy POLQA MOS-LQO scores of the IEM signals as well as the difference between the clean and the noisy POLQA MOS-LQO scores of the OEM signals are shown in Fig. 9.

It can be seen that the decrease in the OEM quality is much greater than the decrease in the IEM quality. As a consequence of noise, half of the OEM sentences were degraded

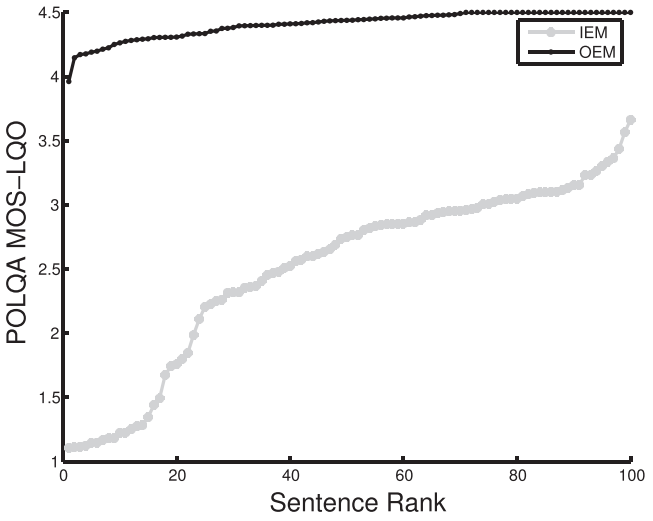

FIG. 7. POLQA MOS-LQO results of clean IEM and OEM signals using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores.

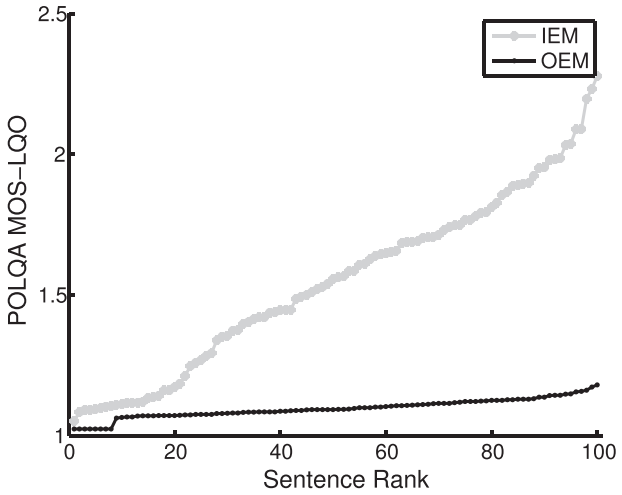

FIG. 8. POLQA MOS-LQO results of noisy IEM and OEM using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores.

by at least 3.39 points on the 5-point POLQA MOS-LQO scale, while half of the IEM sentences were at most degraded by 1.01 points. This confirms that the passive attenuation of the earplug prevents major degradations from noise on the IEM speech. Therefore, in noise, the IEM signals have superior quality relative to the OEM signals. The completely degraded signal captured by the OEM in noisy conditions can be utilized to denoise the relatively superior quality speech signal captured by the IEM, as described in Sec. II C.

B. IEM speech enhancement

1. Adaptive process triggering threshold

The adaptation must be ON only when the user is not speaking. This way the IEM speech is denoised without affecting the speech content. If the adaptation is ON even after the onset of speech by the user, the IEM speech content is affected by the denoising because the filter coefficients are not adapted only to $H(z)$. It is therefore important to have optimal voice activity detection criteria for the adaptive filtering process. The choice of triggering threshold, T_g, was chosen based on the results of one female speaker but validated for four other subjects. A sweep from $T_g = 1.01$ to $T_g = 1.2$ with a step size of 0.01 was done on four different speakers, as described in Sec. II C. The average POLQA MOS-LQO scores with only noise reduction and with BWE are shown in Fig. 10. The results show a clear peak around 1.06–1.07, suggesting that triggering threshold of $T_g = 1.06$ to detect speech activity inside the ear is best and can be extended to more than just one speaker.

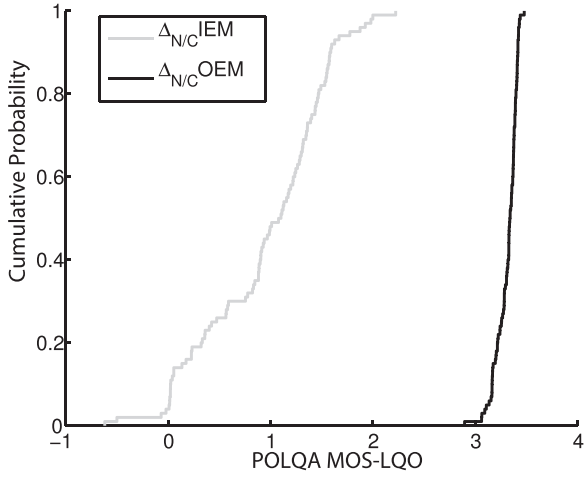

FIG. 9. Cumulative distribution plot of the difference in POLQA MOS-LQO results between the clean and noisy IEM and OEM signals.

2. IEM noise reduction

With a triggering threshold chosen as $T_g = 1.06$, the denoising and BWE techniques described in Sec. II C were performed on the test speech signals. Since the effect of μ and ϵ showed no major changes in performance within a specific window ($0.4 \leq \mu \leq 0.8$ and $0.0001 \leq \epsilon \leq 0.01$), for the denoising phase $\mu = 0.7$ and $\epsilon = 0.001$ were chosen empirically. To show the performance of the denoising using adaptive filtering, a randomly selected denoised IEM signal, IEM NS, is plotted against its corresponding IEM N in Fig. 11. As can be seen, the adaptive filtering process denoises the entire signal, when only noise is present (a), when the user is speaking (b), and when external speech is present (c). The adaptation process stops adapting once the user is speaking and relevant IEM speech content is preserved.

3. Bandwidth extension

Artificial BWE is then applied to the denoised signals. To show the regeneration of the odd harmonics and to compare the spectral content, the spectrograms of the REF signal, the IEM N, the IEM NS, and the bandwidth extended IEM signal are shown in Fig. 12. The noise reduction can be seen, as well as the "noise-like" effects of the BWE. Overall, however, the missing mid and high frequency harmonics lost in the IEM signal are regenerated after the BWE.

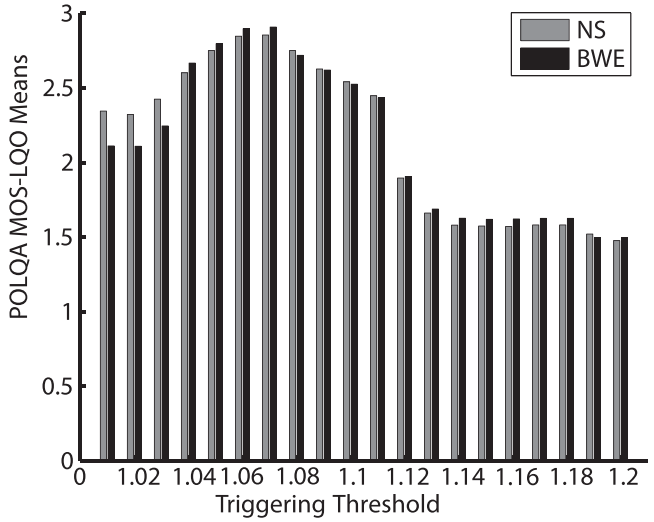

FIG. 10. Average POLQA MOS-LQO scores after denoising (NS) and after BWE over different triggering percentages, showing a peak at $T_g = 1.06 - 1.07$.

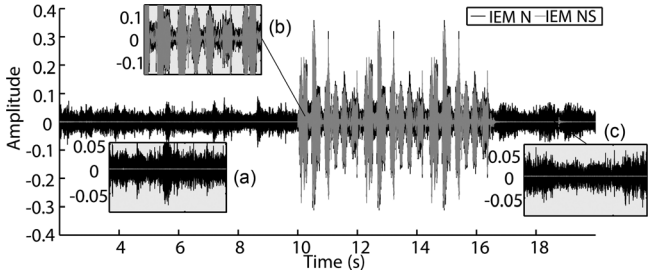

FIG. 11. The IEM NS as compared to IEM N. Zoomed portions of the denoised signal when only noise is present (a), when speech inside the ear is present (b) and when external speech is present (c).

C. Performance evaluation

1. Objective evaluation

To compare POLQA MOS-LQO results, the cumulative distributions of the difference in POLQA MOS-LQO scores between the IEM NS, the IEM N, and the clean IEM signal (IEM C) are plotted in Fig. 13. The same comparison made with the bandwidth extended signals is plotted in Fig. 14. This is done to show if, objectively, the BWE does increase perceived quality. The results show that the denoising enhanced at least half the sentences by 1 point on the POLQA MOS-LQO scale and that at least half the sentences measured no differently than the clean signals after denoising. Objectively, the results comparing the effects of BWE show that BWE increases the quality from the noisy signal by 1.2 points for at least half of the sentences. However, very small improvements of 0.02 and 0.14 points caused by BWE between the denoised signals and the clean signals, respectively, for at least half the sentences is seen. Therefore, results point that both the denoising as well as the BWE enhance the quality of the IEM noisy signal. The mean POLQA scores and p-values from one-dimensional ANOVA tests are shown in Tables I and II, respectively. All results were verified to be normally distributed before the ANOVA tests. Results show a statistically significant enhancement from the noisy signal caused by the denoising and the BWE techniques. There is also a significant improvement from the BWE technique and the IEM C, thus signaling the importance of the higher frequency components for quality perception.

2. Subjective evaluation

The results of the MUSHRA test averaged over 42 participants are shown in Fig. 15. The mean MUSHRA scores and p-values from one-dimensional ANOVA tests are shown in Tables III and IV, respectively. A statistically significant increase in quality can be seen as a consequence of the denoising and the BWE. Objectively and subjectively, there is no statistical significance between the quality of the clean IEM signals and the denoised IEM signals, thus suggesting indistinguishable differences. To confirm this, the log-spectral distance (LSD) between the de-noised IEM and the clean IEM signals was measured. The LSD is defined as follows (Falk *et al.*, 2010):

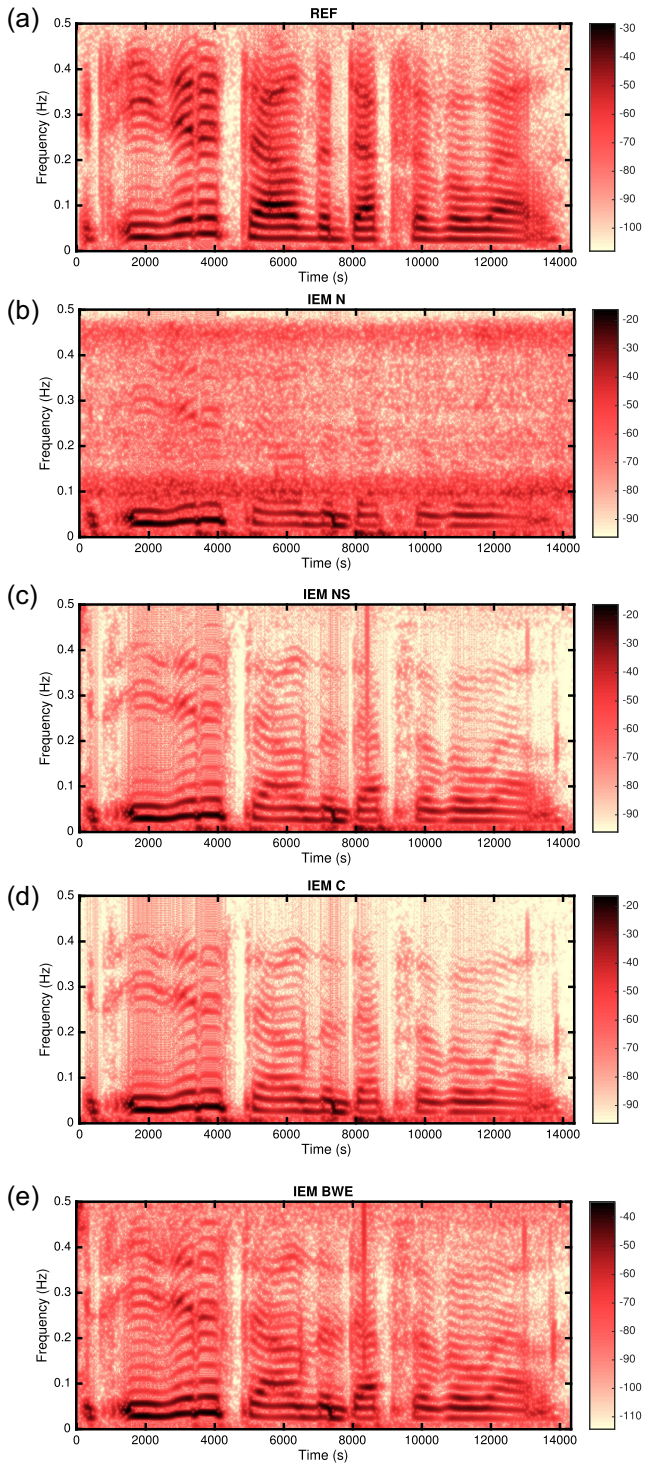

FIG. 12. (Color online) The spectrograms in dB full scale of the sentence "It is easy to tell the depth of a well" of the clean reference signal (a), the IEM N signal (b), the IEM NS signal (c), the original clean IEM (d) for comparison, and the bandwidth extended denoised IEM signal (e).

$$\mathrm{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{s_C(w)}{s_{NS}^{*}(w)}\right]^2 dw}, \tag{3}$$

where s_C and s_{NS} are the clean and denoised IEM speech signals, respectively. As shown in Fig. 16, all LSD values were under 1 dB. The LSD was calculated for 25 ms long,

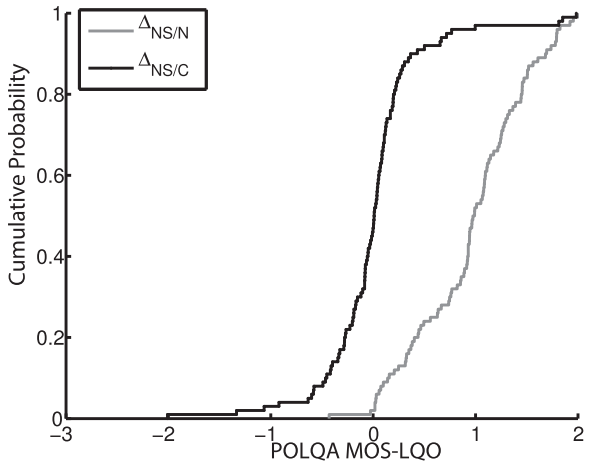

FIG. 13. Cumulative distribution of the difference in POLQA MOS-LQO scores between the denoised and noisy IEM ($\Delta_{NS/N}$), as well as the denoised and clean IEM ($\Delta_{NS/C}$).

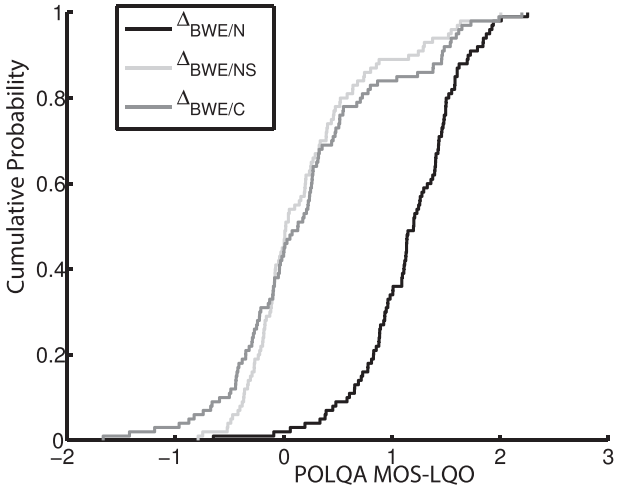

FIG. 14. Cumulative distribution of the difference in POLQA MOS-LQO scores between the bandwidth extended and noisy IEM ($\Delta_{BWE/N}$), the bandwidth extended and denoised IEM ($\Delta_{BWE/NS}$), and the bandwidth extended and clean IEM ($\Delta_{BWE/C}$).

TABLE I. Average POLQA MOS-LQO scores for the IEM N, the IEM NS, the bandwidth extended IEM, and the IEM C.

Signal	Mean POLQA
IEM N	1.559
IEM BWE	2.790
IEM NS	2.655
IEM C	2.757

TABLE II. Statistical significance results based on a 95% confidence interval between the objective evaluation of different stages of enhancement; IEM N signal, the IEM NS, the bandwidth extended IEM, and the IEM C signal.

Signals	p-value	Significant?
N vs NS	$p < 0.0001$	Yes
N vs BWE	$p < 0.0001$	Yes
NS vs BWE	$p < 0.01$	Yes
C vs BWE	$p < 0.01$	Yes
C vs NS	$p = 0.9413$	No

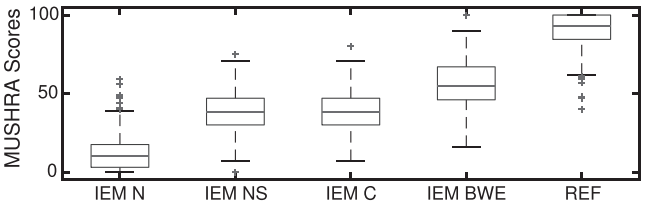

FIG. 15. Box and whisker plot comparing the MUSHRA results of the IEM N, the IEM NS, the IEM C, the bandwidth extended denoised IEM signal and the hidden reference. Table IV shows the statistical significance between relevant signals.

hamming windowed, frames with a 15 ms overlap. In speech coding, two signals with LSD <1 dB are considered to be perceptually indistinguishable (Paliwal and Kleijn, 1995). Therefore subjective results from the MUSHRA listening test confirm the objective trends found using POLQA.

IV. DISCUSSION

Experimental results with POLQA showed a statistically significant improvement in the speech quality between the noisy IEM speech and the enhanced (denoised and bandwidth extended) speech. Looking only at POLQA scores, however, it is not as apparent that the BWE enhances the quality much more than the denoising does. This could be because extending the bandwidth can introduce noise-like features in the high frequencies that could be misinterpreted by the objective measure as noise. The subjective results support this hypothesis. From the MUSHRA results it is evident that the denoised bandwidth extended signal is perceived to have significantly better quality than the denoised IEM without BWE. Extending the bandwidth after denoising results in even better perceived quality than the clean IEM signal. The mean and p-values between the MUSHRA scores of the clean IEM and the denoised IEM show that there was no statistical significance between the two. This suggests that the perceived quality between the denoised IEM and the clean IEM is undistinguishable. In fact, in about 30% of the time, participants gave the clean IEM speech and the denoised IEM speech identical MUSHRA scores.

The proposed approach is speaker independent, computationally simple, robust to noise, and requires no speech training by the user. Utilizing the review of conventional BC speech done by Shin *et al.* (2012), a comparison with the proposed solution is shown in Table V.

A possible limitation of this work is that it is done with speech data from only one female speaker. However, in

TABLE III. Average MUSHRA scores for the IEM N signal, the IEM NS signal, the IEM C signal, the bandwidth extended IEM, and the hidden reference (REF).

Signal	Mean MUSHRA
IEM N	10
IEM NS	38
IEM C	38
IEM BWE	55
REF	93

TABLE IV. Statistical significance results based on a 95% confidence interval between the subjective evaluation of different stages of enhancement; IEM N signal, the IEM NS signal, the bandwidth extended IEM, and the IEM C signal.

Signals	p-value	Significant?
N vs NS	$p < 0.0001$	Yes
N vs BWE	$p < 0.0001$	Yes
NS vs BWE	$p < 0.0001$	Yes
C vs BWE	$p < 0.0001$	Yes
C vs NS	$p = 0.9782$	No

close observation the only criteria that is potentially speaker dependent is the choice of triggering threshold T_g for voice activity detection during the adaptation process. The proposed concept of adaptive filtering for noise reduction has been shown to work in the past and is not speaker dependent (Davis 2002; Martinek and Zidek, 2010). Once denoising is achieved, the BWE process has also been proven to work regardless of the speaker (Iser and Schmidt, 2008). Therefore, the only speaker dependent factor that may arise is the value of the speech detection triggering criteria used in the adaptation process. Since the choice of T_g was made based on results from one speaker but validated on four speakers, it is assumed that this threshold can be extended for use with N number of speakers and should not greatly affect the enhancement process. In this work, the main priority was to evaluate and enhance the quality of the IEM speech. In future work, it is also relevant to measure and evaluate the intelligibility of the IEM speech and how the proposed enhancement process affects it.

V. CONCLUSIONS

Using bone and tissue conducted speech in noisy environments is a reliable way of providing a high SNR speech signal to the listener. The downfall usually lies in the limited bandwidth of the bone and tissue conducted speech. This paper focuses on the enhancement of speech generated from bone and tissue conduction picked up using a communication device equipped with an IEM and an OEM. An adaptive filtering approach is used to denoise the IEM signal using the OEM. Novel voice activity detection criteria using the filter coefficients of the adaptive filter is used to ensure that only noise is reduced while the speech content remains unaffected. Once the signal is denoised the bandwidth of the signal is extended by exploiting the nonlinear characteristics of a cubic operator. Both objective and subjective evaluations show that the BWE of the denoised IEM signal significantly enhances its quality. For factory noise, the techniques shown in this paper provide a simple, speaker independent, non-computationally exhaustive method to enhance the quality of speech picked up using an IEM. Overall, gains of 1.23 (out of 4.5) in POLQA MOS-LQO scores and 45 (out of 100) in MUSHRA scores show the benefits of the proposed speech enhancement solution.

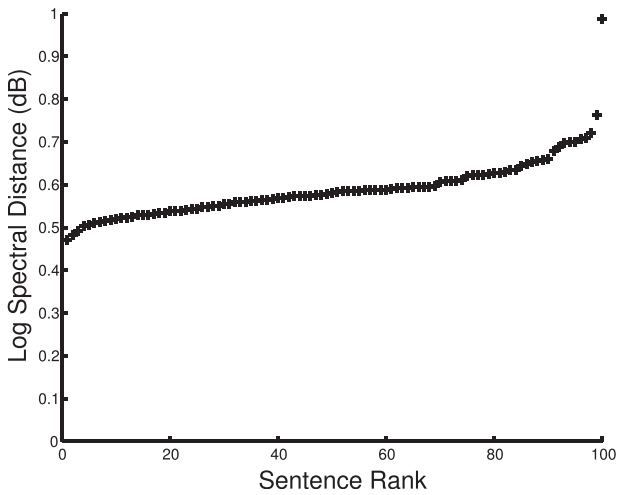

FIG. 16. Log-spectral distance between the IEM C signals and the IEM NS signals, with sentences sorted in ascending order.

TABLE V. Comparison of conventional BC enhancement approaches to the proposed approach.

Approach	Requires training?	Complex?
Equalization (Kondo *et al.*, 2006)	Yes	No
Analysis-and-Synthesis (Vu *et al.*, 2008; Thang *et al.*, 2006)	No	Yes
Probabilistic (Liu *et al.*, 2004)	No	Yes
Proposed	No	No

ACKNOWLEDGMENTS

This work was made possible via funding from the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the NSERC-EERS Industrial Research Chair in In-Ear Technologies. The authors would also like to thank João Felipe Santos for his help with the online subjective evaluation.

Bauer, P., and Fingscheidt, T. (**2008**). "An hmm-based artificial bandwidth extension evaluated by cross-language training and test," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Las Vegas, NV, pp. 4589–4592.

Berger, E. H. (**2003**). *The Noise Manual* (American Industrial Hygiene Assoc., Falls Church, VA), 796 p.

Bou Serhal, R. E., Falk, T. H., and Voix, J. (**2013**). "Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones," Proc. Meet. Acoust. **19**, 040013.

Bouserhal, R. E., Falk, T. H., and Voix, J. (**2015**). "On the potential for artificial bandwidth extension of bone and tissue conducted speech: A mutual information study," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brisbane, Australia, pp. 5108–5112.

Brummund, M. K., Sgard, F., Petit, Y., and Laville, F. (**2014**). "Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effecta)," J. Acoust. Soc. Am. **135**(3), 1433–1444.

Casali, J. G., and Berger, E. H. (**1996**). "Technology advancements in hearing protection circa 1995: Active noise reduction, frequency/amplitude-sensitivity, and uniform attenuation," Am. Indust. Hygiene Assoc. **57**(2), 175–185.

Davis, G. M. (**2002**). *Noise Reduction in Speech Applications*, Vol. 7 (CRC Press, Boca Raton, FL).

Dekens, T., and Verhelst, W. (**2013**). "Body conducted speech enhancement by equalization and signal fusion," IEEE Trans. Audio, Speech, Lang. Process. **21**(12), 2481–2492.
Falk, T. H., Sejdic, E., Chau, T., and Chan, W.-Y. (**2010**). *Spectro-Temporal Analysis of Auscultatory Sounds* (In-Tech Publishing, Croatia), Chap. 5, pp. 93–104.
Gan, W. S., Mitra, S., and Kuo, S. M. (**2005**). "Adaptive feedback active noise control headset: Implementation, evaluation and its extensions," IEEE Trans. Consumer Electron. **51**(3), 975–982.
Iser, B., and Schmidt, G. (**2008**). "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments* (Springer, Heidelberg, Germany), pp. 135–184.
ITU-R, Rec. (**2001**). BS 1534-1, *Method for the Subjective Assessment of Intermediate Sound Quality* (*MUSHRA*) (International Telecommunication Union Geneva, Switzerland).
ITU-T, Rec. (**2011**). P 863, *Perceptual Objective Listening Quality Assessment* (*POLQA*) (International Telecommunication Union, Geneva, Switzerland).
Jax, P., and Vary, P. (**2003**). "On artificial bandwidth extension of telephone speech," Signal Process. **83**(8), 1707–1719.
Kondo, K., Fujita, T., and Nakagawa, K. (**2006**). "On equalization of bone conducted speech for improved speech quality," in *IEEE International Symposium on Signal Processing and Information Technology*, Vancouver, British Columbia, Canada, pp. 426–431.
Kornagel, U. (**2006**). "Techniques for artificial bandwidth extension of telephone speech," Signal Process. **86**(6), 1296–1306.
Lezzoum, N., Gagnon, G., and Voix, J. (**2016**). "Echo threshold between passive and electro-acoustic transmission paths in digital hearing protection devices," Int. J. Industr. Ergon. **53**, 372–379.
Li, M., Cohen, I., and Mousazadeh, S. (**2014**). "Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones," in *IEEE China Summit and International Conference on Signal and Information Processing* (*ChinaSIP*), Chengdu, China, pp. 1–5.
Liu, Z., Zhang, Z., Acero, A., Droppo, J., and Huang, X. (**2004**). "Direct filtering for air-and bone-conductive microphones," in *IEEE 6th Workshop on Multimedia Signal Processing*, Siena, Italy, pp. 363–366.
Manolakis, D. G., Ingle, V. K., and Kogon, S. M. (**2005**). *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*, Vol. 46 (Artech House, Norwood, MA), 816 pp.
Martinek, R., and Zidek, J. (**2010**). "Use of adaptive filtering for noise reduction in communications systems," in *International Conference on Applied Electronics (AE)* (Pilsen, Czech Republic), pp. 1–6.
McBride, M., Tran, P., Letowski, T., and Patrick, R. (**2011**). "The effect of bone conduction microphone locations on speech intelligibility and sound quality," Appl. Ergonom. **42**(3), 495–502.
Murphy, W., Davis, R., Byrne, D., and Franks, J. (**2005**). "Advanced hearing protector study: Conducted at General Motors metal fabricating division," Flint Metal Center, Flint, Michigan.
Nadon, V., Bockstael, A., Botteldooren, D., Lina, J.-M., and Voix, J. (**2015**). "Individual monitoring of hearing status: Development and validation of advanced techniques to measure otoacoustic emissions in suboptimal test conditions," Appl. Acoust. **89**, 78–87.
OSHA. (**1983**). Occupational Noise Exposure: Hearing Conservation Amendment, Final Rule. Occupational Safety and Health Administration, 29CFR191095 Fed Regist **48**(46), pp. 9738–9797.
Paliwal, K., and Kleijn, W. (**1995**). "Quantization of LPC parameters," in *Speech Coding and Synthesis* (Elsevier Science B.V., Amsterdam) pp. 433–466.
Rahman, M. S., and Shimamura, T. (**2011**). "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," in *IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Seoul, Korea (South), pp. 1–4.
Rothauser, E., Chapman, W., Guttman, N., Nordby, K., Silbiger, H., Urbanek, G., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.
Sayed, A. H. (**2003**). *Fundamentals of Adaptive Filtering* (John Wiley & Sons, Vancouver, British Columbia, Canada), 1118 pp.
Seltzer, M. L., Acero, A., and Droppo, J. (**2005**). "Robust bandwidth extension of noise-corrupted narrowband speech," in *INTERSPEECH*, Lisbon, Portugal, pp. 1509–1512.
Shin, H. S., Kang, H.-G., and Fingscheidt, T. (**2012**). "Survey of speech enhancement supported by a bone conduction microphone," in *Proceedings of Speech Communication; 10. ITG Symposium*, VDE, Berlin, Germany, pp. 1–4.
Thang, T. V., Kimura, K., Unoki, M., and Akagi, M. (**2006**). "A study on restoration of bone-conducted speech with mtf-based and lp-based models," J. Signal Process. **10**(6), 407–417.
Turan, M. T., and Erzin, E. (**2013**). "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, British Columbia, Canada, pp. 7049–7053.
Valin, J.-M., and Lefebvre, R. (**2000**). "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *IEEE Proceedings of the Workshop on Speech Coding*, Delavan, WI, pp. 130–132.
Varga, A., and Steeneken, H. J. (**1993**). "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**(3), 247–251.
Voix, J., and Laville, F. (**2009**). "The objective measurement of individual earplug field performance," J. Acoust. Soc. Am. **125**(6), 3722–3732.
Vu, T. tat, Unoki, M., and Akagi, M. (**2008**). "An LP-based blind model for restoring bone-conducted speech," in *Second International Conference on Communications and Electronics* (*ICCE*), Hanoi, Vietnam, pp. 212–217.
Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., and Huang, X. (**2003**). "Air-and bone-conductive integrated microphones for robust speech detection and enhancement," in *IEEE Workshop on Automatic Speech Recognition and Understanding. ASRU'03*, Saint Thomas, U.S. Virgin Islands, pp. 249–254.