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

Overview

Our algorithm performs equalization matching of speech segments recorded from a speaker in **non-ideal, mismatched, real-world recording conditions**.

The algorithm matches the different speech and background spectral balances by using separation algorithms and performing **source-differentiated equalization matching.**

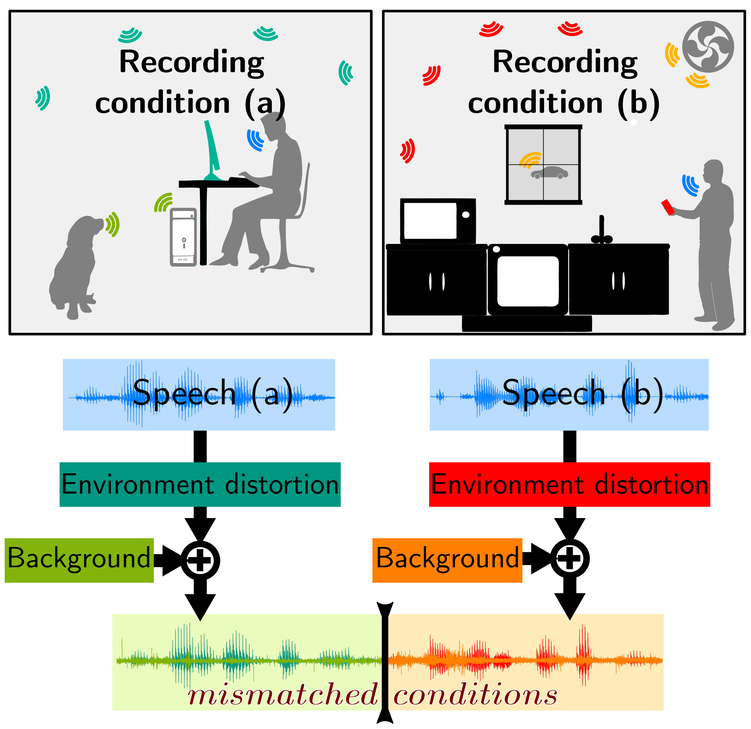
Listening tests show that **our approach significantly outperforms simple equalization matching.**

Context

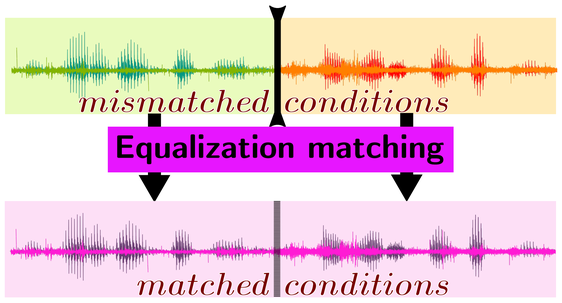
The increasing availability of portable consumer recording devices (smartphones, tablets,...) has led to an explosion of available recorded speech content.

The recorded speech with those devices often display **non-ideal recording conditions,** e.g.:
— **Environment distortion** of the speech such as *reverberation, microphone response, pre-processing*
— Added **background noise**

A problem arises when audio content is recorded in different locations and at different times: the audio has **mismatched recording conditions**, resulting in segments with **different qualities**, and **undesirable transitions**.



Equalization matching



Equalization matching is the operation of **matching the spectral balance of two signal segments.**

In our context, we use equalization matching on two speech segments so that they sound **as if they were recorded in the same conditions.**

Signal model

Models for recorded speech signals (a) and (b):

$$\mathcal{Y}_a(\omega) = \mathcal{H}_a^S(\omega)\,\mathcal{X}_a^S(\omega) + \mathcal{H}_a^N(\omega)\,\mathcal{X}_a^N(\omega)$$
$$\mathcal{Y}_b(\omega) = \underbrace{\mathcal{H}_b^S(\omega)}_{\text{speech distortion}}\,\underbrace{\mathcal{X}_b^S(\omega)}_{\text{speech source}} + \underbrace{\mathcal{H}_b^N(\omega)}_{\text{background distortion}}\,\underbrace{\mathcal{X}_b^N(\omega)}_{\text{background source}}$$

Equalization objective ("make (b) sound like (a)"):

non-equalized: $\mathcal{Y}_b(\omega) = \mathcal{H}_b^S(\omega)\mathcal{X}_b^S(\omega) + \mathcal{H}_b^N(\omega)\mathcal{X}_b^N(\omega)$

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

Source statistic hypothesis:

$$\mathcal{X}_a^S(\omega) \approx \mathcal{X}_b^S(\omega) \text{ and } \mathcal{X}_a^N(\omega) \approx \mathcal{X}_b^N(\omega)$$

Source-differentiated EQ

Simple equalization matching: $\mathcal{Y}_c(\omega) = \mathcal{G}(\omega)\mathcal{Y}_b(\omega)$

Equalization matching conditions:

$$\mathcal{G}(\omega) = \underbrace{\frac{\mathcal{H}_a^S(\omega)}{\mathcal{H}_b^S(\omega)}}_{\text{speech EQ}} \not\equiv \underbrace{\frac{\mathcal{H}_a^N(\omega)}{\mathcal{H}_b^N(\omega)}}_{\text{background EQ}} \quad \boxed{\Rightarrow \text{No solution for } \mathcal{G}}$$

Source-differentiated equalization matching (proposed):

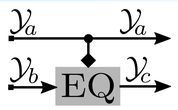
$$\mathcal{Y}_c(\omega) = \underbrace{G^S(\omega)}_{\text{speech EQ}}\mathcal{H}_b^S(\omega)\mathcal{X}_b^S(\omega) + \underbrace{G^N(\omega)}_{\text{background EQ}}\mathcal{H}_b^N(\omega)\mathcal{X}_b^N(\omega)$$

Equalization matching conditions:

$$\mathcal{G}^S(\omega) = \frac{\mathcal{H}_a^S(\omega)}{\mathcal{H}_b^S(\omega)} \text{ and } \mathcal{G}^N(\omega) = \frac{\mathcal{H}_a^N(\omega)}{\mathcal{H}_b^N(\omega)}$$

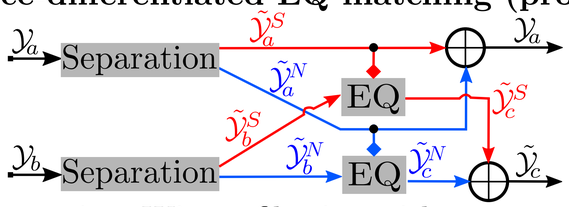
Algorithm

Simple EQ matching:



— EQ: Estimate and apply $\mathcal{G}(\omega) = \mathcal{Y}_a(\omega)/\mathcal{Y}_b(\omega)$

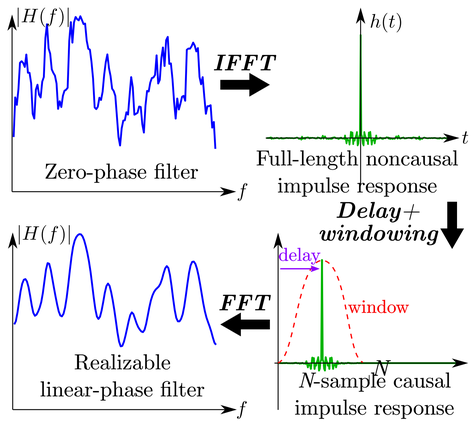
Source-differentiated EQ matching (proposed):



— Separation: Wiener filtering with noncausal estimation of a priori SNR
— EQ: Estimate and apply $\mathcal{G}^i(\omega) = \mathcal{H}_a^i(\omega)/\mathcal{H}_b^i(\omega)$

Realizable filters

From the magnitude, we form realizable filters by:
— zero-pad the signal FFT by N samples;
— form causal filters of length N through windowing and delay from the magnitude (see right).



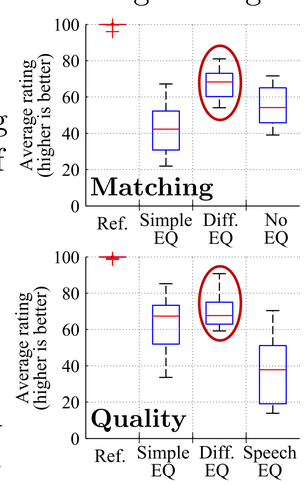
Subjective listening tests

Data: 10 recordings built from the DAPS dataset (real-world speech recordings in varied environments).

Subjects: 12 listeners from the CCRMA community trained in audio processing and/or audio engineering.

Questions:
— Rate the **MATCHING** of beginning and end of the recording
— Rate the overall **QUALITY** of the recording

Results: Our approach **significantly outperforms** simple equalization matching in matching, while **avoiding the loss in quality** usually associated with speech separation algorithms.



Examples: https://ccrma.stanford.edu/~francois/EQM.html

41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, 20–25 March 2016