Remixing music using source separation algorithms to improve the musical experience of cochlear implant users

Jordi Pons[a)]
Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Karl-Wiechert Allee 3, 30625, Hannover, Germany

Jordi Janer
Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra. Roc Boronat 138, 55.310, 08018 Barcelona, Spain

Thilo Rode
HoerSys GmbH, Karl-Wiechert Allee 3, 30625, Hannover, Germany

Waldo Nogueira
Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Karl-Wiechert Allee 3, 30625, Hannover, Germany

(Received 18 January 2016; revised 10 November 2016; accepted 22 November 2016; published online 19 December 2016)

Music perception remains rather poor for many Cochlear Implant (CI) users due to the users' deficient pitch perception. However, comprehensible vocals and simple music structures are well perceived by many CI users. In previous studies researchers re-mixed songs to make music more enjoyable for them, favoring the preferred music elements (vocals or beat) attenuating the others. However, mixing music requires the individually recorded tracks (multitracks) which are usually not accessible. To overcome this limitation, Source Separation (SS) techniques are proposed to estimate the multitracks. These estimated multitracks are further re-mixed to create more pleasant music for CI users. However, SS may introduce undesirable audible distortions and artifacts. Experiments conducted with CI users ($N = 9$) and normal hearing listeners ($N = 9$) show that CI users can have different mixing preferences than normal hearing listeners. Moreover, it is shown that CI users' mixing preferences are user dependent. It is also shown that SS methods can be successfully used to create preferred re-mixes although distortions and artifacts are present. Finally, CI users' preferences are used to propose a benchmark that defines the maximum acceptable levels of SS distortion and artifacts for two different mixes proposed by CI users.

[http://dx.doi.org/10.1121/1.4971424]

[JFL] Pages: 4338–4349

I. INTRODUCTION

Cochlear Implants (CIs) are surgically implanted electronic devices. By stimulating the auditory nerve with a set of electrodes it is possible to restore the sense of hearing for people with profound hearing loss. CI devices have been optimized to improve speech perception but less attention has been paid to improving music perception. This work focuses on improving CI users' musical experience since many of them cannot currently enjoy music.[1,2] The main reason for that is because melody is not well transmitted with these devices. Melody perception is severely limited mainly due to the poor pitch perception achieved with current CI devices.[3–5] However, CI users perceive rhythm and beat as well as normal hearing (NH) people.[6–8] For this reason, CI listeners prefer music with clear rhythm or beat structures. Therefore, less complex music favoring the rhythm/beat can enhance the musical experience with CIs.[9,10] Moreover, since CIs work best in quiet listening environments,[4] CI users can have difficulty tracking the lyrics of a song with accompaniment.[11] However, it has been shown that the recognition of lyrics may become easier if the accompaniment is not complex.[2] Furthermore, CI users find music composed of multiple instruments less pleasant than music played by a single instrument,[10,12] likely because pitch recognition is difficult in a polyphonic scenario.[13] In that sense, reducing music complexity seems like an appropriate way of approaching the music perception problem.[2,9,14] But not all CI users perceive music in the same way. The enjoyment of music varies a lot between implantees[5,9] and depends largely on the experience listening to music.[2,14]

Buyens *et al.*[9] proposed a method to determine enjoyable audio mixes for CI users. They showed that the CI users' music experience can be improved by reducing the complexity re-mixing musical pieces using multitrack recordings. CI listeners rated several re-mixes and it was

[a)]Also at: Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra. Roc Boronat 138, 55.308, 08018 Barcelona, Spain. Electronic mail: jordi.pons@upf.edu

shown that they preferred audio mixes with clear vocals and attenuated instruments. Furthermore, they observed that given an audio mix with clear vocals and attenuated instruments, CI users prefer beat inducers—bass and drums—to be louder than the other instruments. Kohlberg *et al.*[15] also mixed multitrack recordings to reduce their complexity, concluding that modified versions containing only 1–3 instruments were more enjoyable for CI users. But the original multitrack recordings required to create remixes such as those used in these studies are not available for most of the population. Buyens *et al.*[14,16] proposed an approach based on "harmonic percussive sound separation" (HPSS) to separate the vocals, drums, and bass components from mixed music. HPSS assumes that in a Short-Time Fourier Transform (STFT) spectrogram the harmonic components are "smooth in the temporal dimension" and the percussive components are "smooth in the frequency dimension." By adjusting the STFT-spectrogram time-frequency resolution with a long window (between 100 and 500 ms), Buyens *et al.*[17] are capable of separating vocals and drums (percussive components) from the other instruments (harmonic components). Note that the approach of Buyens *et al.*[16] separates vocals and drums, but the bass still remains in the harmonic components. Therefore, Buyens *et al.*[16] still cannot achieve a multitrack separation: vocals, drums, bass, and others. In a recent publication, Buyens *et al.*[14] propose extracting the bass by applying a low pass filter to the harmonic components. However, note that the ideal scenario would be to directly estimate the multitrack (one track per instrument) instead of first separating two audio streams: harmonic (bass and others) and percussive (vocals and drums), and then filtering out the bass from the harmonic stream. Moreover, the approach of Buyens *et al.*[14] uses a stereo binary mask to enhance the separations assuming that vocals, bass, and drums are mixed in the center of the stereo image. As a final remark, the HPSS approach used by Buyens *et al.*[14] is based on the iterative method of Ono *et al.*[17] that allows real time implementations.

We propose using Source Separation (SS) instead of HPSS-based techniques because SS can estimate one track per instrument (source) whereas HPSS cannot. SS algorithms have been used in the CI research community for noise reduction to improve speech perception[18,19] and in the audio research community to isolate audio sources.[20,21] Currently, common SS methods are based on non-negative matrix factorization (NMF)[22,23] or deep learning.[21] NMF is a consolidated state-of-the-art algorithm with a high computational cost due to its iterative nature. In contrast, deep learning is a novel and promising approach allowing faster implementations than NMF. Therefore, deep learning is more suitable for real-time implementations than NMF. However, it has to be taken into account that the estimated multitracks by any SS algorithm may introduce undesirable distortions and artifacts, and these may also depend on the content of the song. These distortions and artifacts are audible for NH listeners and limit the applicability of these algorithms for re-mixing purposes. However, given that CI users seem to better tolerate audio distortions than NH listeners,[24] because they already receive a signal with reduced spectro-temporal resolution, SS algorithms could be a promising tool for re-mixing music for this population.

This manuscript focuses on algorithms for the separation of the vocal components from the background instruments since CI users prefer an audio mix with clear vocals and attenuated instruments.[2,9] However, it has to be noted that SS algorithms are capable of extracting a track per instrument and similar studies could be done considering other music components such as beat inducers: bass,[25] drums,[20] or both together.[22]

This work investigates whether SS algorithms can be used for remixing music for CI users and whether the music mixing preferences found for CI listeners using multitrack recordings[9] are reproducible using estimated multitracks. It is necessary that the undesired distortions and artifacts introduced by these SS algorithms remain tolerable to CI users. Therefore, this work also investigates the maximum levels of acceptable SS distortion/artifacts.

II. METHODS

A. SS

This section introduces the SS algorithms used throughout this manuscript. A NMF[23] as implemented in the Flexible Audio Source Separation Toolbox (FASST)[22] and a deep recurrent neural network (DRNN).[21] The former is the baseline state-of-the-art algorithm used in the perceptual experiments. Therefore, the maximum levels of acceptable distortion and artifacts are measured considering only FASST-NMF. It is chosen for the perceptual experiments because it is a consolidated algorithm in the SS field whereas DRNN is introduced as an example of how novel SS algorithms can be used given that the maximum allowable distortion and artifacts levels are known—with a benchmark defined in Sec. III B 3.

The SS problem in signal processing is defined as the extraction of the original sources from a mixture where several sources have been combined together

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j. \tag{1}$$

Currently, most SS approaches work in the time-frequency domain: the mixture spectrogram $\mathbf{V}$ is approximated with $\hat{\mathbf{V}}$ that is formed by a superposition of J source spectrograms $\mathbf{V}_j$. Where $\mathbf{V}$, $\hat{\mathbf{V}}$, and $\mathbf{V}_j$ are of size $F \times N$, J is the number of sources (i.e., number of different instruments), and F and N are the number of frequency bins and time frames of the spectrogram, respectively. Once the source spectrograms are estimated, a common way to compute the sources is by first applying a soft mask (i.e., Wiener mask[26] or linear mask[21]) and then computing the inverse STFT of the estimated spectrogram. This enforces the constraint that the sum of the predictions is equal to the original mixture. The two SS techniques used (FASST-NMF and DRNN) rely on the previous assumptions.

1. FASST-NMF

NMF is an iterative method that estimates the sources factorizing $\mathbf{V}$ assuming that all components of the model ($\mathbf{w}_j$ and $\mathbf{h}_j$) are non-negative; then Eq. (1) extends to

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j = \sum_{j=1}^{J} \mathbf{w}_j \mathbf{h}_j^T. \quad (2)$$

It is assumed that each of the spectrograms $\mathbf{V}_j$ can be represented by the outer product of basis ($\mathbf{w}_j \in \Re^{F\times 1}$) with a time-varying gain ($\mathbf{h}_j \in \Re^{N\times 1}$).

Each source spectrogram $\mathbf{V}_j$ stands for an instrument. Source spectrograms corresponding to the singing voice and to the other instruments are constrained to $\mathbf{V}_j = \mathbf{w}_j \mathbf{g}_j \mathbf{h}_j$ where $\mathbf{w}_j$ is a fixed dictionary representing the narrowband spectral patterns modeling the harmonic pitch, $\mathbf{g}_j$ represents the $\mathbf{w}_j$ gains, and $\mathbf{h}_j$ describes the time activations that display when each source is activated. The bass and drums source spectrograms are set such that $\mathbf{V}_j = \mathbf{w}_j \mathbf{h}_j$, where $\mathbf{w}_j$ is fixed and pre-trained from isolated bass and drum samples from the Real World Computing Music Database[27] and $\mathbf{h}_j$ describes the time activations.

The factorizations are obtained over an Equivalent Rectangular Bandwidth (ERB)[28] spectrogram, considering all $\mathbf{V}_j$ models together and with random initialization to all the adaptive parameters. The model parameters are estimated from the mixture $\mathbf{V}$ using the iterative Generalized Expectation Maximization[29] algorithm. The vocals are reconstructed by choosing the most energetic source of the ones modeling the other instruments—this assumption holds for all factorized songs in that study—and the remaining sources modeling other instruments are summed up and represent the background instruments. The window for the initial STFT analysis (before the ERB mapping) is set to be a raised sine of size 2048 samples with a 50% overlap and fast Fourier transform (FFT) size of 2048 samples. This implementation is available online[30] and Ozerov *et al.*[22] extensively explain the details of it.

2. DRNN

DRNN is a biologically inspired machine learning scheme of interconnected neurons used to approximate an unknown function from data. It is formalized as follows:

$$\mathbf{y}_{(t)} = \mathbf{W}^{(L-1)} \mathbf{h}_{(t)}^{(L-2)} + \mathbf{b}^{(L-1)}, \quad (3)$$

$$\mathbf{h}_{(t)}^{(l)} = \sigma^{(l)}(\mathbf{W}^{(l)} \mathbf{h}_{(t)}^{(l-1)} + \mathbf{W}_{rec}^{(l)} \mathbf{h}_{(t-1)}^{(l)} + \mathbf{b}^{(l)}), \quad (4)$$

$$\mathbf{h}_{(t)}^{(1)} = \sigma^{(1)}(\mathbf{W}^{(1)} \mathbf{x}_{(t)} + \mathbf{W}_{rec}^{(1)} \mathbf{h}_{(t-1)}^{(1)} + \mathbf{b}^{(1)}), \quad (5)$$

where the weight matrices are denoted by $\mathbf{W}^{(l)} \in \Re^{d_{(l)}\times d_{(l-1)}}$ for feed-forward connections and $\mathbf{W}_{rec}^{(l)} \in \Re^{d_{(l)}\times d_{(l)}}$ for the recurrent connections. The bias vector is $\mathbf{b}^{(l)} \in \Re^{d_{(l)}\times 1}$, the input matrix is $\mathbf{x}_{(t)} \in \Re^{d_{\text{input}}\times 1}$, the hidden activations are denoted as $\mathbf{h}_{(t)}^{(l)} \in \Re^{d_{(l)}\times 1}$, and the output activations as $\mathbf{y}_{(t)} \in \Re^{d_{(L-1)}\times 1}$. l represents the index of the current layer, ranging from $1 \le l \le (L-1)$ and t denotes the current timeframe. Each neuron is characterized by a nonlinear activation function $\sigma^{(l)}$ that can be layer specific. The $\sigma^{(l)}$ used here is the rectifier linear unit (ReLU)

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

A DRNN model was successfully trained by Huang *et al.*[21] to separate the vocals from the background instruments of a mono signal. This implementation is available online[31] and extensively explained by Huang *et al.*[21] This deep recurrent architecture is based on three hidden layers where the second is recurrent, each hidden layer has 1000 units characterized by ReLUs, and the output layer is a linear layer modeling two sources. Moreover, they propose jointly optimizing the net with a soft time-frequency mask attached after the output layer. A context window of three frames ($|\mathbf{V}_{t-1}|$, $|\mathbf{V}_t|$, and $|\mathbf{V}_{t+1}|$) of the magnitude spectra composes the input vector (then, $\mathbf{x}_t \in \Re^{(F\cdot 3)\times 1}$) of the net. A raised sine of 1024 samples is used as the analysis window with a FFT size of 1024 samples and 50% overlap. During this study, the DRNN model was trained using data provided by the Signal Separation Evaluation Campaign 2015 (SISEC-2015).[32] This data, pop western music, have more resemblance to the audio material used for the perceptual experiments, pop western music as well, than the original Chinese karaoke files used by Huang *et al.*[21] Training with the SISEC-2015 dataset is the main modification to the original DRNN model proposed by Huang *et al.*[21]

The model is optimized by back-propagating the gradients through time with respect to the training objectives. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno[33] algorithm is used to train the models from random initialization considering a discriminative cost function[21] together with the Mean Squared Error (MSE) cost function. Such discriminative cost decreases the similarity between the prediction and the targets of other sources while the MSE cost increases the similarity between the target and the prediction of the same source.

3. Evaluation

In the SS research community the separation errors are typically[21,22,25,26] quantified by the Source-to-Distortion Ratio (SDR), the Source-to-Artifacts Ratio (SAR), and the Source-to-Interference Ratio (SIR).[34] However, SIR values are less relevant for this study because when mixing estimated tracks, the interferences are mutually combined and are very difficult to perceive. For this reason, SIR values are omitted in this study. Higher SDR/SAR values represent a better separation quality. Therefore, the maximum levels of distortion/artifacts correspond to the lower bounds of these measures. Computing SDR and SAR values involves two successive steps.[34] In a first step, the estimated source, $\hat{s}_{\text{target}}(n)$, is decomposed

$$\hat{s}_{\text{target}}(n) = s_{\text{target}}(n) + e_{\text{interf}}(n) + e_{\text{noise}}(n) + e_{\text{artif}}(n),$$

where $e_{\text{interf}}(n)$, $e_{\text{noise}}(n)$, $e_{\text{artif}}(n)$ are the interferences, noise, and artifacts error terms, respectively. These error terms are computed by using the BSS Eval Toolbox,[35] an implementation provided by the original authors of these

TABLE I. Demographic information for the CI subjects who took part in the perceptual experiments. Five post-lingual CI subjects (S1–S5) participated in experiment 1 and nine post-lingual CI subjects (S1, S3, and S6–12) participated in experiment 2.

Subject	Age	Gender	CI experience (years)	Etiology	Sound processor	Implant type	Brand
S1	79	Male	13	Sudden hearing loss	CP810	Nucleus CI24R (CS) (Left)	Cochlear
S2	62	Male	8	Ostosclerosis	Freedom	Nucleus CI24RE (CS) (Left)	Cochlear
S3	66	Male	7	Genetics	Opus2	Sonata ti100 (Left)	MEDEL
S4	54	Male	3	Unknown	Opus2	ConcertoFlex EAS 24 (Left)	MEDEL
S5	50	Male	14	Sudden hearing loss	Harmony	Clarion CII (Left)	AB
S6	66	Male	2	Sudden hearing loss	CP810	Nucleus CI24RE (CA) (Left)	Cochlear
S7	73	Male	4	Unknown	CP810	CI512 (Left)	Cochlear
S8	60	Male	6	Unknown	Naida	HIRES90k (Left)	AB
S9	56	Female	7	Sudden hearing loss	CP910	Nucleus CI24RE (CA) (Left)	Cochlear
S10	47	Female	19	Ototoxic	Harmony	Clarion (Right)	AB
S11	67	Female	4	Sudden hearing loss	CP910	Nucleus CI24RE (CA) (Left)	Cochlear
S12	67	Female	6	Sudden hearing loss	Naida	HIRES90k (Right)	AB

measurements. In a second step, SDR and SAR are computed as follows:

$$\mathrm{SDR} = 10\log_{10}\frac{\|s_{\mathrm{target}}(n)\|^2}{\|e_{\mathrm{interf}}(n)+e_{\mathrm{noise}}(n)+e_{\mathrm{artif}}(n)\|^2},$$

$$\mathrm{SAR} = 10\log_{10}\frac{\|s_{\mathrm{target}}(n)+e_{\mathrm{interf}}(n)+e_{\mathrm{noise}}(n)\|^2}{\|e_{\mathrm{artif}}(n)\|^2}.$$

SDR and SAR are given in dBs. The global SDR and SAR measures are defined as their mean value weighted by the length of each song.

B. Perceptual experiments

Two experiments were designed. The first one explored which mixing pre-sets are preferred by CI users; original and estimated multitracks were used for this purpose. In the second experiment, additional music pieces were re-mixed using pre-sets based on the results of the first experiment. CI users were asked, through a pairwise comparison test, about their preference for the different pre-sets using original and estimated multitracks. The methodology described in the following is based on the one presented by Buyens *et al.*[9]—considered the base of this work. By doing so, and given that we use the same audio material, the results presented below are directly comparable to the ones published in Buyens *et al.*[9] Therefore, common methodologies—such as MUltiple Stimuli with Hidden Reference and Anchor[36] (MUSHRA) or Mean Opinion Score[37] (MOS)—are not considered. However, and similarly as in MUSHRA standard, the following perceptual experiments also include anchor mixes. The purpose of the anchor is to make the results more interpretable making sure that minor distortion/artifacts are not rated as having very bad quality.

Five post-lingually deaf CI users (S1–S5) participated in experiment 1. Nine post-lingually deaf CI users (S1, S3, and S6–12) and nine NH subjects (NH1–NH9) participated in experiment 2. CI subjects were native German speakers and had at least 1 year of experience with their CI. Subjects obtaining more than 25% speech intelligibility in noise using the HSM sentence test[38] (signal-to-noise ratio = 10 dB) and having interest in music were selected for the study. Table I shows the CI subjects' details. Six out of 9 NH subjects were males with an average age of 33 yrs old (ranging from 25 to 41). The study participants signed a consent form and were paid for their travel expenses. Ethical approval from the university's local ethics committee was obtained. Subjects had no time constraints for conducting each experiment and could take a break whenever it was necessary. Both experiments used western popular music presented at approximately 60 dB SPL (sound pressure level) through a single loudspeaker (Genelec 8240 APM, Iisalmi, Finland) at a 0° azimuth/elevation angle inside an acoustically treated room.

1. Experiment 1

CI subjects were asked to determine the most enjoyable audio mix using a mixing console that played an infinite loop of the first multitrack presented in Table II. This piece was selected as it is the only one having all accompanying instruments. Each multitrack was presented 10 times in random order in its original and estimated form, giving a total of 20 tests per subject. The mixing console was a web app[39] based on the MT5 multitrack player.[40] It is composed by two sliders that control the gain applied to each track of the song: vocals and instruments. The gain assigned to the sliders ranged from −70 to 0 dB. All sliders were initialized to the lowest value to avoid suggesting *a priori* mixing preferences. Furthermore, a random offset gain ranging from −15 to

TABLE II. Multitrack recordings used for the study. For experiment 1 only the *Dock* multitrack was used. The recordings were kindly provided by Buyens *et al.* (Ref. 9). The + and − signs indicate whether that multitrack does or does not contain the respective instrument. (V-vocals, P-piano, G-guitar, B-bass, D-drums).

ID	Song	V	P	G	B	D
Dock	*Dock of the Bay (Otis Redding)*	+	+	+	+	+
Bef	*Before I Go (Papermouth)*	+	−	+	−	−
Hall	*Hallelujah (Leonard Cohen)*	+	+	−	−	−
Jud2	*Hey Jude (excerpt A) (The Beatles)*	+	−	+	+	−
Jude	*Hey Jude (excerpt B) (The Beatles)*	+	−	+	+	+
Mic2	*Michel (Anouk)*	+	−	+	+	−

0 dB was added independently to each slider track to ensure that subjects did not use visual cues. CI subjects were asked to adjust the sliders to the levels that produced the most enjoyable mix. Before starting the test, subjects performed several training trials until they got familiar with the interface. Experiment 1 allowed CI subjects to propose the mixing pre-sets that they enjoyed most and also allowed the observation of the potential differences in mixing preferences between mixes created using original and estimated multitracks. Before starting the test, the study participants were asked about their level of knowledge on popular modern music on a scale from 1 to 10 (where 10 stands for a high level of musical knowledge) and about their general music enjoyment.

The differences in level settings between tracks were analyzed with the Wilcoxon signed-ranks test. For each multitrack (original or estimated) 50 experiment repetitions (5 subjects $\times$ 10 tests) are available. The Wilcoxon signed-ranks test handles the repetitions by comparing whether the mean of the vocals levels ranks differently from the mean of the instruments levels.

2. Experiment 2

Based on the results from experiment 1 and previous work,[9] different mixing pre-set pairs were defined (see Table IV and the discussion in Sec. IV B 1). The *Standard* pre-set stands for a NH listener mix, with vocals and instruments presented at the same level. The −6 and −12 dBMT pre-sets denote re-mixes obtained from the original multitracks where the vocals are favored attenuating the instruments by 6 and 12 dB, respectively. The −6 and −12 dBSS are defined similarly but using the estimated multitracks from SS.

The experiment was designed (1) to assess whether the SS distortions and artifacts are tolerated by CI subjects or not (comparing *mixSS* with *mixMT* preference percentages), (2) to determine which of the mixing pre-sets presented in Table III are more enjoyable for CI subjects, and (3) to study whether the preference for the pre-sets proposed for CI users change when using SS or not (comparing *mixSS* vs *Standard* preference percentage with *mixMT* vs *Standard* preference percentage).

A pairwise comparison test was used for studies (1), (2), and (3) in experiment 2. Each subject was asked to choose the most enjoyable sound (A or B). When both sounds were equally enjoyable, participants were asked to find any detail that would allow them to decide whether they prefer A or B. If they were not capable of finding it, they were asked to guess. A and B were the same song but with different mixing configurations and the order of the mix assigned to song A or B was randomized, see Table IV. The estimated multitracks were obtained using the FASST-NMF algorithm. Pairs were randomly presented 3 times for each song listed in Table II, giving a total of 108 pair comparisons (6 pairs $\times$ 3 repetitions $\times$ 6 songs). The preference score between mix A and B is indicated in percentages calculated dividing the number of times A or B was preferred by the total number of comparisons for the same pair, A and B are therefore complementary. Before starting the test, CI subjects were asked if they regularly listen to popular western music and about their general music enjoyment. NH subjects were not asked since it is assumed that all NH listeners regularly listen to music because they have no difficulties in listening to it.

For study (1) and (2) the chi-square (χ^2) test with Bonferroni correction was used to determine which of the mixing pre-sets of the pair was preferred. It was applied over the preference percentage score with the null hypothesis (H_0) set such that preference scores were the same for both pre-sets—A and B: 50%–50%. H_0 was set to random guess because the study participants were also instructed to guess in case that they could not perceive any difference between A and B. A significant p-value means that the distributions differed and that a significant preference existed for one of the pre-sets.

For study (3) the χ^2 test was used to determine whether the preference for the pre-sets proposed for CI users changed when using SS or not. The χ^2 test was applied as a goodness of fit test where the H_0 was set to be the preference percentage distribution of *mixMT* vs *Standard*. The test measured how well the preference percentage distribution of *mixSS* vs *Standard* fitted the H_0 distribution. Therefore, *mixSS* vs *Standard* preference percentage was compared to *mixMT* vs *Standard* preference percentage. A significant result means that the distributions differ and that the mixing preferences have changed. This test provides a measure of how much the distortion and artifacts in the estimated SS influenced the user mixing preferences.

C. Audio material

All the multitracks used during the present study (Table II) were kindly provided by Buyens *et al.*[9] who normalized each track to have the same loudness. The ReplayGain[41] standard was used as the loudness measure. Note that by using the same multitracks as Buyens *et al.*,[9] the results

TABLE III. Mixing pre-sets with different instruments-to-vocals ratios used in experiment 2. The *Standard* pre-set is proposed for NH listeners and the −6 and −12 dB pre-sets are proposed for CI users.

Mixing pre-set	Vocals	Instruments
Standard	0 dB	0 dB
−6 dB	0 dB	−6 dB
−12 dB	0 dB	−12 dB

TABLE IV. Pairs of multitrack recordings used in experiment 2. A and B were the same song but with different mixing configurations. The estimated and the original multitrack are denoted here as *mix*SS and *mix*MT, respectively. Mixing pre-sets are defined in Table III.

Song A	Song B
−6 dBSS	−6 dBMT
−12 dBMT	−12 dBSS
Standard	−6 dBSS
−6 dBMT	Standard
Standard	−12 dBSS
−12 dBMT	Standard

presented here are directly comparable to the ones included in their study. They assumed that the summation of the normalized tracks, having the same loudness level, stands for a NH listener studio mix.[9] Throughout this manuscript we refer to this mix as *Standard*.

The monaural input mixtures for the DRNN algorithm were *Standard* mixes. But FASST-NMF can also exploit the panning information from stereo audio files. In order to allow that, the following stereo panning is used for recordings with drums (*Dock* and *Jude*, see Table II):

$$\mathbf{L}: 0.75 \cdot \text{Vocals} + 0.25 \cdot (\text{Piano} + \text{Guitar}) + 0.5 \cdot \text{Drums} + 0.5 \cdot \text{Bass},$$

$$\mathbf{R}: 0.25 \cdot \text{Vocals} + 0.75 \cdot (\text{Piano} + \text{Guitar}) + 0.5 \cdot \text{Drums} + 0.5 \cdot \text{Bass},$$

but for the remaining recordings the following stereo mix is used:

$$\mathbf{L}: 0.75 \cdot \text{Vocals} + 0.25 \cdot (\text{Guitar} + \text{Bass} + \text{Piano}),$$

$$\mathbf{R}: 0.25 \cdot \text{Vocals} + 0.75 \cdot (\text{Guitar} + \text{Bass} + \text{Piano}),$$

where the sum of the channels still provides a *Standard* mix. Note that for some multitrack recordings, not all the tracks described in the above stereo mix setup are present (e.g., *Bef* and *Hall* multitrack recordings have only two tracks). In these cases, the remaining tracks are simply omitted. For example, the stereo mix for the *Bef* multitrack recording is

$$\mathbf{L}: 0.75 \cdot \text{Vocals} + 0.25 \cdot \text{Guitar},$$
$$\mathbf{R}: 0.25 \cdot \text{Vocals} + 0.75 \cdot \text{Guitar}.$$

Note that in these cases where only two tracks are available, panning the vocals to the left allows having a stereo mix—otherwise the mix will be mono. The stereo panning mix was introduced to improve the performance of the FASST-NMF separation. Our goal was to obtain good examples of separation that are valid to assess our hypothesis in the perceptual experiments. In order to keep some coherence among stereo mixes, we decided to always pan the vocals to the left although this is an atypical panning setup—because in commercial releases the vocals are normally placed in the center of the stereo image. Note that this stereo setup allows a consistent comparison among FASST-NMF separations although the FASST-NMF algorithm has an advantage when compared to DRNN since it also uses stereo information. However, in principle, any other panning could be used without decreasing the quality of the separations since FASST-NMF would still be able to exploit the panning information.

In order to allow a fair comparison among the resulting gains of the mixing console in experiment 1, each estimated track was normalized to have the same loudness as its corresponding original track. But for experiment 2, the different loudness levels between presented songs were compensated by applying an overall gain, keeping the mix intact. When setting the estimated tracks to be equally loud as the original track, subjects are not influenced by any loudness difference and the preference decisions are based on the tolerance level that subjects have with respect to SS artifacts and distortions. However, note that the error estimations could be amplified by applying a positive gain. For this reason, negative gains were applied whenever was possible and it was assumed that the estimated track had a similar loudness as the original track. Therefore, the overall gain discussed here is not expected to produce a large effect.

TABLE V. Distortion/artifacts measures (SDR/SAR) used to assess the quality of the FASST-NMF algorithm for each of the songs used in experiment 2. A global SDR/SAR measure, computed as the mean weighted by the length of each song, is also provided. The time required to run and evaluate the algorithm is also included.

(dB)	Target (vocals)		Other (background)	
Song	SDR	SAR	SDR	SAR
Dock	3.713	4.054	11.619	15.585
Bef	3.966	4.031	4.662	12.287
Hall	11.977	12.043	10.575	15.142
Jud2	3.669	3.675	7.359	14.768
Jude	6.906	7.196	14.662	18.013
Mic2	9.880	9.902	11.320	16.463
Global	5.991	6.121	9.928	15.433
Time	4 h 39 min (SS) + 2 min (evaluation)			

III. RESULTS

A. SS

Results are given for FASST-NMF (Table V) and DRNN (Table VI) considering the same dataset used in experiment 2 (Table II) and the SDR and SAR measures introduced in Sec. II A 3.

DRNN is a powerful model to separate vocals from background instruments since similar global SDR and superior SAR values were obtained compared to FASST-NMF. The DRNN algorithm can reduce the FASST-NMF processing time, DRNN is approximately 70 times faster. All runs were computed with a regular laptop: i5 with 4Gb-RAM. Finally, one can observe that DRNN drops significantly its SDR and SAR performance for the *Bef* and *Hall* songs.

TABLE VI. Distortion/artifacts measures (SDR/SAR) used to assess the quality of the DRNN algorithm for each of the songs used in experiment 2. A global SDR/SAR measure, computed as the mean weighted by the length of each song, is also provided. The time required to run and evaluate the algorithm is also included.

dB	Target (vocals)		Other (background)	
Song	SDR	SAR	SDR	SAR
Dock	4.941	10.219	4.719	9.170
Bef	0.727	7.212	−0.015	8.444
Hall	0.580	16.197	2.031	6.755
Jud2	6.839	8.913	5.687	10.790
Jude	9.904	11.779	7.927	13.103
Mic2	6.639	12.715	9.698	10.436
Global	5.574	10.699	5.536	10.078
Time	1 min 49 sec (SS) + 2 min (evaluation)			

B. Perceptual experiments

1. Experiment 1

Subjects were asked to determine their own level of knowledge about western popular modern music (from 1 to 10): S1–3, S2–4, S3–8, S4–5, and S5–7; and all subjects participating in experiment 1 reported to enjoy music in general.

The acquired data were averaged for all CI subjects. For the original multitrack, the preferred instruments level was set significantly lower than the vocals level (Wilcoxon signed-ranks test: p-value $= 0.003$) with a mean instruments-to-vocals ratio of -1.92 dB. For the estimated multitrack with FASST-NMF, the instruments level was set significantly lower than the vocals level (Wilcoxon signed-ranks test: p-value < 0.001) with an instruments-to-vocals ratio mean value of -3.82 dB.

Even though the scope of experiment 1 is to analyze the overall mixing preferences of CI subjects, individual results are also presented (Fig. 1) to illustrate the high variability obtained among subjects. For example, S4 (the less experienced subject with a music knowledge level of five) is the only one having a mean instruments-to-vocals ratio above zero (for the original multitrack). But also a high within subjects variability is observed. For example, subjects having less knowledge about popular western music (S1 and S2) showed a larger variability in their results.

2. Experiment 2

Table VII shows the averaged results for CI subjects and NH subjects, separately. NH subjects significantly preferred mixes created from the original multitrack over mixes created from the estimated. This could be due to the presence of SS distortion/artifacts in the estimated multitracks. NH subjects significantly preferred the *Standard* pre-set over any other.

In the following, the results for studies (1), (2), and (3) that conform experiment 2 are presented.

In Table VII are presented the results for study (1), showing that CI subjects significantly preferred the -12 dBMT over the -12 dBSS pre-set. However, no significant preference was observed when comparing the -6 dBMT and -6 dBSS pre-sets. This could be because the presence of SS distortion/artifacts in vocals were not tolerated for the -12 dB mix but these were tolerable for the -6 dB mix.

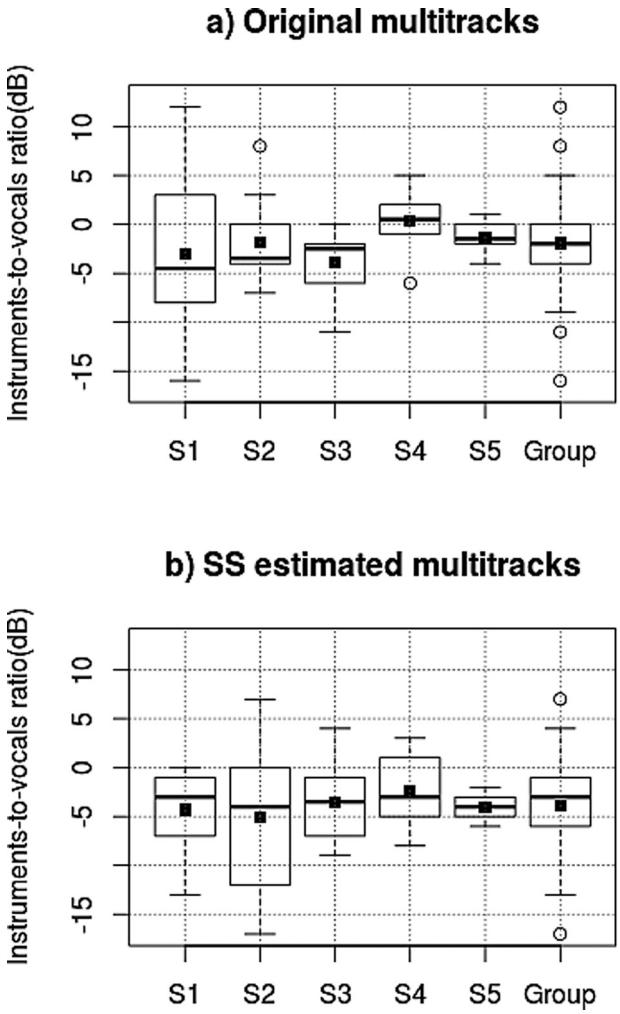

FIG. 1. Results of experiment 1 showing the instruments level relative to vocals (dB) for each subject, in columns. The statistics over the ten repetitions for each subject are summarized with a box-plot where: unfilled circles represent outliers, thick black lines stand for median values, filled squares appear for mean values, the bottom and top of the box are the first and third quartiles, and whiskers represent the maximum 1.5 interquartile range. The last column, *Group*, merges the results for all subjects. The first and second rows depict the instruments level relative to vocals when using the original and the estimated multitracks, respectively.

Individual results for each subject are also presented in Table VIII and provide material for discussing study (2). All CI subjects participating in experiment 2 reported to enjoy music except S6 and S11. In general, a large variability in mixing preferences across CI subjects can be observed. For example, S8 and S3 significantly preferred the *Standard* over

TABLE VII. Experiment 2 showing the preference percentage of each pair for CI and NH subjects. Significant results are highlighted in bold (χ^2 test with H_0 being the equal preference for both conditions, 50%–50%; $\alpha = 0.05$).

A vs B	CI ratings	χ^2 test p-value	NH ratings	χ^2 test p-value
-6 dBSS vs -6 dBMT	46.92%–53.08%	0.4321	32.10%–**67.90%**	**<0.001**
-12 dBSS vs -12 dBMT	40.12%–**59.88%**	**0.0119**	24.08%–**75.92%**	**<0.001**
-6 dBSS vs Standard	54.32%–45.68%	0.2714	25.92%–**74.08%**	**<0.001**
-6 dBMT vs Standard	52.47%–47.53%	0.5297	29.02%–**70.98%**	**<0.001**
-12 dBSS vs Standard	42.60%–57.40%	0.0593	8.03%–**91.97%**	**<0.001**
-12 dBMT vs Standard	52.47%–47.53%	0.5297	12.96%–**87.04%**	**<0.001**

TABLE VIII. Results showing the preference percentage of each CI subject for the pre-sets pairs of experiment 2. Significant results are highlighted in bold (χ^2 test with H_0 being the equal preference for both conditions, 50%–50%; with Bonferroni correction: $m = 9$, $\alpha = 0.0055$). The pop music experience column shows whether a subject regularly listens to popular western music or not.

Subject	Pop music experience	−6 dBSS vs −6 dBMT	−12 dBSS vs −12 dBMT	−6 dBSS vs Standard	−6 dBMT vs Standard	−12 dBSS vs Standard	−12 dBMT vs Standard
S1	NO	77.78%–22.22%	33.33%–66.67%	72.2%–27.78%	**94.44**%–55.56%	72.22%–27.78%	**100**%–0%
S3	YES	50.00%–50.00%	44.44%–55.56%	0%–**100**%	5.56%–**94.44**%	0%–**100**%	0%–**100**%
S6	NO	50.00%–50.00%	27.78%–72.22%	72.22%–27.78%	66.67%–33.33%	61.11%–38.89%	61.11%–38.89%
S7	NO	38.89%–61.11%	55.56%–44.44%	66.67%–33.33%	**88.89**%–11.11%	77.78%–22.22%	77.78%–22.22%
S8	YES	38.89%–61.11%	44.44%–55.56%	11.11%–**88.89**%	0%–**100**%	0%–**100**%	0%–**100**%
S9	YES	27.78%–72.22%	44.44%–55.56%	77.78%–22.22%	38.89%–61.11%	22.22%–77.78%	44.44%–55.56%
S10	YES	50.00%–50.00%	16.67%–**83.33**%	**94.44**%–5.56%	**88.89**%–11.11%	**94.44**%–5.56%	**94.44**%–5.56%
S11	NO	44.44%–55.56%	50.00%–50.00%	38.89%–61.11%	27.78%–72.22%	5.56%–**94.44**%	22.22%–77.78%
S12	YES	44.44%–55.56%	44.44%–55.56%	55.56%–44.44%	61.11%–38.89%	50.00%–50.00%	72.22%–27.78%

all other pre-sets. However, S10 significantly preferred the −6 dB and the −12 dB over the *Standard* pre-set. Also note that four subjects (S6, S9, S11, and S12) showed no significant preference for any of the pre-sets proposed to be more suitable for CI users. If these four subjects are added to the two who significantly preferred the *Standard* mixes over all others (S3 and S8), six out of the nine subjects did not significantly prefer any of the pre-sets proposed to be more suitable for CI users. Individual results also show that only S10 significantly preferred the −12 dBMT pre-set over the −12 dBSS, denoting that some CI subjects can tolerate the SS distortion/artifacts when attenuating the background instruments by 12 dB.

CI subject preferences for the −6 dB pre-set were preserved although the original or the estimated multitracks were used, χ^2 goodness of fit test between −6 dBSS vs Standard (54.32%–45.68%) and −6 dBMT vs Standard (52.47%–47.53%): p-value = 0.6369. However, the preferences significantly changed for the −12 dB mixing pre-set, χ^2 goodness of fit test between −12 dBSS vs Standard (42.60%–57.40%) and −12 dBMT vs Standard (52.47%–47.53%): p-value = 0.011. There is a preference for the −12 dBMT mix over the Standard that is inverted when using −12 dBSS. However, the preference for the −6 dB pre-set is preserved although the original or estimated multitracks are used. These results correspond to study (3) and point out that the mixing preferences changed when using SS for the −12 dB mix. This can be explained by the fact that the SS distortion/artifacts present in vocals were not tolerated for the −12 dB mix but these were tolerable for the −6 dB mix.

3. Defining a benchmark: Rating distortion and artifacts levels

After experiment 2, the objective SDR and SAR measures as well as the subjective preference scores are available. The relation between the objective and subjective measures has been analyzed to study which SDR and SAR levels cause a change in preference from estimated to original multitracks. These may be different for the −6 and −12 dB mixes and are treated separately. The SDR and SAR values are given for the vocals source because for the −6 and −12 dB mixing pre-sets this is the most prominent source, where the distortion and artifacts may be noticeable. Note that for the −12 dB mix, the anchor pre-set, the vocals distortions, and artifacts may be more noticeable since the background is more attenuated.

The resulting data have been analyzed using two models, the linear regression model and the minimum model. One could argue that there is not enough data for fitting the models to the observations. However, simple, easy to interpret, and complementary models were chosen given that the data were scarce. Rather than using a subjective evaluation based on the MUSHRA or MOS tests, a new methodology that links two *de facto* standards is proposed. This novel methodology connects Buyens *et al.*[9] methods and results with the error measures broadly used in the SS research community. With this new methodology, the suitability for remixing music to improve CI users' musical experience of novel SS algorithms can be directly considered.

Results and models are summarized in Fig. 2, where the gray highlighted area (non-preference area, from 29.63% to 70.37%) corresponds to the ratings having a p-value >0.05 of a χ^2 test (with H_0 being the equal preference for both conditions: 50%–50%). Note that the ratings in Fig. 2 never exceed the non-significant upper bound of 70.37%, since this would mean that there is a significant preference for the estimated multitrack over the original multitrack. However, exceeding the lower bound means that the original multitrack is preferred over the estimated multitrack. Higher values of SDR and SAR represent a better separation quality and therefore, defining the maximum allowed distortion/artifacts levels is equivalent to defining the minimum SDR/SAR values.

The linear regression model fits a regression line between the preference ratings for every mix and the SAR and SDR values for the six songs in Table II. The intersection between the linear regression (dark line in Fig. 2) and the lower bound of the non-preference area (left dashed vertical line) gives an approximation of the minimum SDR and SAR values recommended to create re-mixes for CI users using SS. The minimum recommended values for the SDR and SAR measures in the −6 dB mix are 4.9 dB for both measures. For the −12 dB mix the minimum recommended

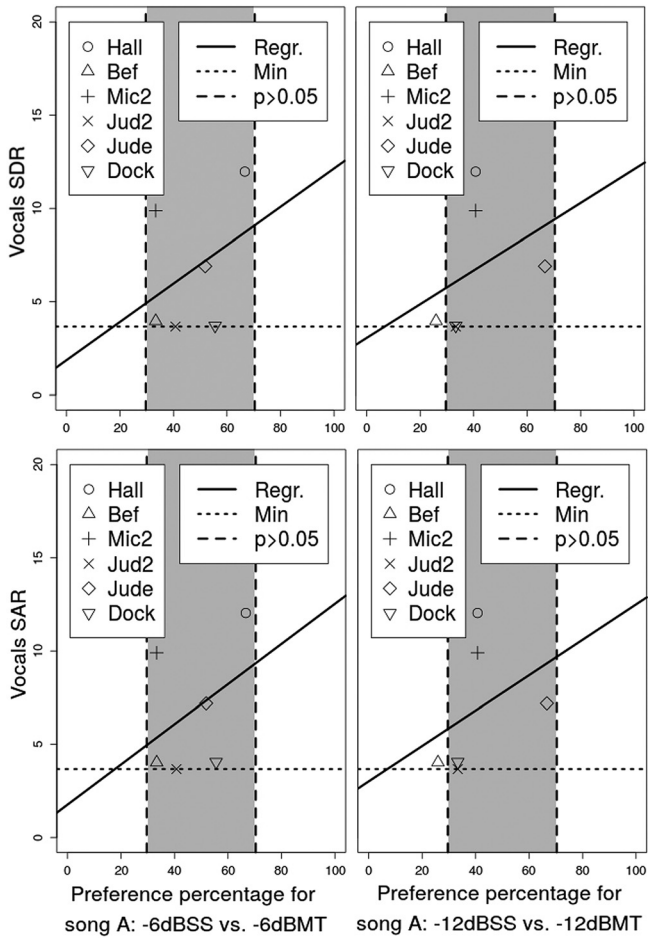

FIG. 2. Results of experiment 2 relating the error levels obtained with the FASST-NMF (SDR and SAR, in rows) with the preference ratings for song A are presented. Song A always represents the *mixSS* (the *mix* created using the estimated multitracks) and song B represents *mixMT* (the *mix* created using the original multitrack). Only the A preference percentages are presented because A and B preference percentages are complementary. Two different mixes are considered: −6 and −12 dB (first and second columns, respectively). Each song is represented by a different symbol. The non-preference area (gray highlighted area from 29.63% to 70.37%) corresponds to the ratings having a p-value >0.05 of a χ^2 test with H_0 being the equal preference for both conditions (50%–50%), the dark line represents the linear regression model, and the dotted line stands for the minimum model defined measuring the minimum SDR/SAR value inside the non-preference area.

values for the SAR and SDR values are 5.7 and 5.8 dB, respectively. A significance test for each linear regression model was performed in order to quantify how well the linear fit represents the data: −6 dB mix with SDR model—p-value = 0.4501 (top left in Fig. 2); −12 dB mix with SDR model—p-value = 0.4852 (top right in Fig. 2); −6 dB mix with SAR model—p-value = 0.421 (bottom left in Fig. 2) and −12 dB mix with SAR model—p-value = 0.4571 (bottom right in Fig. 2). It can be observed that there is no significant correlation between the linear regression models and the data, probably due to the small amount of data available. For this reason, a complementary model is added for defining the benchmark.

The minimum model provides a lower bound value that is recommended not to be exceeded. It is based on measuring the minimum SDR and SAR values inside the non-preference area. The minimum model is represented with a dotted line in Fig. 2. The minimum values for both pre-sets (−6 and −12 dB) correspond to the same song (*Jud2*) and are the same for both measures (SDR and SAR): 3.6 dB.

Given that a small sample of data are considered, both models define a benchmark that provides some intuition about the maximum levels of acceptable distortion and artifacts for each mix (−6 dB and −12 dB).

IV. DISCUSSION

A. SS

Tables V and VI show the levels of distortion and artifacts for FASST-NMF and DRNN, respectively. DRNN is a powerful model to separate vocals from background instruments since similar global SDR and superior SAR values were obtained compared to FASST-NMF. This result is notorious given that DRNN is using monaural signals while FASST-NMF is exploiting stereo information to enhance the separations. It is worth noting that DRNNs achieve a drastic reduction in computational time which is required to achieve real-time implementations for commercial behind-the-ear CI processors. But when comparing the results for each song individually, one can observe that DRNNs achieve a lower performance for the *Bef* and *Hall* songs while the FASST-NMF results are less variable across songs. This observation may be explained by the fact that the *Bef* and *Hall* songs contain two sources while the SISEC-2015 dataset, used for training the DRNNs, contains songs with a full instrumental band playing. These results show the importance of the training database and the high risk of committing over-fitting with data-driven approaches like deep learning. Further experiments with larger training datasets together with the use of techniques that help to prevent over-fitting[42] may help to overcome this DRNN limitation.

B. Perceptual experiments

1. Experiment 1

Results show an instruments-to-vocals ratio mean value of −1.92 dB (p-value = 0.003) when the original multitracks are used and an instruments-to-vocals ratio mean value of −3.82 dB (p-value <0.001) when the estimated multitracks are used. These results are in agreement with the study of Buyens *et al.*,[9] where CI subjects preferred the vocal level set significantly higher than the instruments level with an instruments-to-vocals ratio ranging from −3 to −15 dB. From these results one can conclude that whether using estimated or original multitracks, CI subjects significantly prefer the vocals to be louder than the instruments.

However, individual results show a large variability (Fig. 1—last column: *Group*). When using original multitracks the instruments-to-vocals ratio ranges from −16 to 12 dB, while for the estimated multitracks it ranges from −17 to 7 dB. Such variability could be partially explained by the large dynamic range of the linear sliders, 70 dB. Constraining the search space may help to obtain less variable data. Moreover, in Fig. 1 one can also observe that

results are user dependent. For example, one can observe that the results for S1 and S2 show large variability. However, this variability is smaller than half of the dynamic range assigned to the sliders, indicating that their responses were probably not random.

From the above results and previous work,[9] three mixing pre-sets were designed. The *Standard* pre-set stands for a NH listener mix, with vocals and instruments presented at the same level. The other two pre-sets were configured to be more suitable for CI users with an instruments-to-vocals ratio of −6 and −12 dB. These were selected considering the large variability among individual results (instruments-to-vocals ratio between −17 and 12 dB), to have vocals louder than instruments (instruments-to-vocals ratio <0) and to be consistent with the previous work[9] where the −6 and −12 dB pre-sets were used. Note that for the −12 dB mix the distortions and artifacts in the vocals may be more noticeable since the background is more attenuated. Therefore, the −12 dB pre-set acts as an anchor mix (where the SS distortion/artifacts present in the estimated vocals can be perceived at least for NH) in order to get more interpretable results—similarly as in MUSHRA[36] standard. The mixing pre-sets are summarized in Table III.

2. Experiment 2

Results show that NH listeners have a preference for the *Standard* pre-set and for mixes created from the original rather than the estimated multitracks (Table VII). This result confirms that mixes generated from original multitracks are more enjoyable for NH listeners than SS mixes, revealing that NH subjects do not tolerate SS distortion and artifacts. Furthermore, the assumption that the *Standard* pre-set is most enjoyable for NH listeners is confirmed because these listeners significantly preferred the *Standard* pre-set over any other. Since NH listeners have no difficulties on listening to music and are normally exposed to commercial music with a similar balance between vocals and instruments, NH listeners were more familiar with the *Standard* mix rather than any other mixing pre-set. Therefore, it seems reasonable that NH listeners significantly preferred the *Standard* pre-set over any other because prior musical culture significantly affects people's musical preferences.[43]

In study (1), CI subjects showed no significant preference between the original or the estimated multitracks for the −6 dB mix, but a significant preference for the original multitrack was found for the −12 dB mix (Table VII). Results for study (3) reveal that the mixing preferences for the −12 dB vs Standard changed when estimated multitracks were used but were preserved for the −6 dB vs Standard comparison. Results in studies (1) and (3) suggest that CI subjects can tolerate the SS distortion and artifacts in the estimated vocals when the background is attenuated by 6 dB but not when the background is attenuated by 12 dB. This result is worth highlighting because it means that simplifying the music by re-mixing it using the proposed SS methods is a feasible approach. This is important considering that most of the commercial songs are mono or stereo and require a previous step to estimate the multitrack. However, this claim holds only for mixing pre-sets where the background is attenuated by 6 dB. Yet, other mixing pre-sets might exist (for ratios below 0 dB and between −6 and −12 dB) for which CI listeners do not have a significant preference for mixes created with the original multitracks over those created with the estimated multitracks. Regarding the previous discussion, one can conclude that further research exploring more mixing pre-sets is probably needed.

Experiment 2 individual results (Table VIII) are useful for discussing study (2). There exists a large variability in mixing preferences across CI subjects. This variability can be due to the fact that the mixing preferences are subject-specific or that in some cases the proposed pre-sets did not fit their preference. For example, for six out of nine CI subjects no significant preference was observed for any of the pre-sets proposed for CI users. The choice of the −6 and −12 dB mixing pre-sets (with an anchor mix, −12 dB, that allows a better interpretation of the results) implied using mixing pre-sets that differ significantly from the average results obtained in experiment 1. This could be the reason why most participants showed no preference for any of the proposed pre-sets over the *Standard*. In Table VIII one can also relate the popular music experience of a subject with its mixing preferences. Out of the five subjects that regularly listen to popular music, two have a preference for the *Standard* mix, one has a preference for the mixes with attenuated background and two have no preference for any mixing. Therefore, and differently from NH listeners, subjects familiar with popular music have no common preference for the *Standard* mix. Furthermore, Table VIII shows that some CI subjects can tolerate the SS distortion/artifacts even for the −12 dB mix. Finally, note that S3 preferred a negative instruments-to-vocals ratio in experiment 1 (see Fig. 1) but in experiment 2 S3 preferred the *Standard* mix (see Table VIII). S3 was the unique subject familiar with the song chosen for experiment 1 and moreover, S3 regularly listens to popular western music. We think that in experiment 1 S3 preferred the vocals level to be louder in order to better understand the known lyrics, however in experiment 2 S3 has chosen the *Standard* mix that stands for one of the commercial mixes that this subject normally listens to.

The previously discussed results denote that individual subject-specific mixing pre-sets seem to be the appropriate way to exploit the potential of the SS approach. From the results presented here, one can interpret that general mixing pre-sets may not be helpful for every CI user. A possible way to short-cut these problems in commercial applications using SS is to allow CI users to modify the mixing *in situ*—without assuming any prior preferred mix.

3. Benchmark use case: DRNN

A method for defining a benchmark describing the maximum acceptable levels of distortion in the vocals is proposed in Sec. III B 3. According to the linear model based on the CI listeners' preference ratings, for the −6 dB pre-set the SS algorithms should approximately achieve average SDR and SAR values greater than 4.9 dB. For the −12 dB pre-set the SS algorithms should approximately achieve average

SDR and SAR values greater than 5.7 and 5.8 dB, respectively. However, the minimum SDR and SAR values desirable for the −6 and −12 dB pre-sets, based on the minimum SDR and SAR values inside the non-preference area, should be approximately above 3.6 dB.

In order to emphasize the utility of the proposed benchmark, an additional SS algorithm based on DRNN is introduced as an example in how to use it. In Table VI one can observe that DRNNs can be used to remix the −6 dB pre-set because the global SDR and SAR levels for the vocals are above the average values recommended by the proposed benchmark. However, the SDR levels for the *Bef* and *Hall* songs are below the minimum recommended values. For these songs, it is likely that CI users may be aware of the SS distortions. For the −12 dB pre-set mix, only the *Jud2*, *Jud*, and *Mic2* songs have SDR and SAR measures above the average values recommended by the proposed benchmark. Again, for the other songs it is likely that CI users may be aware of the SS distortion/artifacts.

This study is the first defining a general benchmark considering error measures broadly used by the SS research community. More studies considering not only vocals but also other instruments, with more different mixing pre-sets, taking into account more test songs and with more CI subjects, might further boost the utility of the proposed benchmarking method. These benchmarks may be convenient for directly adopting state-of-the-art SS algorithms to create pilot software for a ready-to-use technology. New SS algorithms may be also tested, especially those that overcome current limitations, such as: real time implementation or separating (and re-mixing) more than one instrument at time.

V. CONCLUSIONS

SS techniques can be used to create re-mixes of popular music that favor vocals over attenuated background instruments to improve CI users' musical experience. However, SS techniques may introduce some distortion and artifacts. A methodology for defining the maximum acceptable levels of distortion and artifacts present in the estimated vocals track is presented. According to our experiments, SS distortion/artifacts were tolerated when the background instruments were attenuated by 6 dB if the average SDR/SAR scores were greater than approximately 4.9 dB. However, some CI subjects tolerated the SS distortion/artifacts when the background instruments were attenuated by 12 dB if the SS algorithm average SDR/SAR scores were greater than approximately 5.7/5.8 dB, respectively. For the mixing pre-sets studied in this manuscript, the multitrack estimations may not exceed the approximate lower SDR/SAR bound of 3.6 dB; otherwise it is likely that CI users may not tolerate the distortion/artifacts. The lower boundaries for the SDR/SAR levels are derived from a non-significant model that fitted a reduced number of experimental data points and therefore, they only provide some intuition about the minimum SDR/SAR levels required to remix music for CI users from estimated multitracks.

The presented results also suggest that CI users may not benefit from general mixing pre-sets, individual subject-specific mixes are needed.

Given that commercial music releases are not available in multitrack format, SS techniques (that allow individual personalized re-mixes) seem to be a promising approach toward a better music appreciation for CI users.

ACKNOWLEDGMENTS

We would like to thank Wim Buyens for sharing the audio excerpts used for this study. Special thanks to the CI users from the German Hearing Center of the Medical University of Hannover and to the NH listeners that volunteered to participate in these extensive experiments. Also thanks to the anonymous reviewers for their suggestions and comments. This work was supported by the DFG Cluster of Excellence EXC 1077/1 "Hearing4all."

[1]W. Nogueira, M. Haro, P. Herrera, and X. Serra, "Music perception with current signal processing strategies for cochlear implants," in *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies*, ACM (2011).
[2]K. Gfeller, A. Christ, K. John, S. Witt, and M. Mehr, "The effects of familiarity and complexity on appraisal of complex songs by cochlear implant recipients and normal hearing adults," J. Music Therapy **40**(2), 78–112 (2003).
[3]E. M. Burns and N. F. Viemeister, "Played-again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**(6), 1655–1660 (1981).
[4]B. C. J. Moore, "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants," Otol. Neurotol. **24**(2), 243–254 (2003).
[5]J. J. Galvin, Q.-J. Fu, and R. V. Shannon, "Melodic contour identification and music perception by cochlear implant users," Ann. N.Y. Acad. Sci. **1169**(1), 518–533 (2009).
[6]Y. Y. Kong, R. Cruz, J. A. Jones, and F. G. Zeng, "Music perception with temporal cues in acoustic and electric hearing," Ear Hear. **25**(2), 173–185 (2004).
[7]H. J. McDermott, "Music perception with cochlear implants: A review," Trends Amplif. **8**(2), 49–82 (2004).
[8]V. Looi, H. McDermott, C. McKay, and L. Hickson, "Music perception of cochlear implant users compared with that of hearing aid users," Ear Hear. **29**(3), 421–434 (2008).
[9]W. Buyens, B. van Dijk, M. Moonen, and J. Wouters, "Music mixing preferences of cochlear implant recipients: A pilot study," Int. J. Audiol. **53**(5), 294–301 (2014).
[10]K. Gfeller, D. Jiang, J. Oleson, V. Driscoll, and J. F. Knutson, "Temporal stability of music perception and appraisal scores of adult cochlear implant recipients," J. Am. Acad. Audiol. **21**(1), 28–34 (2010).
[11]K. Gfeller, A. Christ, J. F. Knutson, S. Witt, K. T. Murray, and R. S. Tyler, "Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients," J. Am. Acad. Audiol. **11**(7), 390–406 (2000).
[12]V. Looi, H. McDermott, C. McKay, and L. Hickson, "Comparisons of quality ratings for music by cochlear implant and hearing aid users," Ear Hear. **28**(2), 59S–61S (2007).
[13]P. J. Donnelly, Z. Guo Benjamin, and J. L. Charles, "Perceptual fusion of polyphonic pitch in cochlear implant users," J. Acoust. Soc. Am. **126**(5), EL128–EL133 (2009).
[14]W. Buyens, B. van Dijk, J. Wouters, and M. Moonen, "A stereo music pre-processing scheme for cochlear implant users," IEEE Trans. Biomed. Eng. **62**(10), 2434–2442 (2015).
[15]G. D. Kohlberg, D. M. Mancuso, D. A. Chari, and A. K. Lalwani, "Music engineering as a novel strategy for enhancing music enjoyment in the cochlear implant recipient," Behav. Neurol. **501**, 829680 (2015).
[16]W. Buyens, B. Van Dijk, J. Wouters, and M. Moonen, "A harmonic/percussive sound separation based music pre-processing scheme for cochlear implant users," in *Proceedings of the 21st European Signal Processing Conference* (*EUSIPCO*) (2013), pp. 1–5.

[17]N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of 16th European Signal Processing Conference* (*EUSIPCO*) (2008), pp. 1–4.
[18]K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," J. Acoust. Soc. Am. **123**(4), 2379–2390 (2008).
[19]W. Nogueira, T. Gajecki, B. Krger, J. Janer, and A. Bchner, "Development of a sound coding strategy based on a deep recurrent neural network for monaural source separation in cochlear implants?," in *Proceedings of the 12th ITG Conference on Speech Communication* (2016).
[20]A. Roebel, J. Pons, M. Liuni, and M. Lagrange, "On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (2015), pp. 414–418.
[21]P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *International Society for Music Information Retrieval (ISMIR)* (2014).
[22]A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," IEEE Trans. Audio, Speech, Lang. Process. **20**(4), 1118–1133 (2012).
[23]D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Advances in Neural Information Processing Systems* (*NIPS*) (2001), pp. 556–562.
[24]W. Nogueira, M. Lopez, T. Rode, S. Doclo, and A. Buechner, "Individualizing a monaural beamformer for cochlear implant users," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (2015), pp. 5738–5742.
[25]R. Marxer and J. Janer, "Low-latency bass separation using harmonic-percussion decomposition," in *Proceedings of the 16th International Conference on Digital Audio Effects* (*DAFx*) (2013).
[26]J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodriguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," EURASIP J. Adv. Signal Process. **2013**(1), 1–16.
[27]M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music genre database and musical instrument sound database," Int. Soc. Music Inf. Retrieval **3**, 229–230 (2003).
[28]B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753 (1983).
[29]A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Stat. Soc. Ser. B **39**(1), 1–38 (1977).
[30]http://bass-db.gforge.inria.fr/fasst/ (Last viewed January 18, 2016).
[31]https://github.com/posenhuang/deeplearningsourceseparation (Last viewed January 18, 2016).
[32]http://sisec.inria.fr/professionally-produced-music-recordings/ (Last viewed January 18, 2016).
[33]D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," Math. Program. **45**(1–3), 503–528 (1989).
[34]E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Lang. Process. **14**(4), 1462–1469 (2006).
[35]http://bass-db.gforge.inria.fr/bss_eval/ (Last viewed January 18, 2016).
[36]Recommendation, I. T. U. R., Bs. 1534-1, "Method for the subjective assessment of intermediate sound quality (MUSHRA)" (International Telecommunications Union, Geneva, 2001).
[37]Recommendation, I. T. U. T., P. 800.1. "Mean opinion score (MOS) terminology" (International Telecommunication Union, Geneva, 2006).
[38]I. Hochmair-Desoyer, E. Schulz, L. Moser, and M. Schmidt, "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users," Am. J. Otol. **18**(6), S83–S83 (1997).
[39]https://github.com/jordipons/MT5 (Last viewed January 18, 2016).
[40]M. Buffa, A. Hallili, and P. R. Gonin, "MT5: A HTML5 multitrack player for musicians," *First Web Audio Conference* (2015).
[41]D. Robinson, "Replay Gain—A proposed standard," Online document (2001), http://wiki.hydrogenaud.io/index.php?title=ReplayGain_1.0_specification.
[42]N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Machine Learn. Res. **15**(1), 1929–1958 (2014).
[43]G. Soley and E. E. Hannon, "Infants prefer the musical meter of their own culture: A cross-cultural comparison," Develop. Psychol. **46**(1), 286–292 (2010).