EQUALIZATION MATCHING OF SPEECH RECORDINGS IN REAL-WORLD ENVIRONMENTS

François G. Germain[†*] *Gautham J. Mysore*[⊗] *Takako Fujioka*[†]

[†]Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA
[⊗]Adobe Research, San Francisco, CA, USA

fgermain@stanford.edu gmysore@adobe.com takako@ccrma.stanford.edu

ABSTRACT

When different parts of speech content such as voice-overs and narration are recorded in real-world environments with different acoustic properties and background noise, the difference in sound quality between the recordings is typically quite audible and therefore undesirable. We propose an algorithm to equalize multiple such speech recordings so that they sound like they were recorded in the same environment. As the timbral content of the speech and background noise typically differ considerably, a simple equalization matching results in a noticeable mismatch in the output signals. A single equalization filter affects both timbres equally and thus cannot disambiguate the competing matching equations of each source. We propose leveraging speech enhancement methods in order to separate speech and background noise, independently apply equalization filtering to each source, and recombine the outputs. By independently equalizing the separated sources, our method is able to better disambiguate the matching equations associated with each source. Therefore the resulting matched signals are perceptually very similar. Additionally, by retaining the background noise in the final output signals, most artifacts from speech enhancement methods are considerably reduced and in general perceptually masked. Subjective listening tests show that our approach significantly outperforms simple equalization matching.

***Index Terms*—** Equalization matching, speech enhancement, voice-overs.

1. INTRODUCTION

Equalization is a common audio effect used to manipulate the spectral balance of an audio signal. In audio mixing, it can be used to emphasize/de-emphasize a frequency range in a signal or to compensate for distortion introduced by hardware or post-processing [1, 2]. On the other hand, in the mastering stage, equalization (EQ) is used to affect the general tonal balance of the track.

Being able to match the perceptual rendering of two audio tracks through equalization is often desirable. For example, one task of audio mastering engineers is to match the tonal balance between different tracks in order to maintain some uniformity [2]. Another application of interest corresponds to the scenario in which segments that are meant to be played back sequentially have been recorded separately in different recording environments (e.g., hardware, location). The change in environment changes the spectral balance of the signal and it can be desirable to smooth the rendering between segments. As a consequence, many professional audio software packages provide *equalization matching* audio effects [3, 4, 5]. To the best of our knowledge, such equalization is often performed by estimating some form of spectral balance for a track and applying it to another.

In this paper, we focus on an application of growing interest: with the increased availability of mobile devices equipped with recording hardware, more and more speech content is recorded on-the-fly for various purposes. Examples include voice-overs, narration, audio stories, and educational content. Such content is often recorded in non-ideal environments with imperfect sound insulation resulting in audio with noticeable background noise and coloration. It is often convenient to record different segments of the content at different times and with different recording setups (e.g., location, recording layout, hardware). This results in mismatched background noise and coloration between those segments. Consequently, the resulting audio track exhibits noticeable transitions between those segments and considerably degrades the listener's experience when playing back the entire track.

Using a simple matching approach where we apply a single equalization filter is not generally suitable to smooth out those differences, as the spectral balance between the background noise and speech content of two different segments are substantially different. Alternatively, prior removal of the background before applying a single equalization filter to the speech alone is also generally unsuitable since either 1) incomplete removal of the background results in a similar issue of mismatched spectral balance between residual background and speech, or 2) complete removal of the background is often associated with noticeable audio artifacts in the speech signal. We present a method that allows for differentiated equalization of the speech and background noise sources and subsequent remixing of them in order to produce a convincing matching with low levels of audio artifacts. In Sec. 2, we present a formulation of the problem and detail our approach. In Sec. 3, we detail our practical implementation. In Sec. 4, we present the results of subjective listening tests validating our approach.

2. EQUALIZATION MATCHING

We consider two different real-world recordings y_a and y_b containing a speech y^S and a background noise y^N source:

$$y_a(t) = y_a^S(t) + y_a^N(t) \text{ and } y_b(t) = y_b^S(t) + y_b^N(t) \quad (1)$$

In the remainder of the paper, we use the equivalent frequency domain representation of the signals:

$$\mathcal{Y}_a(\omega) = \mathcal{Y}_a^S(\omega) + \mathcal{Y}_a^N(\omega) \text{ and } \mathcal{Y}_b(\omega) = \mathcal{Y}_b^S(\omega) + \mathcal{Y}_b^N(\omega) \quad (2)$$

Each source in those mixtures is likely to have been the result of an original signal distorted due to different phenomena occurring during the recording process, such as room reverberation and

[*]Part of this work was performed while interning at Adobe Research.

microphone/sensor response. Additionally, some recording devices can include some additional automated post-processing (e.g., noise suppression). We approximate those distortions as a series of linear operations, and summarize them as a linear filter $\mathcal{H}^i$ applied to each source $i \in \{S, N\}$. The mixtures $\mathcal{Y}_a$ and $\mathcal{Y}_b$ can then be written as:

$$\mathcal{Y}_m(\omega) = \mathcal{H}_m^S(\omega)\mathcal{X}_m^S(\omega) + \mathcal{H}_m^N(\omega)\mathcal{X}_m^N(\omega) \text{ for } m \in \{a, b\} \quad (3)$$

It is generally impractical to discriminate the original source from its distortion. For example, while extensive research exists to estimate and attenuate distortion introduced by reverberation [6, 7, 8], the process can introduce undesirable artifacts. Moreover, when dereverberation is applied to multiple signals recorded in different recording environments, each signal often retains enough reverberation to make them sound quite different in terms of spectral balance, leaving the equalization matching problem unsolved.

An alternative objective can be to match the two sound mixtures $\mathcal{Y}_a$ and $\mathcal{Y}_b$ so as to create a new mixture $\mathcal{Y}_c$ from $\mathcal{Y}_b$ that sounds like it was recorded in the same conditions as $\mathcal{Y}_a$. In our framework, this objective can be accomplished by swapping the distortion filters (i.e., replacing $\mathcal{H}_b^S$ and $\mathcal{H}_b^N$ respectively with $\mathcal{H}_a^S$ and $\mathcal{H}_a^N$). In other words, the matching algorithm aims at generating the mixture $\mathcal{Y}_c$ such that:

$$\mathcal{Y}_c(\omega) = \mathcal{H}_a^S(\omega)\mathcal{X}_b^S(\omega) + \mathcal{H}_a^N(\omega)\mathcal{X}_b^N(\omega) \quad (4)$$

2.1. Simple matching

To the best of our knowledge, existing equalization matching systems typically use a simple equalization matching process as described in this section. This process estimates a single equalization profile (i.e., a linear filter $\mathcal{G}$) for $\mathcal{Y}_a$ and applies it to the signal $\mathcal{Y}_b$ to get the matched signal $\mathcal{Y}_c$ as $\mathcal{Y}_c(\omega) = \mathcal{G}(\omega)\mathcal{Y}_b(\omega)$, meaning:

$$\mathcal{Y}_c(\omega) = \mathcal{G}(\omega)\mathcal{H}_b^S(\omega)\mathcal{X}_b^S(\omega) + \mathcal{G}(\omega)\mathcal{H}_b^N(\omega)\mathcal{X}_b^N(\omega) \quad (5)$$

The matching equations for any $\mathcal{X}_b^S(\omega)$ and $\mathcal{X}_b^N(\omega)$ are then:

$$\mathcal{H}_a^i(\omega) = \mathcal{G}(\omega)\mathcal{H}_b^i(\omega) \text{ for } i \in \{S, N\} \quad (6)$$

In practical cases, it is not possible to estimate satisfactorily the filters $\mathcal{H}_a^S$, $\mathcal{H}_b^S$, $\mathcal{H}_a^N$ and $\mathcal{H}_b^N$. However, we can often assume that the sources from the two tracks have similar statistics (i.e., similar overall spectral content). For example, the average spectral balance of speech segments from the same person are likely to be similar, and background can often be approximated as filtered white noise. We can then assume $\mathcal{X}_a^S(\omega) \approx \mathcal{X}_b^S(\omega)$ and $\mathcal{X}_a^N(\omega) \approx \mathcal{X}_b^N(\omega)$. The matching equations then become:

$$\mathcal{H}_a^i(\omega)\mathcal{X}_a^i(\omega) \approx \mathcal{G}(\omega)\mathcal{H}_b^i(\omega)\mathcal{X}_b^i(\omega) \text{ for } i \in \{S, N\} \quad (7)$$

We can see that neither (6) nor (7) have solutions for $\mathcal{G}(\omega)$ in general as the matching equations are contradictory in both cases, meaning it would not be possible to find a filter that would produce the correct spectral balance in both speech and background. Our best guess is to compute the filter $\mathcal{G}$ as the ratio:

$$\mathcal{G}(\omega) = \mathcal{Y}_a(\omega)/\mathcal{Y}_b(\omega) \quad (8)$$

2.2. Source-differentiated matching

In the previous section, we see that performing a proper matching is made difficult by the fact that the equalization filter affects both sources, a well-known fact in audio mastering [1]. Ideally, we need

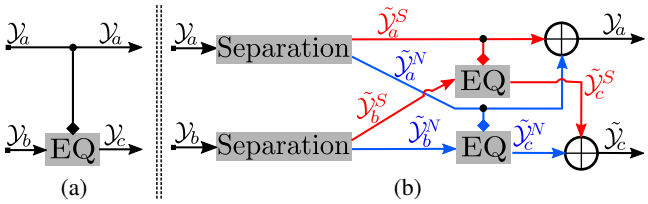

Fig. 1. Structure of the (a) simple and (b) source-differentiated equalization matching systems.

to perform differentiated equalization on each source before mixing them, using two equalization filters $\mathcal{G}^S$ and $\mathcal{G}^N$ so that:

$$\mathcal{Y}_c = \mathcal{G}^S(\omega)\mathcal{H}_b^S(\omega)\mathcal{X}_b^S(\omega) + \mathcal{G}^N(\omega)\mathcal{H}_b^N(\omega)\mathcal{X}_b^N(\omega) \quad (9)$$

The matching equations are then expressed as:

$$\mathcal{H}_a^i(\omega) = \mathcal{G}^i(\omega)\mathcal{H}_b^i(\omega) \text{ for } i \in \{S, N\} \quad (10)$$

With the hypotheses $\mathcal{X}_a^S(\omega) \approx \mathcal{X}_b^S(\omega)$ and $\mathcal{X}_a^N(\omega) \approx \mathcal{X}_b^N(\omega)$, we get the following matching equations:

$$\mathcal{H}_a^i(\omega)\mathcal{X}_a^i(\omega) \approx \mathcal{G}^i(\omega)\mathcal{H}_b^i(\omega)\mathcal{X}_b^i(\omega) \text{ for } i \in \{S, N\} \quad (11)$$

To solve those two equations, we would need access to each source. This is impractical in the desired context, as both sources were recorded simultaneously. The proposed approach is to use a source separation approach to obtain an approximate separation of the two sources so that both segments are separated as follows:

$$\mathcal{Y}_m(\omega) = \tilde{\mathcal{Y}}_m^S(\omega) + \tilde{\mathcal{Y}}_m^N(\omega) \text{ for } m \in \{a, b\} \quad (12)$$

In speech enhancement literature, a common approach is to consider that the Fourier coefficients of the S and N sources are the realization of independent identically-distributed zero-mean random processes with a given distribution (e.g., Gaussian, Laplace) [9]. The linear distortion of the source does not affect the nature of the distribution, meaning for example that if $\Re(\mathcal{X})$ and $\Im(\mathcal{X})$ follow a Gaussian distribution, then so will also $\Re(\mathcal{H}\mathcal{X})$ and $\Im(\mathcal{H}\mathcal{X})$ (and reciprocally) [10]. As a result, common speech enhancement approaches can be used to estimate $\tilde{\mathcal{Y}}_a^S$ and $\tilde{\mathcal{Y}}_a^N$ (respectively $\tilde{\mathcal{Y}}_b^S$ and $\tilde{\mathcal{Y}}_b^N$) from $\mathcal{Y}_a$ (resp. $\mathcal{Y}_b$) as $\tilde{\mathcal{Y}}_a^S$ and $\tilde{\mathcal{Y}}_a^N$ (resp. $\tilde{\mathcal{Y}}_b^S$ and $\tilde{\mathcal{Y}}_b^N$) so that:

$$\tilde{\mathcal{Y}}_m^i(\omega) \approx \mathcal{H}_m^i(\omega)\mathcal{X}_m^i(\omega) \text{ for } m \in \{a, b\} \text{ and } i \in \{S, N\} \quad (13)$$

We can then rewrite the matching equations (10), so that we compute $\mathcal{G}^S(\omega)$ and $\mathcal{G}^N(\omega)$ as the ratios:

$$\mathcal{G}^i(\omega) = \tilde{\mathcal{Y}}_a^i(\omega)/\tilde{\mathcal{Y}}_b^i(\omega) \text{ for } i \in \{S, N\} \quad (14)$$

3. IMPLEMENTATION

In practice, the algorithm is applied in the short-time Fourier transform (STFT) domain. We therefore replace (ω) with (ω, t) in our notation for the signals. In the STFT domain, our equalization filter can capture and compensate for timbral coloration and for reverberation with a duration lesser or equal to the length of an STFT frame.

The structure of the simple equalization system is shown in Fig. 1.a. It outputs $\mathcal{Y}_c(\omega, t) = \mathcal{G}(\omega)\mathcal{Y}_b(\omega, t)$ using $\mathcal{G}$ as defined in (8). In comparison, the structure of the source-differentiated equalization system is presented in Fig. 1.b. The two input mixtures $\mathcal{Y}_a$ and $\mathcal{Y}_b$ are each processed through a separation unit whose outputs follows (12). $\tilde{\mathcal{Y}}_b^S$ (respectively $\tilde{\mathcal{Y}}_b^N$) goes through an EQ unit where its equalization is matched to the equalization of $\tilde{\mathcal{Y}}_a^S$ (resp. $\tilde{\mathcal{Y}}_a^N$) using $\mathcal{G}^S$ (resp. $\mathcal{G}^N$) as defined in (14) so that we get the output:

$$\tilde{\mathcal{Y}}_c(\omega, t) = \mathcal{G}^S(\omega)\tilde{\mathcal{Y}}_b^S(\omega, t) + \mathcal{G}^N(\omega)\tilde{\mathcal{Y}}_b^N(\omega, t) \quad (15)$$

3.1. Separation unit

Source separation is an active field of audio research [11, 12, 13, 14, 15]. In our scenario of interest, we use algorithms derived from the field of speech enhancement [9], which aim at removing relatively stationary background noise signals from speech signals, such that the background can be recovered as a residual from the algorithm.

We use a separation algorithm based on Wiener filtering [16] as we found it to be most suitable for our application. In particular, it was found to separate efficiently most of the background noise, resulting in an improved estimation and application of the differentiated background equalization without significant performance loss in the speech equalization. Other methods were found to achieve a separation balance less suitable to a good performance for our approach, mostly due to insufficient background removal from the enhanced speech, resulting in matching issues similar to those we encounter in the simple equalization (Sec. 2.1) with audible mismatched background still remaining with the speech. The separation we achieve with Wiener filtering introduces noticeable artifacts in the enhanced speech signal due to strong noise suppression. However, in the context of differentiated equalization, the impact of those artifacts is considerably mitigated by the fact that neither component of the signal is removed, but rather remixed with a modified spectral balance.

Wiener filtering relies on modeling the STFT coefficients of the two sources $\mathcal{Y}^S(\omega,t)$ and $\mathcal{Y}^N(\omega,t)$ as zero-mean Gaussian distributions with variances $\lambda^S(\omega,t)$ and $\lambda^N(\omega,t)$, and requires 1) to compute an estimate of the background noise spectral profile $\hat{\lambda}^N(\omega,t)$, and 2) to compute the *a priori* signal-to noise ratio (SNR) $\xi(\omega,t)$:

$$\xi(\omega,t) = \lambda^S(\omega,t)/\lambda^N(\omega,t) \tag{16}$$

For our offline algorithm, we use the non-causal algorithm described in [17] to estimate ξ. Then, we estimate the STFT speech frames through the weighting $|\mathcal{L}(\omega,t)|$ of the mixture frames:

$$\mathcal{Y}^S(\omega,t) = |\mathcal{L}(\omega,t)|\mathcal{Y}(\omega,t) = \frac{\xi(\omega,t)}{\xi(\omega,t)+1}\mathcal{Y}(\omega,t) \tag{17}$$

3.2. Equalization (EQ) unit

To find the equalization filters $\mathcal{G}^S$ and $\mathcal{G}^N$ based on our assumptions, we can estimate their magnitude frequency response using (14) as:

$$|\mathcal{G}^i(\omega)| = |\tilde{\mathcal{Y}}^i_a(\omega)|/|\tilde{\mathcal{Y}}^i_b(\omega)| \text{ for } i \in \{S,N\} \tag{18}$$

In the STFT domain, the coefficients are modeled as realizations of the sources' random variables, leading to the possible estimators:

$$|\mathcal{G}^i(\omega)| = \left(\frac{E[|\tilde{\mathcal{Y}}^i_b(\omega,t)|^k]}{E[|\tilde{\mathcal{Y}}^i_a(\omega,t)|^k]}\right)^{1/k} \text{ for } i \in \{S,N\} \tag{19}$$

In practice, for a given k, we estimate the filters as empirical averages over the STFT frames:

$$|\hat{\mathcal{G}}^i(\omega)| = \left(\frac{\frac{1}{T_b}\sum_{t=1}^{T_b}|\tilde{\mathcal{Y}}^i_b(\omega,t)|^k}{\left[\frac{1}{T_a}\sum_{t=1}^{T_a}|\tilde{\mathcal{Y}}^i_a(\omega,t)|^k\right]+\epsilon}\right)^{1/k} \text{ for } i \in \{S,N\} \tag{20}$$

where ϵ is a small value added to avoid numerical problems for very low energy bands. Acknowledging the limitations of the separation units, we know that 1) speech is not present in all spectral frames, and 2) the background noise estimated as residual can be highly distorted in the presence of speech. For this reason, we estimate the filter for a given source by using only the STFT frames for which the estimated energy of that source is higher than the estimated energy of the other source. Our assumptions regarding the different random processes do not allow us to estimate the correct phase of the filters, so we consider only linear-phase filters with magnitude $|\hat{\mathcal{G}}^i(\omega)|$. These filters have the advantage of preserving the transients of the signals, but they cannot compensate for the temporal smearing resulting from distortions such as long reverberations. Hence, while the overall timbre of the two signals can be matched, the temporal structure of their distortion remains unaffected. For the simple EQ matching, the system has a single EQ unit (Fig. 1.a) outputting $\mathcal{Y}_c(\omega,t) = \mathcal{G}(\omega)\mathcal{Y}_b(\omega,t)$. We compute $|\mathcal{G}(\omega)|$ from (8) with an estimator similar to (20) with $\mathcal{Y}_a$ and $\mathcal{Y}_b$ in place of $\tilde{\mathcal{Y}}^i_a$ and $\tilde{\mathcal{Y}}^i_b$.

3.3. Realizable filters

Multiplying directly the STFT frames with an arbitrary zero-phase spectral weighting for separation (with $|\mathcal{L}(\omega,t)|$) or EQ (with $|\mathcal{G}(\omega)|$) corresponds to using non-realizable filters [9], resulting in time-aliasing artifacts and non-consistent STFTs [18, 19, 20]. We convert weightings in realizable ones using the following process:

- We zero-pad the signal frames (of length M) used to compute the STFT by L samples to get $M+L$ frequency bands.
- Given a zero-phase weighting $|\mathcal{H}(\omega)|$, we 1) compute the inverse discrete Fourier transform (DFT) of $|\mathcal{H}(\omega)|$, 2) circularly shift the resulting impulse response by $\frac{L}{2}$ samples to the right, 3) time-limit the impulse response with a window of length L centered at the $\left(\frac{L}{2}+1\right)$th sample, so that only the first L taps are non-zero, and 4) compute the DFT of the windowed impulse response to get a realizable weighting $\mathcal{H}(\omega)$.

This process effectively smooths out $|\mathcal{H}(\omega)|$ in a way depending on the window. Applying the resulting (linear-phase) weighting avoids time-aliasing artifacts, while its phase results in an added delay of $\frac{L}{2}$ samples to the output that we compensate for after processing.

3.4. Energy matching and signal-to-noise ratio

The described processing has the advantage of providing approximate energy matching for each source as, for $i \in \{S,N\}$, we have:

$$\sum_\omega |\mathcal{Y}^i_a(\omega)|^2 \approx \sum_\omega |\mathcal{G}^i(\omega)\mathcal{Y}^i_b(\omega)|^2 = \sum_\omega |\tilde{\mathcal{Y}}^i_c(\omega)|^2 \tag{21}$$

It is also interesting to notice that the signal-to-noise ratio (SNR) of $\tilde{\mathcal{Y}}_c$ also roughly matches the SNR of $\mathcal{Y}_a$ as:

$$\frac{\sum_\omega |\mathcal{Y}^S_a(\omega)|^2}{\sum_\omega |\mathcal{Y}^N_a(\omega)|^2} \approx \frac{\sum_\omega |\mathcal{G}^S(\omega)\mathcal{Y}^S_b(\omega)|^2}{\sum_\omega |\mathcal{G}^N(\omega)\mathcal{Y}^N_b(\omega)|^2} = \frac{\sum_\omega |\tilde{\mathcal{Y}}^S_c(\omega)|^2}{\sum_\omega |\tilde{\mathcal{Y}}^N_c(\omega)|^2} \tag{22}$$

Hence, the matching process results in a significant increase in SNR in the second audio segment if the SNR of signal a is significantly higher than the SNR of signal b. Due to the artifacts arising from the separation unit, this increase in SNR can result in suboptimal quality in the output audio signal as those artifacts are not well masked. This issue can be mitigated by altering the SNR of the output audio signal at remixing to improve the masking. To do so, we apply a remixing gain γ to the equalized background audio as:

$$\mathcal{Z}_a = \mathcal{Y}^S_a + \gamma\mathcal{Y}^N_a \text{ and } \tilde{\mathcal{Z}}_c = \tilde{\mathcal{Y}}^S_c + \gamma\tilde{\mathcal{Y}}^N_c \tag{23}$$

The SNR of the remixed signals $\mathcal{Z}_a$ and $\tilde{\mathcal{Z}}_c$ is then given by:

$$\mathrm{SNR}_z = \frac{1}{\gamma^2}\frac{\sum_\omega |\mathcal{Y}^S_a(\omega)|^2}{\sum_\omega |\mathcal{Y}^N_a(\omega)|^2} = \frac{\mathrm{SNR}_y}{\gamma^2} \tag{24}$$

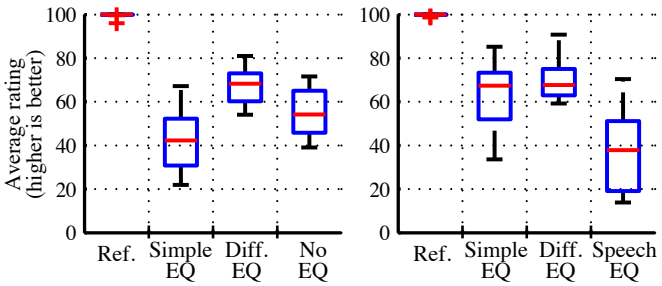

Fig. 2. Distribution of each listener's average rating across the 10 combination examples for the 4 different cases tested in the matching task (left) and the quality task (right).

4. SUBJECTIVE LISTENING TESTS

Traditional objective metrics for speech enhancement [9] and audio source separation [21, 22] are not well suited to measure equalization matching performance due to the nature of the task. For this reason, we conduct MUSHRA-like [23] listening tests based on the MUSHRAM interface[1] [24] to assess the performance of our system.

4.1. Data

For our experiments, we used the data from the DAPS dataset [25]. This dataset provides recordings of 5 scripts read by 20 different speakers (10 male, 10 female) in 10 different real-world recording environments, i.e., combinations of a room (e.g., living room) and a device (e.g., iPad). The room *balcony* was not used as the level and quality of the background do not correspond to our target scenario.

This provides us with 1000 unique combinations of speaker, script and environment (room+device). For the recordings corresponding to a given speaker and script, we extract 20 seconds of audio so that about 0.5 seconds of background noise is present (required for our separation algorithm) at the beginning of the segment and that the transition at 10 seconds has low speech energy. We then replace the last 10 seconds of that signal with the corresponding audio from another environment (room+device), and apply our matching algorithms to match the second half of the signal to the first half. Each signal is sampled at 44.1kHz. For the STFT, we use 1024-sample long Hann windows with 50% overlap and zero-padded by 1024 samples. To estimate the different equalization filters, we use $k = 1$ and $\epsilon = 10^{-3}$, and we time-limit them using a 1024-sample Hann window. Finally, we extract the 10-second segment around the environment transition. Audio examples are available at `https://ccrma.stanford.edu/~francois/EQM.html`.

For the listening tests, 10 combination examples (speaker, script, environment no.1, environment no.2) are picked at random from our dataset. 10 unique speakers (5 male, 5 female) are each present exactly once and each of the 5 scripts are present exactly twice in the examples. We pick 10 unique pairs of environment no.1 and environment no.2 so that room no.1 and room no.2 are different.

4.2. Tasks and subjects

In the listening tests, subjects were asked to rate audio files while listening over headphones, answering two distinct questions in two independent consecutive tasks. 12 subjects were recruited among the student and alumni community at Stanford University. All subjects are between the age of 18 and 35, have no reported hearing problem, and have a prior general knowledge of audio processing and/or audio engineering. Half of the subjects took the quality task first, while the other half took the matching task first.

In the matching task, subjects are asked to judge how "matched" each sound file is, in the sense that the beginning and the end of the file sound like they have been recorded in the same environment (room+device). For that task, for each of the 10 examples, the listeners are presented with 4 cases: 1) material that was recorded in a single environment no.1 (reference), 2) the same material with the first half recorded in environment no.1 and the second half in environment no.2 and then matched to no.1 using the simple EQ algorithm (Sec. 2.1), 3) the same material with the first half recorded in environment no.1 and the second half in environment no.2 and then matched to no.1 using the proposed algorithm (Sec. 2.2), and 4) for the same material with the first half recorded in environment no.1 and the second half in environment no.2 and left unchanged.

In the quality task, subjects are asked to judge the quality of each sound file, in terms of perceived distortion, unnaturalness, and unpleasantness without consideration of the change in recording environment. For that task, for each of the 10 examples, the listeners are also presented with 4 cases: 1), 2) and 3) are identical to 1), 2) and 3) for the previous task, while 4) is the same material processed similarly to 3) except that we keep only the equalized speech part without remixing the equalized background noise in the signal.

4.3. Results

For each listener, we compute the average of their ratings for each given case across the 10 selected combination examples. The distribution of those averages is presented in Fig. 2. A one-way repeated-measures ANOVA [26] is conducted to compare the matching and the quality average rating distributions for the three non-reference cases. The difference between the different means is found to be significant for the matching task ($F(2, 22) = 36.694, p < 10^{-4}$) and the quality task ($F(2, 22) = 19.224, p < 10^{-4}$). We also run paired-sample t-tests between the differentiated EQ and the other cases, we find the improvement of the differentiated EQ to be significant in matching over the simple EQ ($p = 1.9 \times 10^{-5}$) and the unequalized file ($p = 1.6 \times 10^{-3}$) and in quality over the speech-only EQ ($p = 8 \times 10^{-5}$). The difference in quality between simple and differentiated EQ is found to be non-significant.

These results demonstrate the ability of differentiated EQ to improve the matching between audio segments recorded in two distinct environments while the simple EQ filter fails to equalize the signals, actually degrading the overall sense of matching. Additionally, we found that a few of the tested files using differentiated EQ were well enough matched to be misclassified as reference. They also demonstrate the interest of remixing back the background noise in the final mixture in order to mask some of the distortion artifacts introduced by the different units in the differentiated EQ system.

5. CONCLUSION

In this paper, we presented a novel algorithm for performing equalization matching of speech recordings in real-world environments. Our algorithm separates speech and background noise, independently equalizes them, and finally remixes them. This approach disambiguates the competing matching equations of the two sources, while limiting the presence of artifacts. It significantly outperforms the traditional method of using a single equalization filter.

Acknowledgment—The authors warmly thank Professor Julius O. Smith (Stanford University) for funding the listening tests.

[1] `http://c4dm.eecs.qmul.ac.uk/downloads/#mushram`

6. REFERENCES

[1] R. A. Katz, *Mastering Audio: The Art and the Science*, Focal Press, Burlington, MA, 3rd edition, 2014.

[2] S. Savage, *Mixing and Mastering in the Box: the Guide to Making Great Mixes and Final Masters on your Computer*, 2014.

[3] Apple Inc., *Logic Pro X Effects for Mac OSX*, 2013.

[4] FabFilter Software Instruments, *Pro-Q^2 Manual*, 2014.

[5] iZotope, Inc., *Ozone 6 Help Documentation*, 2014.

[6] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement.*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[7] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[8] D. Liang, M. D. Hoffman, and G. J. Mysore, "Speech dereverberation using a learned speech model," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1871–1875.

[9] P. C. Loizou, *Speech enhancement: theory and practice*, vol. 30, CRC Press, Boca Raton, FL, 2007.

[10] S. M. Ross, *A First Course in Probability*, Pearson Prentice Hall, 2008.

[11] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, 2006.

[12] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind speech separation*, Springer, 2007.

[13] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[14] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, 2010.

[15] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[16] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, vol. 2, pp. 629–632.

[17] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[18] J. B. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[19] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proceedings of the 2008 Workshop on Statistical and Perceptual Audition (SAPA)*, 2008, pp. 23–28.

[20] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, 2013.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[23] International Telecommunication Union, "BS.1534-2: Method for the subjective assessment of intermediate quality levels of coding systems," 2014.

[24] E. Vincent, M. Jafari, and M. Plumbley, "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.

[25] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[26] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, John Wiley & Sons, 2012.