KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Bernardi G., van Waterschoot T., Wouters J., Moonen M. (2017), **Subjective and Objective Sound-Quality Evaluation of Adaptive Feedback Cancellation Algorithms** *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, May 2018, pp. 1010-1024
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	http://ieeexplore.ieee.org/abstract/document/8295151/
Journal homepage	http://ieeexplore.ieee.org/xpl/aboutJournal.jsp?punumber=6570655
Author contact	giuliano.bernardi@esat.kuleuven.be *+ 32 (0)16 321797*
IR	Klik hier als u tekst wilt invoeren.

(article begins on next page)

Subjective and Objective Sound-Quality Evaluation of Adaptive Feedback Cancellation Algorithms

Giuliano Bernardi, *Student Member, IEEE,* Toon van Waterschoot, *Member, IEEE,* Jan Wouters, and Marc Moonen, *Fellow, IEEE*

***Abstract*—Objective measures are widely used for the perceptual sound-quality evaluation of audio signal processing algorithms. Nevertheless, the use of subjective-evaluation measures remains relevant, in particular when application-specific objective measures are lacking.**

In this paper, we present a perceptual sound-quality evaluation of different algorithms for adaptive feedback cancellation (AFC), with both speech and music signals. Three algorithms are compared: the block normalized least mean squares (BNLMS) algorithm, the prediction-error method (PEM)-based frequency-domain adaptive filter (PEM-FDAF), and the PEM-based frequency-domain Kalman filter (PEM-FDKF).

The subjective evaluation results for the tested algorithms suggest that there is a large difference in statistical significance, and a corresponding large effect size, between the PEM-FDKF and the other algorithms, when using speech signals. A smaller statistical significance, and a lower effect size, is reported when using music signals.

The subjective evaluation results are then compared to the results obtained with several objective measures. The correlation between subjective and objective scores shows that objective measures can be effectively used to predict the sound-quality degradation caused by acoustic feedback and AFC artifacts.

***Index Terms*—Adaptive feedback cancellation (AFC), acoustic feedback control, prediction-error method (PEM), frequency-domain adaptive filter (FDAF), Kalman filter**

I. Introduction

ACOUSTIC feedback is a common problem encountered in many acoustic systems, such as public address systems and hearing aids (HAs), when there exists an unwanted coupling between the loudspeaker and the microphone within the system. A system in a feedback configuration can give rise to an unstable behavior which, in acoustics terms, might translate into a spectrum of different phenomena affecting sound quality, such as signal distortion and howling [1]. This is why feedback cancellation is of considerable importance in such systems, especially in the case of HAs where small microphone-to-loudspeaker distances are involved. In the past decades, many different algorithms have been developed to cope with acoustic feedback. A common class of acoustic feedback control algorithms includes the algorithms evaluated in this paper usually referred to as adaptive feedback cancellation (AFC) algorithms. The general idea in an AFC algorithm is that of modelling the unknown true feedback path $F_{\mathrm{t}}(q,t)$, where the subscript $_{\mathrm{t}}$ stands for 'true', with an estimated adaptive filter $\hat{F}(q,t)$, where t and q^{-1} are the discrete time index and the delay operator, respectively, i. e. $q^{-k}u(t) = u(t-k)$. The loudspeaker signal $u(t)$ is then filtered by $\hat{F}(q,t)$ and the resulting output $\bar{y}[t|\hat{\mathbf{f}}(t)]$ is subtracted from the microphone signal $y(t)$, in order to remove the unwanted contribution of the feedback signal $x(t)$. The microphone signal $y(t)$ then approximates the source signal $v(t)$, modeled as the summation of the noise signal $n(t)$ and the anechoic signal $a(t)$ filtered with the room impulse response (RIR) $R(q,t)$, that is processed by the forward path $G(q,t)$ and provided to the patient through the loudspeaker, cf. Fig. 1. The main problem faced by AFC algorithms is caused by the correlation between the source signal $v(t)$ and the loudspeaker signal $u(t)$, leading to a biased estimate of the feedback path [2]. In order to overcome this problem, many variations of the AFC algorithm involving different decorrelation approaches have been proposed, e. g. [3], [4], [5], [6], [7], [8].

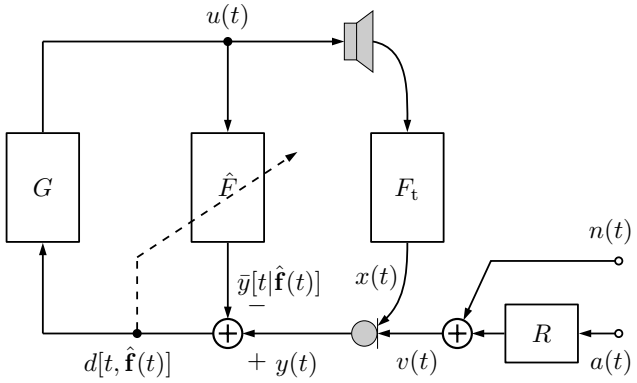


Fig. 1: General AFC scenario.

A widely studied approach to provide a suitable decorrelation between the source signal $v(t)$ and the loudspeaker signal $u(t)$ includes the application of the prediction-error method (PEM) [9]. The resulting algorithm, the PEM-based adaptive feedback canceller (PEM-AFC) [10], yields good convergence

This research work was carried out at the ESAT Laboratory and at the ExpORL Laboratory of KU Leuven, in the frame of the IWT O&O Project nr. 150432 'Advances in Auditory Implants: Signal Processing and Clinical Aspects', the KU Leuven Internal Funds C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', and the KU Leuven Impulsfonds IMP/14/037. The scientific responsibility is assumed by its authors.

G. Bernardi and M. Moonen are with the Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, B-3001 Leuven, Belgium (e-mail: giuliano.bernardi@esat.kuleuven.be; marc.moonen@esat.kuleuven.be).

T. van Waterschoot is with the Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, B-3001 Leuven, Belgium, and the Department of Electrical Engineering, ESAT-ETC, C-Media Research Lab, B-3000 Leuven, Belgium (e-mail: toon.vanwaterschoot@esat.kuleuven.be).

J. Wouters is with the Department of Neurosciences, ExpORL, KU Leuven, B-3000 Leuven, Belgium (e-mail: jan.wouters@med.kuleuven.be).

performance and limited perceptual distortions, when compared to algorithms employing other decorrelation approaches [1], [11]. Various modifications of the PEM-AFC have been developed by changing some of the underlying assumptions and/or implementation strategies; e. g., among others, the PEM-based adaptive filtering with row operations (PEM-AFROW) [12] makes the algorithm suitable for long feedback paths, the PEM-based frequency-domain adaptive filter (PEM-FDAF) [11], [13] offers a frequency domain implementation of the feedback path adaptation, and the PEM-based frequency-domain Kalman filter (PEM-FDKF) [14] replaces the standard normalized least mean squares (NLMS) type adaptive filter with a Kalman filter.

Traditional performance measures for AFC algorithms are the misadjustment, i. e. the normalized error between the true and estimated feedback path, and the increase in the maximum stable gain (MSG), i. e. the additional gain maintaining system stability upon the adaptive filter insertion [15]. These measures give information about the feedback system under investigation and are especially advantageous in simulated scenarios. A few recent studies, e. g. [16], [17], [18], have tackled the problem of objectively evaluating AFC algorithms in a more general way, that can be used to objectively evaluate the AFC algorithms of different commercial HAs. Not only did these studies include objective measures to evaluate instability (e. g. the MSG), but also objective measures evaluating oscillation and distortions.

This collection of different objective measures are helpful to evaluate several aspects related to the performance of a tested AFC algorithm, but are not readily interpretable in terms of perceptual sound quality [1]. Among different perceptually meaningful sound attributes, sound quality plays an important role in HA [19], [20]. Although sound quality can be heavily impacted by feedback and AFC artifacts, only a handful of sound-quality objective measures have been validated in a feedback scenario [16], [21]. Many other sound-quality objective measures were not specifically developed for feedback and AFC artifacts, and their typical application is that of assessing the sound-quality degradation caused by acoustic noise or by transmission channel effects (i. e. linear filtering) [22], [23]. The efficacy of these objective measures in a feedback/AFC scenario has not being investigated.

The goal of this paper is twofold. First, we present the results of a perceptual subjective sound-quality evaluation of three different AFC algorithms in order to validate the hypothesis of superior performance of the PEM-FDKF over the PEM-FDAF and the block normalized least mean squares (BNLMS) algorithm, as observed in simulations [14]. Second, the test signals used to subjectively evaluate sound quality, are tested using eight different existing objective measures, in order to evaluate the predictive capabilities of these objective measures.

The paper is organized as follows. In Section II, we describe different aspects of the current study, related to both the subjective and the objective evaluations. In Section III, we describe how the subjective evaluation has been carried out, focusing on the design details and choices of the experiment. In Section IV, we describe the different objective measures used to perform the objective evaluation. In Section V, we describe the performance criteria used to evaluate the predictive capabilities of the objective measures. In Section VI, we describe the results of the subjective and the objective evaluations, as well as the correspondence between the two. Finally, the conclusions are drawn in Section VII.

II. Materials and Methods

In this section, we give a detailed description of different aspects of the current study, i. e. the chosen AFC algorithms, the signal recordings and preprocessing, the employed speech and music material, and the post processing introduced to limit the annoyance of unstable loudspeaker signals.

A. Algorithms

Three AFC algorithms are considered in the current study: the BNLMS, the PEM-FDAF and the PEM-FDKF.

The BNLMS algorithm [24] is one of the simplest adaptive algorithms that can be used for AFC; other algorithms, such as affine projection and recursive least squares (RLS), exist but their increased complexity still represents a major hurdle in HAs applications [1]. We use a frame version, instead of the standard NLMS algorithm, in order to simplify the comparison with the other two algorithms which, being frequency-domain implementations, also operate in a frame fashion. The BNLMS was implemented using the `LMS Filter` block in Simulink.

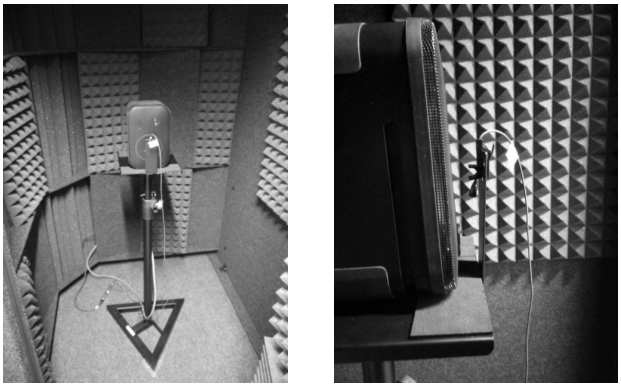
The PEM-FDAF [25] integrates the decorrelating properties of the PEM, a widely studied technique for AFC [10], [13], [26], [27], [12], into the frequency-domain adaptive filter (FDAF). An AFC algorithm based on the PEM assumes, in its simplest form, the source signal $v(t)$ to be the output of an autoregressive system, fed with a white noise input signal. In such a case, estimating the unknown source signal model and prefiltering the adaptive filter input and desired signal with the inverse signal model leads to an increased decorrelation between the source signal $v(t)$ and the loudspeaker signal $u(t)$, thus reducing the estimation bias.

Finally, the PEM-FDKF [14] integrates the PEM into the frequency-domain Kalman filter (FDKF) [28], and has been shown to provide superior performance over the PEM-FDAF [14]. The PEM-FDKF exploits the properties of the FDKF, leading to a faster convergence and to an optimal estimation of the step-size parameter [28], simplifying the design choices one has to make in order to use the algorithm.

The three algorithms were implemented in Simulink Real-time and run on a Speedgoat Audio real-time target machine (https://www.speedgoat.ch). The sampling frequency was set to $f_\mathrm{s} = 16\,\mathrm{kHz}$ and the resulting sampling time $T_\mathrm{s} = 62.5\,\mu\mathrm{s}$ was well above the task execution time of the algorithms, being lower than $2\,\mu\mathrm{s}$.

For all three algorithms, the frame size and the discrete Fourier transform (DFT) size were set to M, and the frame shift was set to R, where $M = 2R = 128$. The adaptive filter order was fixed to $n_{\hat{F}} = R - 1 = 63$, i. e. a $4\,\mathrm{ms}$ length.

Other algorithm-specific parameters were as follows: the stepsize of the BNLMS was either fixed to $\mu_\mathrm{anch} = 2 \times 10^{-2}$, when the algorithm was used to generate the anchor signal (cf. Section III-A), or to $\mu_\mathrm{BNLMS} = 2 \times 10^{-3}$. The adaptive stepsize of the PEM-FDAF was defined as in [14], with the fixed stepsize $\mu_\mathrm{FDAF} = 2 \times 10^{-2}$, and the regularization



(a) From the front. (b) From the side.

Fig. 2: Feedback scenario.

TABLE I: Test signals IDs, descriptions and employed gains.

ID	Speech/music material and description	$G(q,t)$ [dB]
$\mathtt{S_{f1}}$	LIST\list01\{09, 08}	K_2
$\mathtt{S_{f2}}$	LIST\list01\{10, 05}	K_2
$\mathtt{S_{f3}}$	——	K_3
$\mathtt{S_{f4}}$	LIST list01\{03, 01, 04}	K_2
$\mathtt{S_{f5}}$	LIST\list04\{01, 04, 05}	K_2
$\mathtt{S_{f6}}$	LIST\list04\{09, 07}	K_2
$\mathtt{S_{m1}}$	LISTm\list01\{01, 02, 03}	K_2
$\mathtt{S_{m2}}$	LISTm\list03\{01, 02}	K_2
$\mathtt{S_{m3}}$	LISTm\list03\{03, 04}	K_2
$\mathtt{S_{m4}}$	LISTm\list03\{05, 06}	K_2
$\mathtt{S_{m5}}$	LISTm\list03\{07, 08}	K_2
$\mathtt{S_{m6}}$	——	K_3
$\mathtt{M_1}$	B. White - *You're the first, the last, my everything* - Rhythm and blues, disco	K_1
$\mathtt{M_2}$	B. Bragg - *A New England* - Folk rock	K_1
$\mathtt{M_3}$	——	K_2
$\mathtt{M_4}$	Los Mensajeros de las Libertad - *El Contrapunto* - Latin American folk	K_1
$\mathtt{M_5}$	Ska-P - *El Gato Lopez* - Ska, punk	K_1
$\mathtt{M_6}$	M. Ravel - *Bolero* - Classical	K_1
$\mathtt{M_7}$	W. A. Mozart - *Magic Flute Overture* - Classical	K_1
$\mathtt{M_8}$	J. S. Bach - *Partita for Violin No. 2* - Classical	K_1

constant $\delta = 5 \times 10^{-5}$. The initial value of the error covariance matrix and the transition factor of the PEM-FDKF were set, respectively, to $\mathbf{P}(0) = 1 \times 10^{-4}\,\mathbf{I}_{M\times M}$ and $\alpha = 0.999$. For more details about the PEM-FDAF and the PEM-FDKF we refer to [25] and [14]. These parameters were chosen such that each algorithm would remain stable for most part of all the source signals employed in the study. At the same time, the values were chosen not to cause the typical erratic behavior of a poorly-tuned adaptive filter.

Finally, to calculate the prefilter for the PEM-FDAF and the PEM-FDKF, the filter order was fixed to 15, cf. [14], while the estimation window length was set to 160 [11].

B. Signal recordings and preprocessing

The loudspeaker and the microphone were, respectively, a JBL CONTROL 1 PRO loudspeaker and an AKG CK 97-O microphone, placed in close proximity within a sound booth, see Fig. 2. This was done to limit the length of the feedback path to similar values encountered in HAs applications [1]. The forward path transfer function $G(q,t)$ was modeled using a gain and a delay stage. The gain was kept constant throughout each measurement, but a total of three different constant values were employed in different measurements. The three values were: $K_1 = K_{\mathrm{MSG}}$, $K_2 = K_{\mathrm{MSG}} + 10\,\mathrm{dB}$, and $K_3 = K_{\mathrm{MSG}} + 15\,\mathrm{dB}$, where K_{MSG} was the MSG of our setup when no AFC was included, cf. Table I. The delay of the forward path transfer function $G(q,t)$, $d_G = 80\,\mathrm{samples}$, was kept constant in all the measurements.

An automated script handled the recording of the different loudspeaker signals originating from the different recording combinations, i. e. varying speech/music signals, AFC algorithms, forward path gains and other algorithm parameters. The loudspeaker was driven by a signal coming from one of the Speedgoat output ports, passed through an RME Fireface UCX soundcard, a Samson SERVO 200 amplifier and, finally, reaching the loudspeaker using the feed-through of the sound booth. The loudspeaker signals were stored in the flash memory of the Speedgoat, and subsequently transferred to the connected PC running Simulink.

The microphone was connected, using the feed-through of the sound booth, to a EURORACK MX 3242X mixer, providing the necessary phantom power and, from there, to one of the Speedgoat input ports.

The source signal, $v(t)$ in Fig. 1, was an anechoic speech or music signal, i. e. $a(t)$, filtered with a measured RIR and added to the microphone signal directly in the Simulink model. The signal-to-noise ratio (SNR) of the source signal $v(t)$ with respect to the background noise within the booth (stemming from the lab equipment outside the booth), i. e. $n(t)$, was fixed to $40\,\mathrm{dB}$ in all the recordings.

The frequency responses of the loudspeaker and microphone used in the setup, were objectively different than the frequency responses of the same components in a standard HA. In particular, the loudspeaker was much bigger and, as a consequence, had a much stronger low-frequency response. Additionally, the coupling of a HA receiver with the ear, provides a further filtering of the loudspeaker signal, that we did not account for in our scenario. Initial tests showed some problems related to this strong low-frequency behavior of the loudspeaker leading, e. g., to instabilities at low frequencies. The AFC algorithms tested in this study cannot cope with strong feedback at low frequencies, due to the limited number of adaptive filter coefficients [11]. Therefore, loudspeaker signals were filtered with an A-weighting filter [29], introducing a strong reduction of the frequencies below $500\,\mathrm{Hz}$, in order to partially account for the discrepancies with a real HA scenario.

C. Speech and music material

Both speech and music signals were used in the study. The speech material was taken from the LIST [30] and LISTm [31] corpora, consisting of Dutch sentences uttered by a female and a male speaker, respectively. Each signal used in the study was obtained by merging either two or three utterances from one of the lists of the corpora; this was done in order to have speech signals of length between 8 and $10\,\mathrm{s}$, as suggested in the listening test standard [32], cf. Section III-A. Five signals were taken from each of the corpora and processed using gain

K_2, i. e. $\mathtt{S_{f1}}$ to $\mathtt{S_{f5}}$ and $\mathtt{S_{m1}}$ to $\mathtt{S_{m5}}$; one of the five signals in both corpora (randomly chosen) was also processed using an additional forward path gain K_3, i. e. $\mathtt{S_{f3}}$ and $\mathtt{S_{m6}}$. In order to limit the number of stimuli and thus the duration of the listening test, only few signals were processed using gain K_3. Therefore, no statistical analysis can be done using the level as an independent variable. Nevertheless, the signals processed with gain K_3 were kept to offer a qualitative perspective on the results at higher gains. Further details of the utterances used to obtain the different speech signals employed in the sound-quality evaluation can be found in Table I.

The music material included a total of eight music signals, of different genres, drawn from different sources. Five music signals (including one repeated at two forward path gains), i. e. $\mathtt{M_1}$ to $\mathtt{M_6}$, were taken from the corpus associated with the Music Information Retrieval Evaluation eXchange (MIREX 2006) challenge [33]. Two signals, i. e. $\mathtt{M_7}$ to $\mathtt{M_8}$, were excerpts of classical music which have already been used to test AFC algorithms [1], [16]. All the signals were processed using a forward path gain K_1, and one of the signals was also processed using a forward path gain K_2. The choice of using a lower forward path gain (except for one recording) compared to the case of the speech material was motivated by literature results showing that AFC convergence is generally worse for music signals than for speech signals [1].

The implementations of the PEM-based algorithms used in this study rely on a low-order linear prediction model of the source signal and are, hence, not very good models for music signals [27]. Nevertheless, music signals were included given the raising evidence of hearing-impaired listeners enjoying music thanks to the advances in HA technology [34].

D. Signals post-processing

Due to the signal-dependent behavior of the AFC algorithms, guaranteeing stability of all the algorithms for all the different signals and the different parameters used in the listening test was found to be challenging. This was especially true for the anchor and the baseline algorithms (i. e. the BNLMS and the PEM-FDAF), which exhibited multiple instability periods during the recordings associated with different source signals. Some of the recordings were corrupted with artifacts that were too loud and, as a consequence, too disturbing for the sound-quality evaluation procedure.

Therefore, the test signals used in the listening test were obtained from the screened and post-processed loudspeaker signals, by means of targeted gain reductions in those time intervals corresponding to an unstable signal segment. Different gain reductions were applied, sometimes multiple times per each signal, using multiples of $-5\,\mathrm{dB}$. The gain reductions were applied in two speech signals processed by the BNLMS, in three speech signals processed by the PEM-FDAF, and in all the speech signals as well as in two music signals used as anchors.[1]

[1]A detailed summary of the gain reduction values applied to the different loudspeaker signals, including the time intervals at which the gains were applied, is provided at the following link https://lirias.kuleuven.be/bitstream/123456789/599139/5/fbMUSHRApp.csv.

The post-processing was only applied to the anchor and to the baseline algorithms since processing the signal through these three algorithms led, to a different extent, to an unstable behavior in some cases; the PEM-FDKF did not exhibit any strong instability for the processed files and, thus, no post-processing was applied. Due to the way the reference signal was obtained, no feedback (and, naturally, no AFC) artifacts were present hence no post-processing was needed.

Although the inclusion of the post-processing stage does necessarily alter the outcome of the listening test, since it might have lead to higher scores for the baseline algorithms and the anchor than without any post-processing, we would like to point out that the main hypothesis we aim to test here is the superior performance of the PEM-FDKF. Therefore, we believe that our methodology is valid, even though it might lead to a more conservative assessment of the perceptual improvements of the PEM-FDKF over the baseline algorithms. Nevertheless, in Section VI-A we report the results of both the cases including and excluding the signals where post-processing was applied.

III. Subjective Evaluation

In this section, we give a detailed description of different aspects related to the subjective evaluation, i. e. the actual listening test and the chosen statistical analysis.

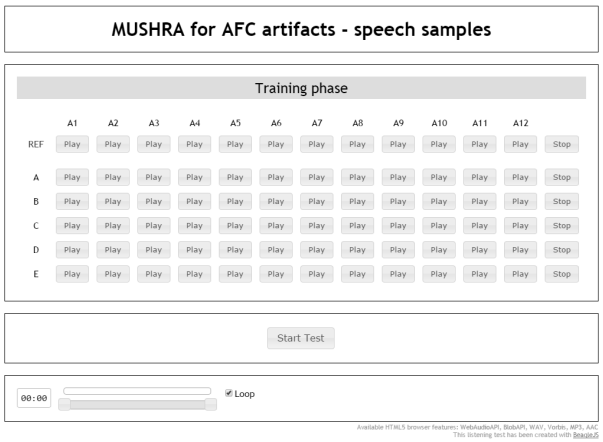
A. Listening test

The subjective sound-quality evaluation was performed using the Multi Stimulus Tests with Hidden Reference and Anchor (MUSHRA) [32]. The MUSHRA was preferred over other standards for sound-quality evaluation, e. g. [35], since the predicted impairments where supposed to be, at least, intermediate. The implementation of the MUSHRA was adapted from `beaqleJS`, a Javascript and HTML5 based framework for subjective sound-quality evaluation [36]. The adaptation included the addition of the MUSHRA training phase and several small edits.

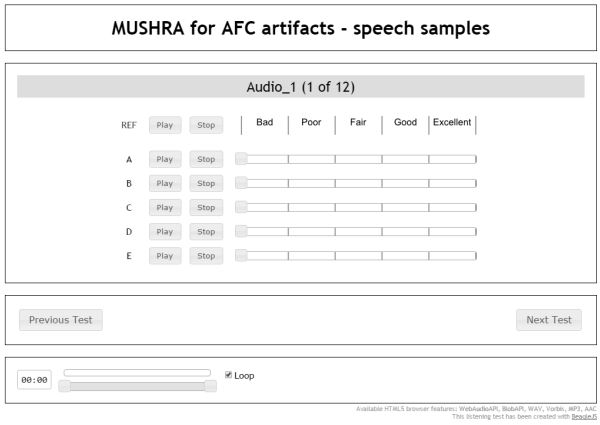
A total of 21 subjects were tested of which 12 were native Dutch speakers (the language of the speech corpora).[2] The subjects were 24 to 32 years old, and not all of them were experienced listeners, a recommended but not necessary condition for the test [32]. Prior to the actual listening test, the subjects were screened through an audiometric test run on a PC, using the APEX 3 program developed at KU Leuven ExpORL [37], connected to the RME Fireface UCX soundcard and a pair of standard audiometric headphones, the Sennheiser HDA200. The calibration of the setup was performed using the B&K 4153 artificial ear, including a B&K 4192 microphone. Upon testing, all the subjects were confirmed to be normal hearing.

The listening test was run on the same PC used for the audiological screening but, due to the small dimensions of the listening booth, an Apple iPad was used as an external user interface. The presentation level of the different test signals was $65\,\mathrm{dB\ SPL}$, calibrated using the noise calibration signal of the

[2]All participants gave written informed consent and ethical approval for the study was obtained from the Medical Ethics Committee UZ/KU Leuven.



(a) Screenshot of the training interface.



(b) Screenshot of the test interface.

Fig. 3: Screenshots of the training and test interfaces of the listening test.

LISTm corpus, and the signals were presented diotically via the same headphones used for the audiological screening. All the subjects acknowledged that the listening level was comfortable before starting the test.

The listening test was divided in two main parts based on the kind of audio material employed in each part (speech and music, respectively). Each of these two parts was additionally divided into a training and a test phase. The training phase was conducted first, and provided the subjects with all the conditions that were subsequently presented during the test phase, see Fig. 3a. This acclimation phase exposes the test subjects to all types of artifacts that are corrupting the test signals they were requested to assess in the actual test. The test phase consists in the presentation of the test signals, one at a time, processed through the different algorithms at test (cf. infra). The subjects must rate the sound quality associated with the different conditions compared to a reference (which is always accessible) on a scale from 0 to 100, with higher ratings indicating better sound quality, see Fig. 3b.[3] The average duration of the listening test, including a brief description of the test ($15\,\mathrm{min}$) as well as training and test for both parts ($62\,\mathrm{min} \pm 11\,\mathrm{min}$) was roughly $1\,\mathrm{h}\ 15\,\mathrm{min}$.

[3]The numerical values were not displayed in the test user interface, but the subject were explicitly told that the grading bar ranged from 0 to 100.

For each of the test signals, the subjects were presented with five conditions:

1) Three conditions corresponding to the loudspeaker signals obtained by running the three algorithms presented in Section II-A, i. e. the BNLMS, the PEM-FDAF, and the PEM-FDKF, possibly including a signal post-processing;
2) One hidden reference condition obtained by running the source signal through an "opened" feedback model, i. e. where the loudspeaker signal was logged right after the forward path stage but no signal was played by the actual loudspeaker, in order to avoid any feedback into the microphone signal. In this way, the loudspeaker signal presented the same characteristics, in terms of forward path gain, SNR, microphone transfer function, and A-weighting as the other signals; however, no feedback (and, naturally, no AFC) artifacts were present. The same signal was used both as the reference condition and as the hidden reference condition in the test [32].
3) One anchor specifically designed for this test, consisting of the loudspeaker signal obtained by running the BNLMS with non-optimal parameters (i. e. $\mu_{\mathrm{anch}} = 2 \times 10^{-2}$), possibly including a signal post-processing. This value was chosen manually in order to generate significant feedback and AFC artifacts, e. g. whistling and entrainment [1], in all the loudspeaker signals. Therefore, the anchor should be easily recognized as a worst-case condition. No additional anchor obtained with the low-pass filtering described in [32] was used since the sound-quality degradation caused by a low-pass filter does not resemble the sound-quality degradation caused by feedback and AFC artifacts.

The subjects were not told which of the five conditions corresponded to which algorithms. They were, however, instructed that a hidden reference was included as one of the five conditions and that at least one of the conditions should be rated 100. When giving the test instructions, we stressed that the references would not be clean, natural-sounding test signals and, despite that, they should be associated with the best achievable sound quality in the considered experiment, i. e. equivalent to 100 in the rating scale. We chose not to use clean test signals for the references as the subjects would have likely had a much harder time to grade the specific impact of feedback and AFC artifacts, and they could have graded (perhaps not even consciously) the actual degradation from the clean condition not related to feedback and AFC artifacts.

B. Statistical analysis

The results, presented in Section VI-A, from speech and music signals are analyzed separately. The chosen analysis technique is a mixed-design analysis of variance (ANOVA) [32], carried out with algorithm (excluding anchor and reference, i. e. the BNLMS, the PEM-FDAF, and the PEM-FDKF) and test signal as within-subject factors, while the status of the listener as a native speaker (often referred to, in the literature, as "nativeness") of the test subject is included as a between-subject factor. The three factors (named 'effects' for consistency with the literature) will be indicated, respectively,

as AL, SI, and NA. The decision of using non-native Dutch speakers was due to convenience. However, considering the effects of the subject's nativeness in other psychoacoustics task, e. g. when evaluating speech intelligibility [38], NA was introduced in order to quantify possible biases. Although NA is not expected to lead to differences in the case of music signals, the effect is kept in the analysis for the sake of consistency with the analyses performed on speech signals. In addition to the standard reported measures from an ANOVA analysis, i. e. degrees of freedom, F-values, and p-values p_A, the effect size will also be reported. The chosen effect size was the generalized eta-squared η^2_G, as it can be calculated along the ANOVA, and it was included given the increasing importance acknowledged to this metric for scientific reporting of statistical data [39], [40], [41]. The recommendations from [39] were used to interpret the effect of η^2_G: small if $\eta^2_G \leq 0.02$, medium if $\eta^2_G \approx 0.13$, and large if $\eta^2_G \geq 0.26$. In order to account for departures from sphericity, a Huynh-Feldt correction was applied to the p-values obtained from the ANOVA [32].

Along with the mixed-design ANOVA results, when the AL effect is found to be significant, the results of a Tukey's HSD test on the considered effect is also presented. This post hoc analysis can be used to find statistically significant differences, through the p-value p_T, between the results of the different algorithms and, hence, rank the algorithms in terms of overall sound quality. In this case the chosen effect size was the Cohen's d [39], using the pooled standard deviation between each pair as a normalization, whose values can be interpreted as follows: small if $0.2 < d < 0.3$, medium if $d \approx 0.5$, and large if $d > 0.8$ [39].

Whereas (mixed-design) ANOVAs are widely used to analyze MUSHRA data, especially within communities working in audio processing, based on their robustness to assumption violations [42], [43], [44], newer and more robust methods have been recently derived. Wilcox [45] collected several recent innovations in statistical methods to specifically deal with non normality and heteroscedasticity, as well as to achieve better control of type I errors. Most of these methods are now available in a number of R packages, such as WRS, WRS2, and welchADF [46], [47], among others. Therefore, an analysis was carried out with one of these packages, welchADF [48], implementing Johansen's general formulation of the Welch-James's test with approximate degrees of freedom with trimmed means, Winsorized variances and the use of bootsrap, as a way to verify the analysis carried out with the standard mixed-design ANOVA mentioned earlier. The results of the omnibus test are given in terms of the p-values p^{o}_{WJ}. The same statistics are applied as a post-hoc test to the AL effect, when the latter is found to be significant, and the results of the post-hoc test is returned in terms of Hochberg corrected p-values p^{p}_{WJ}.

IV. Objective Evaluation

Different objective measures chosen from the literature were applied to the postprocessed files used for the listening test: the HA audio quality index (HAAQI) [49], the HA speech intelligibility index (HASPI) [50], the HA speech quality index (HASQI) [19], [21], the perceptual evaluation of audio quality (PEAQ) [51], the perceptual-model-based quality prediction method (PEMO-Q) [52], the perceptual evaluation of speech quality (PESQ) [23], [53], the frequency-weighted log-spectral signal distortion (SD) [16], and the short-time objective intelligibility (STOI) [54]. All these objective measures are intrusive, meaning that they rely on a reference signal to compute the degradation. Not all of the objective measures are designed for sound-quality evaluation and, among those that were, most were not designed to evaluate the sound quality degradation caused feedback and AFC artifacts. Additionally, some of the objective measures have been designed to assess perceptual aspects of speech, not music, signals. We are aware that using these objective measures for music signals might not be the best strategy, due to the difference in the underlying auditory models (especially concerning the frequency ranges at which they operate). Nevertheless, we also computed the results for music signals in order to probe the robustness of the objective measures in non-standard conditions.

All the results obtained from the objective measures were normalized between 0 and 1, to simplify the comparison between different objective measures. Given that the SD does not provide a bounded value like the other objective measures, we mapped the maximum value found in the simulations to 1.

A. HAAQI, HASPI, and HASQI

The HAAQI, HASPI and HASQI are three indexes developed by Kates and Arehart to predict audio quality, speech intelligibility and speech quality, respectively [49], [50], [19], [21]. These three indexes are all relying on an internal representation of the reference and processed signal with an auditory model, including a modelling stage to account for hearing impairment.

The auditory model was initially adopted by HASQI Version 1 [19] to predict speech quality variations due to noise, nonlinear distortions, and linear filtering in normal hearing (NH) and hearing impaired (HI) subjects. Subsequently, an extended version of the index, HASQI Version 2 [21], was introduced to cope with the degradations caused by frequency compression, noise suppression, speech vocoding, acoustic feedback and feedback cancellation, and speech combined with modulated noise. HASPI [50] adapts part of the auditory model used by HASQI in order to predict speech intelligibility. HAAQI [49] uses the HASQI auditory model with different parameters fitted to the data of an extensive music-quality rating experiment.

Although the three indexes can be used as predictors of quality and intelligibility for both NH and HI subjects, they are only used to predict NH scores, given the nature of our subjective data. In this paper, we used three MATLAB implementations of HAAQI, HASPI, and HASQI Version 2, respectively, kindly provided by Prof. Kates.

B. PEAQ

The PEAQ is a perceptual measure standardized by the International Telecommunications Union (ITU) [51]. It relies on the combination of different auditory model outputs and has been developed to evaluate high-quality (i. e. low-distortion) digital codec systems. Two versions of PEAQ have been

described in the literature: a basic and an advanced one. No implementation of PEAQ is freely available, but there exist at least three non-conformant open source versions. In this paper, we used an implementation of the advanced PEAQ by Holters *et al*. [55].

C. PEMO-Q

The PEMO-Q [52] returns the perceptual similarity measure (PSM) between a reference and a processed signal. It is based on a quantitative model of the "effective" auditory processing [56], that has been psychoacoustically validated in diverse tasks. For this reason, the PEMO-Q gives a higher prediction accuracy over PEAQ in several tasks [52]. In this paper, we used a MATLAB version of PEMO-Q from the PEASS toolkit [57].

D. PESQ

The PESQ [23], [53], [58], as well as its wideband extensions [59], is part of an ITU Standard and is widely used to objectively assess the perceptual sound quality of speech. The PESQ score is obtained by processing both the reference and processed signal with an auditory model and combining different intermediate scores weighted from telephony data. In this paper, we used an implementation of the wideband PESQ built from the binaries provided by the ITU.

E. SD

The SD is an easy-to-compute objective measure to assess speech quality [16], [1]; it relies on the short-term power spectral density (PSD) of the reference and the processed signals, $P_r(f,t)$ and $P_u(f,t)$ respectively, and on a psychoacoustically-motivated frequency weighting, i. e. w_{ERB} [16], based on auditory critical bands in the Nyquist interval:

$$SD(t) = \sqrt{\int_0^{f_s/2} w_{\mathrm{ERB}} \left(10 \log_{10} \frac{P_u(f,t)}{P_r(f,t)}\right)^2 \mathrm{d}f}. \quad (1)$$

The mean value of SD over the entire signal was used in the evaluation. SD has been used to evaluate sound quality in different papers analyzing AFC algorithms, e. g. [16], [17], [18], [1], [60], based on the good correlation results between SD and MSG found in [16]. However, its performance has not yet been validated with sound quality subjective data.

F. STOI

The STOI is a correlation-based speech intelligibility measure that works on the temporal envelopes of short speech frames [54]. In recent years STOI has become extremely popular for the assessment of different kinds of speech enhancement algorithms due to its simplicity and good performance. Therefore, it was included in our batch of objective measures, even though it is neither designed for sound quality evaluation nor for sound quality deterioration caused by feedback and AFC artifacts. In this paper, we used a MATLAB implementation of STOI from the webpage of the author (http://www.ceestaal.nl/code/).

V. Performance Criteria

Five commonly-used performance criteria, chosen based on the analysis in [61], were used to evaluate the predictive capabilities of the objective measures.

Three different correlation coefficients were computed. First, the Pearson correlation coefficient r was calculated to quantify linear relationships between objective scores O and subjective scores S. Second, the Spearman rank correlation coefficient r_{spear} was calculated to quantify how well O rank compared to S.[4] Third, an additional Pearson correlation coefficient r_{sig} was calculated, assuming a possible nonlinear relationship between the scores O and S, in the following way:

1) A sigmoidal mapping was found between the pairs of scores O and S, i. e.
$$S = \frac{1}{1 + e^{-(\alpha_1 O + \alpha_2)}} \times 100\,\%, \quad (2)$$
through nonlinear regression; this results in $\hat{\alpha}_1$ and $\hat{\alpha}_2$, estimates of α_1 and α_2, respectively;
2) The scores O were mapped to the sigmoidal-mapped objective scores O_{sig} using $\hat{\alpha}_1$ and $\hat{\alpha}_2$, ;
3) The Pearson correlation coefficient r_{sig} was calculated between O_{sig} and S.

Finally, the per-condition (i. e. per-algorithm) root-mean-square error (RMSE) and its extension, the so-called epsilon-insensitive root-mean-square error (ε-RMSE), were used. The ε-RMSE extends the RMSE in the sense that it assesses the prediction error between O_{sig} and S by taking into account the variability of S, as recommended in the ITU-T Rec. P.863 [62]. The decision of including the RMSE, along with the ε-RMSE, was made since, in some cases, a value of ε-RMSE $= 0$ was obtained. This was due to the high variability of the subjective data and the specific formulation of the ε-RMSE.

After computing the five performance criteria for each of the objective measures, a test to assess the statistical indifference among pairs of results was performed. The chosen statistical test for the correlation-based criteria was the Fisher transformation z-test [61]; for the RMSE-based criteria, the test suggested in [61], based on the test proposed in [63], was chosen, i. e.:

$$T_{ij} = \max\left(0, \frac{\varepsilon\text{-RMSE}_i^2}{\varepsilon\text{-RMSE}_j^2} - F(0.05, C, C)\right), \quad (3)$$

with $F(0.05, C, C)$ being the F-value calculated at a 0.05 significance level and $C = 5$, the considered number of conditions (i. e. algorithms), as both degrees of freedom.[5] The hypothesis of statistical indifference is accepted if $T_{ij} = 0$, while it is rejected (i. e. ε-RMSE$_i$ performs statistically significantly worse than ε-RMSE$_j$) if $T_{ij} > 0$ [62], [63].

VI. Results and Discussion

The results and discussion in this section is divided into two subsections: the first discussing the statistical analysis

[4] Of the eight chosen objective measures, the SD is the only one increasing when the distortion increases; thus, the resulting correlation coefficients are negative. For this reason, we used the absolute value of the coefficients.

[5] The test requires the relation ε-RMSE$_i > \varepsilon$-RMSE$_j$ to hold [63].

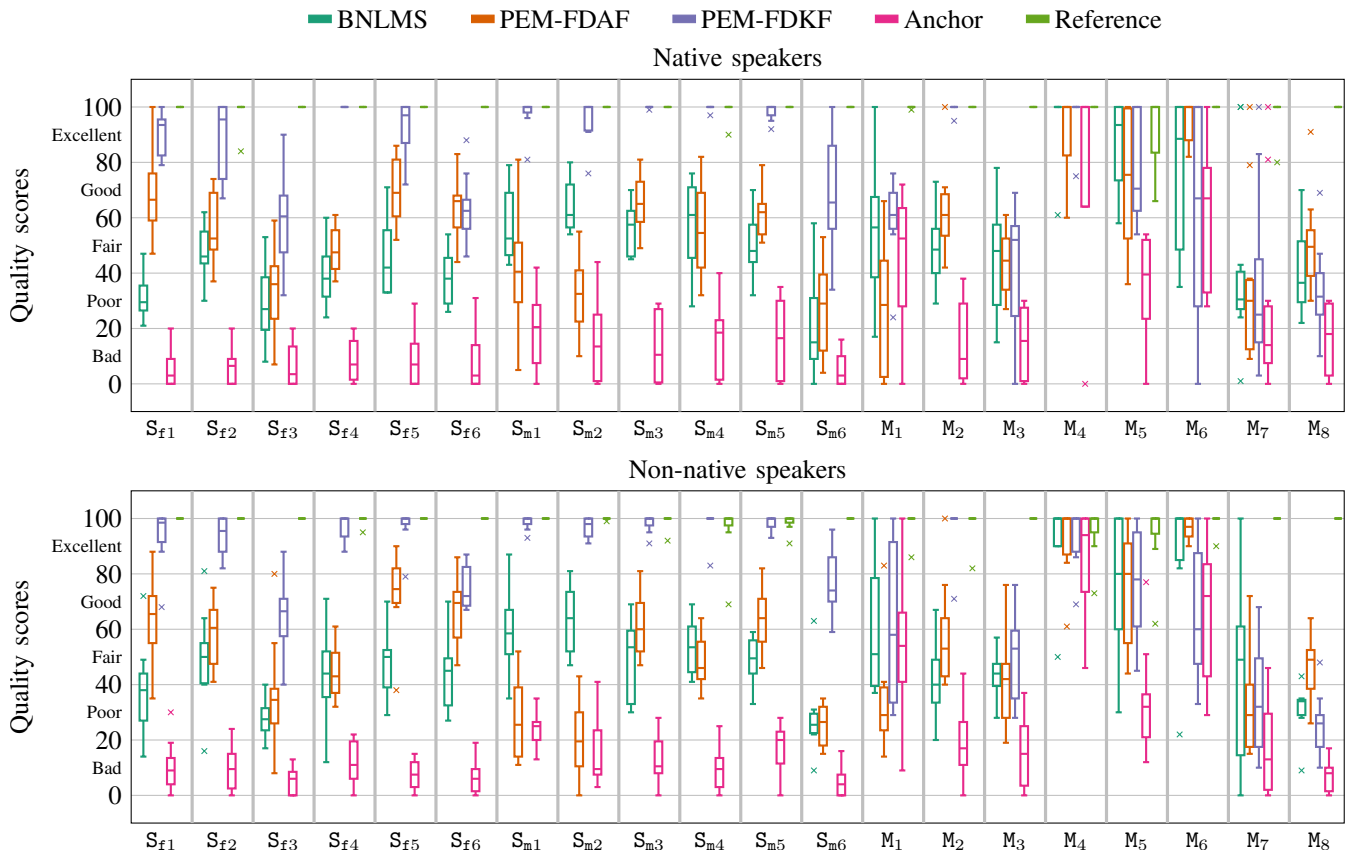


Fig. 4: Box plots of the subjective scores for the five different conditions (i. e. algorithms) across all different test signals. For each test signal, the algorithms plotted from left to right reflect the ordering expressed in the legend. The results from the native speakers and the non-native speakers are kept separated.

for the subjective scores alone, and the second discussing the relationships between the subjective and the objective scores.

A. *Subjective-evaluation measures*

The subjective scores collected with the MUSHRA were first post-screened, according to the standard [32], in order to remove those subjects (if any) assigning very high scores to the anchor, and/or very low scores to the hidden reference. One of the subjects from the native-speaker group had to be removed as he consistently rated the hidden reference lower than 90; this, possibly, indicates that the subject did not properly understand the instructions. Thus the total number of subjects used for the subsequent analysis was 20: 11 native speakers and 9 non-native speakers.

The results for native and non-native speakers, obtained for each test signal, are shown by means of boxplots in Fig. 4. For speech signals (i. e. $\mathtt{S}_{\mathtt{f}1}$ to $\mathtt{S}_{\mathtt{m}6}$), the boxplots for both groups are, in general, consistent across test signals, as indicated by a monotonic increase in the scores of, in ascending order, anchor, BNLMS, PEM-FDAF, PEM-FDKF, and reference. Some cases (i. e. $\mathtt{S}_{\mathtt{f}6}$, $\mathtt{S}_{\mathtt{m}1}$, $\mathtt{S}_{\mathtt{m}2}$, and $\mathtt{S}_{\mathtt{m}4}$ for native speakers and $\mathtt{S}_{\mathtt{m}1}$, $\mathtt{S}_{\mathtt{m}2}$, and $\mathtt{S}_{\mathtt{m}4}$ for non-native speakers) show a slight variation in the mentioned ascending order. These variations might be due to specific signal-dependent effects of the different algorithms but they do not impact on the average order of the five conditions over speech signals, as described later when discussing the results of the post-hoc tests. The increase of the gain from K_2 to K_3 in $\mathtt{S}_{\mathtt{f}2}$ to $\mathtt{S}_{\mathtt{f}3}$ and in $\mathtt{S}_{\mathtt{m}5}$ to $\mathtt{S}_{\mathtt{m}6}$ can be clearly seen as all three tested algorithms show a strong reduction in the median subjective score and an increase variance. Therefore, a strong effect might be expected with gain variations at high forward path gain values. However, as mentioned in Section II-C, since not all the source signals were processed with K_3, no statistical analysis could be carried out using the level as an independent variable. For music signals (i. e. $\mathtt{M}_1$ to $\mathtt{M}_8$), the monotonic increase in the scores of, in ascending order, anchor, BNLMS, PEM-FDAF, PEM-FDKF, and reference cannot be seen and it is noticeable how the three algorithms do not seem to consistently yield large subjective-score differences. In particular, two music signals, $\mathtt{M}_4$ and $\mathtt{M}_6$, return the highest degree of confusion between the different conditions, with more than 25 % of the subjects assigning a score ≥ 90 to the anchor.[6] The reason of this high variability, reflecting the subjects' inability to recognize the differences between the different processing algorithms, can be explained by considering the nature of those two music signals: both are characterized by a dynamic wideband content. Therefore, the

[6]According to [32], this indicates that the anchor processing might have not introduced enough degradation. For this reason, a separate statistical analysis will be carried out excluding $\mathtt{M}_4$ and $\mathtt{M}_6$.

TABLE II: Results of the statistical analyses. The first two columns indicate the analysis number and the effect variable The following four columns report the results of the mixed-design ANOVA analyses, i. e. the degrees of freedom, the F-values, the effect sizes η_G^2, and the corresponding Huynh-Feldt corrected p-values p_A. The last three columns report the results of the Welch-James's tests with approximate degrees of freedom, i. e. the degrees of freedom, the values of the Welch-James's statistics T_{WJ}, and the corresponding p-values p^{o}_{WJ}. The three significance codes applied to the p-values are the following: $p < .001$ (**boldface**), $.001 < p < .05$ (underlined), $.05 < p < .1$ (squiggly underlined), and $p > .1$ (standard typeface).

		mixed-design ANOVA				Welch-James's test		
Analysis	Effect	(DF_n, DF_d)	F	η_G^2	p_A	(DF_n, DF_d)	T_{WJ}	p^{o}_{WJ}
1 (Speech)	NA	$(1, 18)$	4.66×10^{-4}	6.08×10^{-6}	9.83×10^{-1}	$(1, 273)$	4.63×10^{-1}	4.86×10^{-1}
	AL	$(2, 36)$	7.33×10^{2}	7.21×10^{-1}	$\mathbf{6.89\times10^{-30}}$	$(2, 182)$	1.28×10^{3}	$\mathbf{< 1\times10^{-6}}$
	SI	$(11, 198)$	3.27×10^{1}	4.06×10^{-1}	$\mathbf{4.56\times10^{-24}}$	$(11, 109)$	4.00×10^{1}	$\mathbf{< 1\times10^{-6}}$
	NA:AL	$(2, 36)$	2.42	8.46×10^{-3}	1.03×10^{-1}	$(2, 182)$	3.02	5.21×10^{-2}
	NA:SI	$(11, 198)$	6.95×10^{-1}	1.43×10^{-2}	6.67×10^{-1}	$(11, 109)$	1.22	2.38×10^{-1}
	AL:SI	$(22, 396)$	3.19×10^{1}	3.66×10^{-1}	$\mathbf{4.85\times10^{-49}}$	$(22, 107)$	1.53×10^{1}	$\mathbf{< 1\times10^{-6}}$
	NA:AL:SI	$(22, 396)$	8.26×10^{-1}	1.47×10^{-2}	6.43×10^{-1}	$(22, 107)$	5.25×10^{-1}	9.4×10^{-1}
2 (Speech w/o $\mathtt{S_{f1}}$, $\mathtt{S_{f4}}$, $\mathtt{S_{m1}}$, $\mathtt{S_{m2}}$, and $\mathtt{S_{m6}}$)	NA	$(1, 18)$	5.54×10^{-4}	7.53×10^{-6}	9.81×10^{-1}	$(1, 158)$	5.68×10^{-1}	4.63×10^{-1}
	AL	$(2, 36)$	4.79×10^{2}	6.69×10^{-1}	$\mathbf{1.14\times10^{-26}}$	$(2, 107)$	5.90×10^{2}	$\mathbf{< 1\times10^{-6}}$
	SI	$(6, 108)$	3.15×10^{1}	4.16×10^{-1}	$\mathbf{3.05\times10^{-13}}$	$(6, 70)$	4.87×10^{1}	$\mathbf{< 1\times10^{-6}}$
	NA:AL	$(2, 36)$	7.43×10^{-1}	3.13×10^{-3}	4.83×10^{-1}	$(2, 107)$	1.22	3.13×10^{-1}
	NA:SI	$(6, 108)$	6.14×10^{-1}	1.37×10^{-2}	6.29×10^{-1}	$(6, 70)$	1.51	1.83×10^{-1}
	AL:SI	$(12, 216)$	1.49×10^{1}	1.84×10^{-1}	$\mathbf{2.24\times10^{-19}}$	$(12, 65)$	8.92	$\mathbf{< 1\times10^{-6}}$
	NA:AL:SI	$(12, 216)$	2.27×10^{-1}	3.42×10^{-3}	9.94×10^{-1}	$(12, 65)$	4.48×10^{-1}	9.35×10^{-1}
3 (Music)	NA	$(1, 18)$	1.69×10^{-3}	2.78×10^{-5}	9.68×10^{-1}	$(1, 125)$	2.11×10^{-2}	8.86×10^{-1}
	AL	$(2, 36)$	4.58	1.11×10^{-2}	1.69×10^{-2}	$(2, 87)$	3.83	1.8×10^{-2}
	SI	$(7, 126)$	5.95×10^{1}	5.58×10^{-1}	$\mathbf{1.84\times10^{-30}}$	$(7, 74)$	2.86×10^{2}	$\mathbf{< 1\times10^{-6}}$
	NA:AL	$(2, 36)$	4.04×10^{-1}	9.85×10^{-4}	6.71×10^{-1}	$(2, 87)$	1.93×10^{-1}	8.43×10^{-1}
	NA:SI	$(7, 126)$	6.32×10^{-1}	1.32×10^{-2}	6.96×10^{-1}	$(7, 74)$	6.49×10^{-1}	7.11×10^{-1}
	AL:SI	$(14, 252)$	1.98×10^{1}	2.34×10^{-1}	$\mathbf{2.68\times10^{-24}}$	$(14, 69)$	2.65×10^{1}	$\mathbf{< 1\times10^{-6}}$
	NA:AL:SI	$(14, 252)$	5.77×10^{-1}	8.84×10^{-3}	8.31×10^{-1}	$(14, 69)$	3.35×10^{-1}	9.8×10^{-1}
4 (Music w/o $\mathtt{M_4}$ and $\mathtt{M_6}$)	NA	$(1, 18)$	3.73×10^{-2}	6.91×10^{-4}	8.49×10^{-1}	$(1, 84)$	7.99×10^{-2}	7.68×10^{-1}
	AL	$(2, 36)$	1.79×10^{1}	5.23×10^{-2}	$\mathbf{4.57\times10^{-6}}$	$(2, 57)$	1.42×10^{1}	$\mathbf{< 1\times10^{-6}}$
	SI	$(5, 90)$	5.87×10^{1}	5.68×10^{-1}	$\mathbf{1.87\times10^{-19}}$	$(5, 58)$	4.02×10^{2}	$\mathbf{< 1\times10^{-6}}$
	NA:AL	$(2, 36)$	1.42×10^{-1}	4.38×10^{-4}	8.65×10^{-1}	$(2, 57)$	2.44×10^{-1}	7.58×10^{-1}
	NA:SI	$(5, 90)$	6.76×10^{-1}	1.49×10^{-2}	5.92×10^{-1}	$(5, 58)$	7.93×10^{-1}	5.17×10^{-1}
	AL:SI	$(10, 180)$	3.04×10^{1}	2.59×10^{-1}	$\mathbf{1.96\times10^{-24}}$	$(10, 53)$	3.56×10^{1}	$\mathbf{< 1\times10^{-6}}$
	NA:AL:SI	$(10, 180)$	5.79×10^{-1}	6.63×10^{-3}	7.74×10^{-1}	$(10, 53)$	2.95×10^{-1}	9.81×10^{-1}

typical narrowband artifacts of an AFC scenario are likely to be masked by the simultaneous wideband frequency content. A similar confusion is noticed in the scores of $\mathtt{M_1}$ where, however, the average scores (excluding the reference) tend to gather around the center of the sound-quality scale. This is due to the fact that $\mathtt{M_1}$ includes a very fast sequence of harmonic sounds (mainly violin or synthesized violin notes) followed by short silence segments and the resulting feedback is difficult to cancel with the considered algorithms, leading to a similar sound-quality degradation between the three algorithms and the anchor. However, unlike in $\mathtt{M_4}$ and $\mathtt{M_6}$, the perceptual difference with the reference was easy to notice. The increase of the gain from K_1 to K_2 in $\mathtt{M_2}$ to $\mathtt{M_3}$ causes an effect that is more difficult to interpret than for speech signals. The subjective scores show a very limited increase for the BNLMS, a small decrease for the PEM-FDAF, and a very strong decrease (including a large increase in the variance) for the PEM-FDKF.

The results of the statistical analysis are reported in Table II in terms of signal type, effect, results from the mixed-design ANOVA (i. e. degrees of freedom, F-values, effect sizes η_G^2, and p-values p_A), and results of the Welch-James's tests with approximate degrees of freedom (i. e. degrees of freedom, values of the Welch-James's statistics T_{WJ}, and p-values p^{o}_{WJ}). The analysis for the speech signals was done twice: once including all the speech signals, and once excluding those speech signals that required a post-processing in either of the baseline conditions, i. e. $\mathtt{S_{f1}}$, $\mathtt{S_{f4}}$, $\mathtt{S_{m1}}$, $\mathtt{S_{m2}}$, and $\mathtt{S_{m6}}$, from now referred to as postprocessed speech signals. The analysis for the music signals was done twice: once including all the signals from $\mathtt{M_1}$ to $\mathtt{M_8}$, and once excluding $\mathtt{M_4}$ and $\mathtt{M_6}$, considering the problems of $\mathtt{M_4}$ and $\mathtt{M_6}$ explained in Footnote 6.

1) Speech signals: The first mixed-design ANOVA was carried out using all the speech signals as SI within-subject factor, cf. top portion of Table II. The analysis illustrates that the scores by native and non-native speakers were not significantly different. A significant effect with $p_A < .001$ was found for AL, SI and the interaction effect between AL and SI (AL:SI). According to the guidelines found in literature [39], the effect size for AL, SI and AL:SI, can be considered as large. For instance, in the case of AL, $\eta_G^2 = 7.21 \times 10^{-1}$ means that, considering the data from a given algorithm, we can explain $72\,\%$ of the subjective scores variability for that given algorithm evaluation. Therefore, we can confidently claim that different algorithms lead to different results. The significant

effect on SI was to be expected, due to the signal-dependent nature of feedback and AFC artifacts. No significant effect was found for the two remaining two-way interactions, NA:AL and NA:SI, and for the single three-way interaction NA:AL:SI, and all the corresponding effect sizes were found to be small. The results of the Welch-James's test, in terms of p-values significance, confirm the results of the mixed-design ANOVA for all the effects but the NA:AL, see Table II. Given that the AL effect was found to be significant for both analyses, a Tukey's HSD test and a post-hoc Welch-James's test were applied to assess the statistical difference between the three tested algorithms. The mean value of the scores for the three algorithms were as follows: 46.4 for the BNLMS, 51.9 for the PEM-FDAF, and 89.5 for the PEM-FDKF. The p-values p_T obtained from the Tukey's HSD test, the p-values p^{p}_{WJ} obtained from the post-hoc Welch-James's test, and the effect sizes d were as follows: [$p_T < .001$, $p_{WJ} < .001$, $d = 0.29$] between the BNLMS and the PEM-FDAF, [$p_T < .001$, $p_{WJ} < .001$, $d = 2.56$] between the BNLMS and the PEM-FDKF, and [$p_T < .001$, $p_{WJ} < .001$, $d = 2.03$] between the PEM-FDAF and the PEM-FDKF.

The second mixed-design ANOVA was carried out using only seven of the twelve speech signals (i. e. excluding the postprocessed speech signals) as SI within-subject factor, cf. second portion from the top of Table II. The analysis illustrates that the scores by native and non-native speakers were not significantly different. A significant effect with $p_A < .001$ was found for AL, SI and AL:SI. The effect size for AL and SI is large, while it is small for AL:SI, As in the previous case where all the speech signals were included, we can confidently claim that different algorithms lead to different results. No significant effect was found for the three remaining multi-way interactions, and all the corresponding effect sizes were found to be small. The results of the Welch-James's test, in terms of p-values significance, confirm the results of the mixed-design ANOVA. The mean value of the scores for the three algorithms were as follows: 47.0 for the BNLMS, 59.1 for the PEM-FDAF, and 87.6 for the PEM-FDKF. The results of the two post-hoc tests were as follows: [$p_T < .001$, $p_{WJ} < .001$, $d = 0.75$] between the BNLMS and the PEM-FDAF, [$p_T < .001$, $p_{WJ} < .001$, $d = 2.43$] between the BNLMS and the PEM-FDKF, and [$p_T < .001$, $p_{WJ} < .001$, $d = 1.61$] between the PEM-FDAF and the PEM-FDKF.

Overall, both analyses with the speech signals suggest that there is a strong statistically significant effect for AL, SI and AL:SI, backed up by large effect sizes. According to the post-hoc tests results, the PEM-FDKF outperforms the other two algorithms in terms of subjectively-evaluated sound quality, given the very low p-values and the large effect sizes. As mentioned in Section II-D, the inclusion of the postprocessed speech signals might lead to a more conservative subjective-scores difference between the PEM-FDKF and the BNLMS/PEM-FDAF. Therefore, it is expected that an even bigger difference would be found, had the postprocessed speech signals not been included. The improvement of the PEM-FDAF over the BNLMS is significant in both analyses, but the effect size varies from small to large once the postprocessed speech signals are removed. Therefore, even though there seems to be an existing difference, strong evidence cannot be drawn from our experiment. An experiment with increased number of speech signals, or the inclusion of more speech signals processed with a lower forward path gain might strengthen such evidence.

2) Music signals: The third mixed-design ANOVA was carried out using the music signals as SI within-subject factor, cf. third portion of Table II. The analysis illustrates that the scores by native and non-native speakers were not significantly different. A significant effect was found for AL ($p_A < .05$) SI ($p_A < .001$) and the interaction effect AL:SI ($p_A < .001$). In this case, the effect size for both SI and AL:SI is large, while the effect size for AL is small. Compared to the analysis for the speech signals, the higher level of the p-value for AL ($p_A < .05$ versus $p_A < .001$) is likely caused by the higher variability of the subjective scores and the similarity of the performance of the different algorithms. Nevertheless, the null hypothesis (i. e. equal mean for the three algorithms) could still be rejected using a significance threshold of $5\,\%$. However, the very small effect size suggests that the overall score difference caused by the three tested algorithms might be, in reality, negligible. No significant effect was found for the three remaining multi-way interactions, and all the corresponding effect sizes were found to be small. The results of the Welch-James's test, in terms of p-values significance, confirm the results of the mixed-design ANOVA. The mean value of the scores for the three algorithms were as follows: 62.3 for the BNLMS, 61.7 for the PEM-FDAF, and 65.9 for the PEM-FDKF. The results of the two post-hoc tests were as follows: [$p_T = 9.13 \times 10^{-1}$, $p_{WJ} = 3.63 \times 10^{-1}$, $d = 0.02$] between the BNLMS and the PEM-FDAF, [$p_T = 3 \times 10^{-2}$, $p_{WJ} = 8.17 \times 10^{-2}$, $d = 0.13$] between the BNLMS and the PEM-FDKF, and [$p_T = 9.09 \times 10^{-3}$, $p_{WJ} = 6.86 \times 10^{-3}$, $d = 0.15$] between the PEM-FDAF and the PEM-FDKF.

The fourth mixed-design ANOVA was carried out using only six of the eight music signals (i. e. without $\mathtt{M}_4$ and $\mathtt{M}_6$) as SI within-subject factor, cf. bottom portion of Table II. The analysis illustrates that the scores by native and non-native speakers were not significantly different. A significant effect with $p_A < .001$ was found for AL, SI, and the interaction effect AL:SI. The effect size for both SI and AL:SI is still large, while effect size for AL is still small. Compared to the previous case, the removal of $\mathtt{M}_4$ and $\mathtt{M}_6$ leads to a lower p-value for AL ($p_A < .001$ versus $p_A < .05$), but the effect size remains small. This, again, suggests that the overall score difference caused by the three tested algorithms might be, in reality, negligible. No significant effect was found for the three remaining multi-way interactions, and all the corresponding effect sizes were found to be small. The results of the Welch-James's test, in terms of p-values significance, confirm the results of the mixed-design ANOVA. The mean value of the scores for the three algorithms were as follows: 55.3 for the BNLMS, 53.1 for the PEM-FDAF, and 62.5 for the PEM-FDKF. The results of the two post-hoc tests were as follows: [$p_T = 3.44 \times 10^{-1}$, $p_{WJ} = 1.73 \times 10^{-1}$, $d = 0.08$] between the BNLMS and the PEM-FDAF, [$p_T < .001$, $p_{WJ} < .001$, $d = 0.27$] between the BNLMS and the PEM-FDKF, and [$p_T < .001$, $p_{WJ} < .001$,

$d = 0.35$] between the PEM-FDAF and the PEM-FDKF.

Overall, both analyses with the music signals suggest that there is a strong statistically significant effect for SI and AL:SI, backed up by large effect sizes. For AL, the p-values are $< .001$ only when removing M_4 and M_6, but the effect sizes remain small. The post-hoc tests suggest that there might be a slight perceptual sound-quality improvement when using the PEM-FDKF over the PEM-FDAF and the BNLMS, but the effect is minimal. This suggests that the statistically significant improvement obtained for the PEM-FDKF over the other two algorithms in terms of subjectively-evaluated sound quality might not be relevant. The removal of M_4 and M_6 does decrease the p-values but does not, substantially, change the effect size. On the other hand, there is no significant difference between the perceived sound quality when using either the PEM-FDAF or the BNLMS. Unlike for speech signals, no postprocessing was necessary for music signals.

B. Objective measures and performance criteria

The results of the comparisons between the different objective scores and the subjective scores are shown in the scatter plots in Fig. 5. Each of the subplots depicts the scores for a single objective measure plotted against the subjective scores. Overlayed to each scatter plot is the sigmoid function fitted on the data. The scores obtained using either speech or music signals are kept separated (first two and last two columns from the left, respectively).

The results for speech signals show that, for most of the objective measures, there is a good visual correspondence between objective and subjective scores. For three of the objective measures, i. e. HASPI, PEMO-Q, and STOI, it can be observed how most of the objective scores are located in the right-hand side of the plot, suggesting that the predictions yield values in a narrower range.[7] This effect alone is not enough to dismiss the predictive capabilities, but it makes it harder to visually evaluate how good the prediction is. The remaining objective measures yield scores distributed along the whole range. It can also be observed how for some of the objecive measures, e. g. HAAQI, HASQI, PEAQ, PESQ, and mean(SD), the results from single algorithms are lumped on a side of the sigmoid fitting (see, e. g., the PEM-FDKF results for HAAQI). Despite relying on a similar auditory model, HASPI and HASQI yield quite different results. Both models attempt to quantify changes in the modulation of the speech envelope between the reference and degraded signal, through nonlinear processing. However, in the case of HASQI, the changes in the long-term signal spectrum between the reference and degraded signal are quantified, too. We think this difference yield to the large gap in the results between HASPI and HASQI.

The results for music signals show a similar outcome: a good visual correspondence between objective and subjective scores in most cases, except for HASPI and PEMO-Q. STOI, in this case, yields scores more widely distributed than for the speech signals.

[7]HASPI and STOI are speech intelligibility objective measures, but in this study they were only correlated to subjective sound-quality scores and exposed to non-standard artifacts. Therefore, despite their high values, we cannot claim that they are effective to measure speech intelligibility in these conditions.

TABLE III: Results obtained using the different performance criteria when comparing objective and subjective scores, separately calculated for speech and music signals. With the **boldface** we indicate the best performances (statistically indifferent) computed for r_{sig} and ε-RMSE.

Signal type	Objective measure	$\lvert r\rvert$	$\lvert r_{\mathrm{spear}}\rvert$	r_{sig}	RMSE	ε-RMSE
speech	HAAQI	0.95	0.96	**0.96**	7.60	**0.06**
	HASPI	0.56	0.95	0.58	34.58	25.61
	HASQI	0.90	0.94	0.92	11.85	4.44
	PEAQ	0.88	0.94	**0.94**	7.39	-
	PEMO-Q	0.89	0.94	**0.93**	9.81	1.23
	PESQ	0.94	0.95	**0.95**	7.45	0.39
	mean(SD)	0.78	0.90	0.88	12.08	3.35
	STOI	0.88	0.95	0.92	14.15	3.64
music	HAAQI	0.81	0.85	0.84	2.74	-
	HASPI	0.73	0.85	0.72	22.29	15.29
	HASQI	0.87	0.89	**0.87**	4.17	-
	PEAQ	0.87	0.90	**0.90**	2.91	-
	PEMO-Q	0.83	0.82	0.81	6.64	**3.01**
	PESQ	0.92	0.93	**0.92**	5.43	-
	mean(SD)	0.76	0.81	0.79	3.87	-
	STOI	0.86	0.87	**0.85**	5.69	**2.53**

In order to quantify these aspects, the results in terms of the different performance criteria, defined in Section V, are shown in Table III. The best performances, separately calculated for speech and music signals, are reported in bold (up to statistical indifference) for two performance criteria, as done in [61], i. e. r_{sig} and ε-RMSE. The cases that yielded a value of ε-RMSE $= 0$ are indicated by '-'.

The results for speech signals, in terms of correlation-based criteria, show that HASPI performs the worst in terms of r and r_{sig}. The mean(SD), too, performs poorly compared to the other objective measures, all yielding values of $|r| > 0.8$ and values of $r_{\mathrm{sig}} \geq 0.9$. Nevertheless, the monotonicity is maintained by all objective measures, all yielding $|r_{\mathrm{spear}}| \geq 0.9$. The best-performing objective measure, using r_{sig} as a predictor, is the HAAQI reaching $r_{\mathrm{sig}} = 0.96$, statistically indifferent from the values of PEAQ, PEMO-Q, and PESQ. In terms of RMSE-based criteria, the situation is unchanged, with those objective measures yielding high correlations also yielding low RMSE and ε-RMSE. In this case, too, the HAAQI yields the best performance with a ε-RMSE $= 0.06$, statistically different from all the other ε-RMSE values. Although HAAQI and PEAQ were mainly designed to predict subjective scores for music signals, they seem to yield very good predictive capabilities for speech signals. These results suggest that HAAQI, PEAQ, PEMO-Q, or PESQ might be a better choice over SD to objectively evaluate the speech quality when using AFC algorithms. Unexpectedly, STOI, too, performed well even though it has primarily been used to predict speech intelligibility.

The results for music signals, in terms of correlation-based criteria, show, too, that HASPI performs the worst for all three criteria, followed by the mean(SD). The best-performing objective measure is PEAQ reaching $r_{\mathrm{sig}} = 0.90$, statistically indifferent from the values of HASQI, PESQ and STOI. In terms of RMSE-based criteria, the RMSE values are, in this

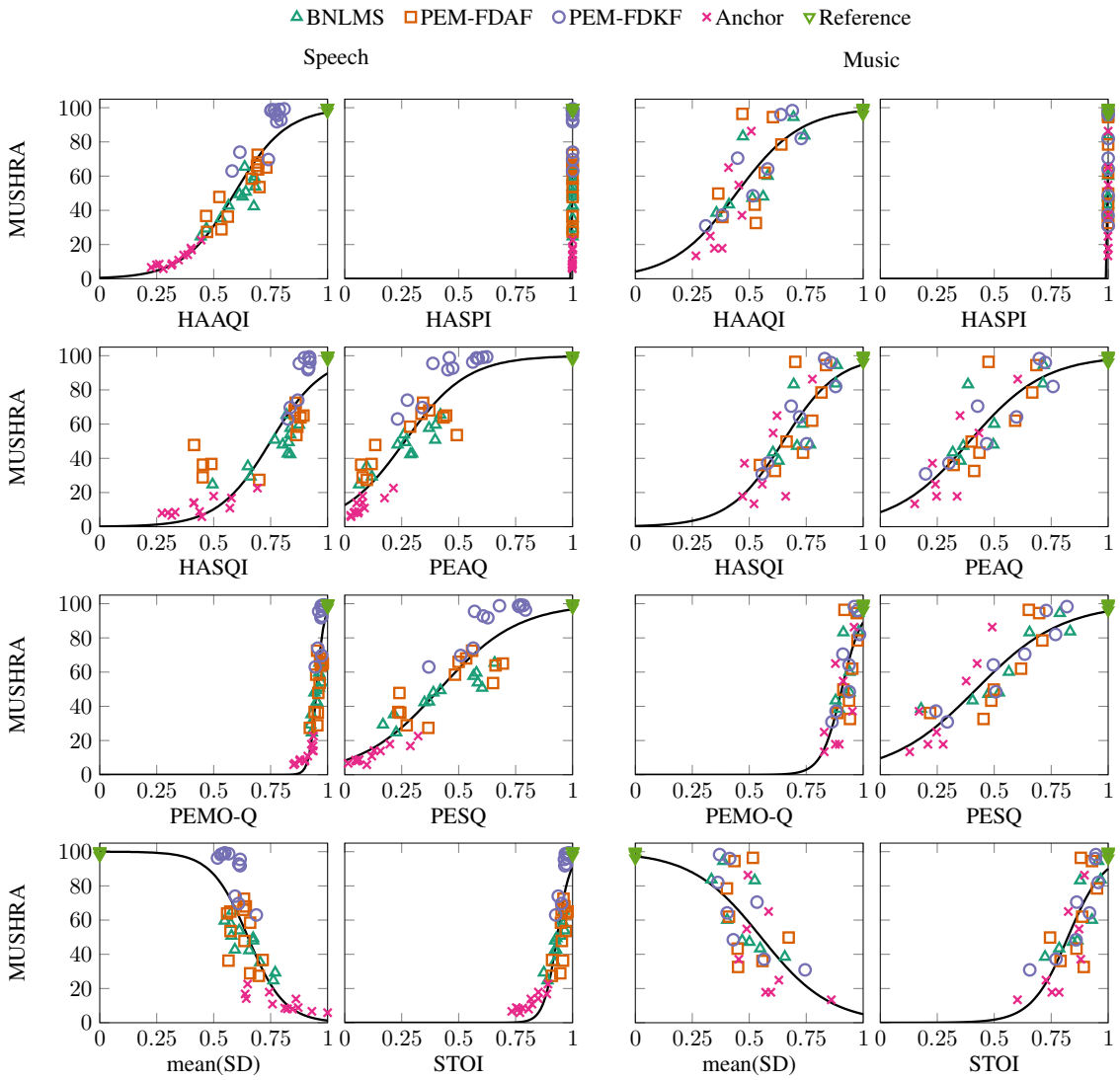


Fig. 5: Scatter plots of subjective scores versus objective scores, including the sigmoid mapping calculated for the set of data within each scatter plot.

case, lower than those found for speech signals. This can be understood by considering the aforementioned phenomenon involving the outliers lumped on one side of the sigmoid fittings. These outliers are more common in speech signals than in music signals, and the RMSE is very sensitive to outliers, leading to lower RMSE values compared to those found for speech signals. When comparing the RMSE values with the correlation-based measures, we see again a correspondence in the objective measures yielding high correlations and low RMSE and ε-RMSE. However, the ε-RMSE is less informative in this case as for most of the objective measures the variability is too large to yield a useful ε-RMSE. PEMO-Q and STOI yields the best, statistically indifferent, performance in terms of ε-RMSE.

Overall, these results point out that several existing objective measures can effectively be used to predict the perceptually-evaluated sound quality of signals subject to acoustic feedback and processed by AFC algorithms. Although some objective measures perform better than others, in terms of a given performance criteria, the differences might be very limited and not univocally represented by the results from all the other performance criteria. Therefore, a single objective measure cannot be selected as the best performing one, in terms of all the different criteria and for both speech and music signals. Some of the objective measures tested in this study are widespread, e. g. PEAQ, PESQ, and STOI, and their use could simplify the assessment of AFC algorithms working in conjunction with other processing stages usually included in a HA, e. g. noise reduction and compression algorithms. Nevertheless, further tests including more subjects and test signals are necessary to successfully generalize the outcomes of this study.

VII. Conclusions

In this study, we have presented an experiment to assess the perceptually-evaluated sound quality of three different AFC algorithms processing speech and music signals.

The analysis of the collected data have shown that both native and non-native speakers perform in a statistically indif-

ferent way for both music and speech signals. The analysis on speech signals and on a partial set of the music signals has shown that both SI and AL cause a highly significant effect, but the effect size for AL is large only for speech signals. The effect on AL has subsequently been studied with a post hoc analysis, in order to find a ranking of the algorithms performance: the PEM-FDKF has been found to be the algorithm providing best performance in both cases, where the large effect size for speech signals suggests the generalizability of the result.

The subjective scores have been compared to objective scores obtained with a set of well-known objective measures. A very high correlation, using different correlation coefficients, has been found when using the objective measures designed to assess perceptual sound quality. Low RMSE-based values have been found for those objective measures yielding high correlations. These results suggest that different existing objective measures might be used to predict the sound-quality degradation caused by acoustic feedback and AFC artifacts, for both speech and music signals, potentially reducing the testing times and costs of assessing AFC algorithms.

ACKNOWLEDGMENT

The authors would like to thank Mr. João Felipe Santos and Prof. Tom Francart for helpful discussions, and Prof. James Kates for making available the implementations of the HAAQI, HASPI, and HASQI measures.

REFERENCES

[1] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288 –327, Feb. 2011.

[2] J. Hellgren and U. Forssell, "Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 906 –913, Nov. 2001.

[3] M. Guo, S. Jensen, and J. Jensen, "Novel acoustic feedback cancellation approaches in hearing aid applications using probe noise and probe noise enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2549 –2563, Nov. 2012.

[4] C. R. C. Nakagawa, S. Nordholm, and W. Y. Yan, "Feedback cancellation with probe shaping compensation," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 365–369, Mar 2014.

[5] G. Schmidt and T. Haulick, "Signal processing for in-car communication systems," *Signal Process.*, vol. 86, no. 6, pp. 1307 – 1326, 2006.

[6] M. Guo, S. H. Jensen, J. Jensen, and S. L. Grant, "On the use of a phase modulation method for decorrelation in acoustic feedback cancellation," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug 2012, pp. 2000–2004.

[7] F. J. van der Meulen, S. Kamerling, and C. P. Janse, "A new way of acoustic feedback suppression," in *Preprints AES 104th Conv.*, Amsterdam, The Netherlands, May 1998.

[8] M. G. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 443–453, Jul. 2000.

[9] L. Ljung, Ed., *System Identification: Theory for the User (2nd Edition)*. Upper Saddle River, NJ: Prentice Hall, 1999.

[10] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3749–3763, Oct. 2005.

[11] A. Spriet, G. Rombouts, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids," *J. Franklin Instit.*, vol. 343, no. 6, pp. 545 – 573, Aug 2006.

[12] G. Rombouts, T. van Waterschoot, K. Struyve, and M. Moonen, "Acoustic feedback cancellation for long acoustic paths using a nonstationary source model," *IEEE Trans. on Signal Process.*, vol. 54, no. 9, pp. 3426–3434, Sep 2006.

[13] J. M. Gil-Cacho, T. van Waterschoot, M. Moonen, and S. H. Jensen, "Wiener variable step size and gradient spectral variance smoothing for double-talk-robust acoustic echo cancellation and acoustic feedback cancellation," *Signal Process.*, vol. 104, pp. 1–14, Nov 2014.

[14] G. Bernardi, T. van Waterschoot, J. Wouters, and M. Moonen, "Adaptive feedback cancellation using a partitioned-block frequency-domain Kalman filter approach with PEM-based signal prewhitening," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 9, pp. 1480–1494, Sep. 2017.

[15] J. Maxwell and P. Zurek, "Reducing acoustic feedback in hearing aids," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 304–313, Jul. 1995.

[16] A. Spriet, M. Moonen, and J. Wouters, "Evaluation of feedback reduction techniques in hearing aids based on physical performance measures," *J. Acoust. Soc. Amer.*, vol. 128, no. 3, pp. 1245–1261, Sep. 2010.

[17] N. Madhu, J. Wouters, A. Spriet, T. Bisitz, V. Hohmann, and M. Moonen, "Study on the applicability of instrumental measures for black-box evaluation of static feedback control in hearing aids," *J. Acoust. Soc. Amer.*, vol. 130, no. 2, pp. 933–947, Aug 2011.

[18] T. Bisitz, N. Madhu, A. Spriet, J. Wouters, M. Moonen, B. Kollmeier, and V. Hohmann, "Relation between different measures of feedback in hearing aids and gain measures derived from percentile analysis," *Acta Acust. united Ac.*, vol. 98, no. 4, pp. 676–684, Jul. 2012.

[19] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *J. Audio Eng. Soc*, vol. 58, no. 5, pp. 363–381, 2010.

[20] S. Kochkin, "Customer satisfaction with hearing instruments in the digital age," *Hear. J.*, vol. 58, pp. 30–37, Sep. 2005.

[21] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) version 2," *J. Audio Eng. Soc*, vol. 62, no. 3, pp. 99–117, 2014.

[22] P. C. Loizou, *Speech Enhancement. Theory and Practice*. Taylor and Francis, 2007.

[23] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Geneva, Switzerland, Feb. 2001.

[24] G. Clark, S. Mitra, and S. Parker, "Block implementation of adaptive digital filters," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 3, pp. 744–752, Jul. 1981.

[25] J. M. Gil-Cacho, T. van Waterschoot, M. Moonen, and S. H. Jensen, "A frequency-domain adaptive filter (FDAF) prediction error method (PEM) framework for double-talk-robust acoustic echo cancellation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2074–2086, Dec. 2014.

[26] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, and S. H. Jensen, "Improved prediction error filters for adaptive feedback cancellation in hearing aids," *Signal Process.*, vol. 93, no. 11, pp. 3062 – 3075, 2013.

[27] T. van Waterschoot and M. Moonen, "Adaptive feedback cancellation for audio applications," *Signal Process.*, vol. 89, no. 11, pp. 2185 – 2201, 2009.

[28] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140 – 1156, Jun. 2006.

[29] IEC Rec. 61672-1:2013, "Electroacoustics - Sound level meters - Part 1: Specifications," International Electrotechnical Commission, Geneva, Switzerland, Sep. 2013.

[30] A. van Wieringen and J. Wouters, "LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for flanders and the netherlands," *Int. J. Audiol.*, vol. 47, no. 6, pp. 348–355, 2008.

[31] S. Jansen, R. Koning, J. Wouters, and A. van Wieringen, "Development and validation of the leuven intelligibility sentence test with male speaker (LIST-m)," *Int. J. Audiol.*, vol. 53, no. 1, pp. 55–59, 2014.

[32] ITU-T Rec. BS.1534-3, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," International Telecommunication Union, Geneva, Switzerland, Oct. 2015.

[33] "Music Information Retrieval Evaluation eXchange (MIREX 2006)," 2007, Date last accessed: Mar. 01, 2017. [Online]. Available: http://www.music-ir.org/mirex/wiki/2006:Main_Page

[34] M. R. Leek, M. R. Molis, L. R. Kubli, and J. B. Tufts, "Enjoyment of music by elderly hearing-impaired listeners," *J. Am. Acad. Audiol.*, vol. 19, no. 6, pp. 519–526, 2008.

[35] ITU-T Rec. BS.1116-3, "Method for the Subjective Assessment of Small Impairments in Audio Systems," International Telecommunication Union, Geneva, Switzerland, Feb. 2015.

[36] S. Kraft and U. Zölzer, "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conf.*, Karlsruhe, Germany, Nov. 2014.

[37] T. Francart, A. Van Wieringen, and J. Wouters, "APEX 3: a multi-purpose test platform for auditory psychophysical experiments," *J. Neurosci. Methods*, vol. 172, no. 2, pp. 283–293, 2008.
[38] S. J. van Wijngaarden, H. J. M. Steeneken, and T. Houtgast, "Quantifying the intelligibility of speech in noise for non-native listeners," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1906–1916, 2002.
[39] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1988.
[40] G. M. Sullivan and R. Feinn, "Using effect size–or why the P value is not enough," *J. Grad. Med. Educ.*, vol. 4, no. 3, pp. 279–282, 2012.
[41] M. Schoeffler, A. Silzle, and J. Herre, "Evaluation of spatial/3D audio: Basic audio quality versus quality of experience," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 75–88, Feb. 2017.
[42] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Rev. Edu. Res.*, vol. 42, no. 3, pp. 237–288, Sep. 1972.
[43] L. M. Lix, J. C. Keselman, and H. J. Keselman, "Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test," *Rev. Edu. Res.*, vol. 66, no. 4, pp. 579–619, Dec. 1996.
[44] H. Finch, "Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated," *Methodology*, vol. 1, no. 1, pp. 27–38, Sep. 2005.
[45] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing, 3rd Edition*. Waltham, MA: Academic Press, 2012.
[46] P. Mair and R. Wilcox, *WRS2: A Collection of Robust Statistical Methods*, May 2017, R package version 0.9-2. [Online]. Available: https://cran.r-project.org/web/packages/WRS2/index.html
[47] P. J. Villacorta, *welchADF: Welch-James Statistic for Robust Hypothesis Testing under Heterocedasticity and Non-Normality*, Oct. 2017, R package version 0.2. [Online]. Available: https://cran.r-project.org/web/packages/welchADF/index.html
[48] ——, "The welchADF package for robust hypothesis testing in unbalanced multivariate mixed models with heteroscedastic and non-normal data," *R J.*, Oct. 2017.
[49] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 354–365, Feb. 2016.
[50] ——, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.
[51] ITU-R Rec. BS.1387-1, "Method for objective measurements of perceived audio quality," International Telecommunication Union, Geneva, Switzerland, Nov. 2001.
[52] R. Huber and B. Kollmeier, "PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.
[53] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, May 2001, pp. 749–752.
[54] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time and frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.
[55] M. Holters and U. Zölzer, "GstPEAQ - an open source implementation of the PEAQ algorithm," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, vol. 4, Trondheim, Norway, Nov. 2015.
[56] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, Jun 1996.
[57] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
[58] ITU-T Rec. P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," International Telecommunication Union, Geneva, Switzerland, Nov. 2003.
[59] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, Geneva, Switzerland, Nov. 2007.
[60] G. Bernardi, T. van Waterschoot, J. Wouters, M. Hillbratt, and M. Moonen, "A PEM-based frequency-domain Kalman filter for adaptive feedback cancellation," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 270–274.
[61] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.
[62] ITU-T Rec. P.863, "Perceptual objective listening quality assessment," International Telecommunication Union, Geneva, Switzerland, Sep. 2014.
[63] ITU-R SG 12, "Statistical evaluation procedure for P.OLQA v.1.0." International Telecommunication Union, Geneva, Switzerland, Mar. 2009.



Giuliano Bernardi (S'12) was born in Asolo, Italy in 1987. He received the M.Sc. in Engineering Acoustics from Denmark Technical University (DTU), Denmark, in 2011, the MEng in Bioengineering from University of Padua, Padua, Italy, in 2012, and the Ph.D. in engineering science from KU Leuven, Belgium, in 2018.

Currently, he is a Postdoctoral Researcher at KU Leuven focusing on acoustic feedback control and noise reduction strategies for hearing-aid applications. His research interests include audio signal processing, speech perception and psychoacoustics, real-time audio signal processing, and speech enhancement.



Toon van Waterschoot (S'04, M'12) received MSc (2001) and PhD (2009) degrees in Electrical Engineering, both from KU Leuven, Belgium, where he is currently a tenure-track Assistant Professor and Consolidator Grantee of the European Research Council (ERC). He has previously also held teaching and research positions at Delft University of Technology in The Netherlands and the University of Lugano in Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

He has been serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing, and as a Guest Editor for Elsevier Signal Processing. He is a Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference in Leuven, Belgium (2016), and has been serving on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016) and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017). He is a member of EURASIP, IEEE, ASA, and AES.



Jan Wouters was born in 1960. He received the Master's and Ph.D. degrees in physics from the University of Leuven, Katholieke Universiteit(KU) Leuven, Leuven, Belgium, in 1982 and 1989, respectively, with intermission for officer military service.

From 1989 to 1992, he was a Postdoctoral Research Fellow with the Belgian National Fund for Scientific Research, Institute of Nuclear Physics, Catholic University of Louvain, Louvain-la-Neuve and at NASA Goddard Space Flight Center, USA. Since 1993, he has been a Professor at the Neurosciences Department, KU Leuven, and has been a Full Professor since 2005. His research interests include audiology and the auditory system, signal processing for cochlear implants, and hearing aids.

Dr. Wouters is the Editorial Board of the *International Journal of Audiology*, the *Journal of Communication Disorders*, and the *Journal B-ENT*. He is the President of the European Federation of Audiological Societies and the Belgian Audiological Society. He is a Member of the International Collegium for ORL (CORLAS), and a Board Member of the International Collegium for Rehabilitative Audiology.



Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received journal best paper awards from the IEEE Transactions on Signal Processing (with Geert Leus and with Daniele Giacobello) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012).

He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in IEEE Signal Processing Magazine (2012-2014), and has been a member of the editorial board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, EURASIP Journal on Applied Signal Processing and EURASIP Signal Processing.