AUDITORY FILTERBANKS BENEFIT UNIVERSAL SOUND SOURCE SEPARATION

Han Li[1,2] Kean Chen[1] Bernhard U. Seeber[2]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, 710072, Xi'an, China

[2] Audio Information Processing, Technical University of Munich, 80290, Munich, Germany

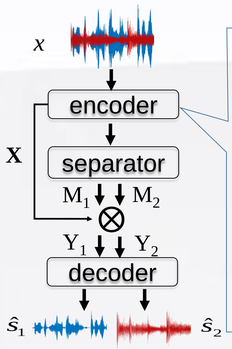
Contact information: lihan@mail.nwpu.edu.cn

ICASSP2021 TORONTO Canada June 6–11, 2021 Metro Toronto Convention Centre

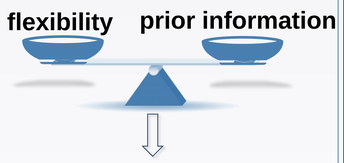
Paper number 1948

Parameterized Gammatone filterbanks for universal source separation

➢ Framework for source separation



Filterbank	Fixed	Free
Properties	set manually before training	learned with the network
Pros	intuitive interpretation	high flexibility
Cons	limited performance	easy to be disturbed by noise



Parameterized Gammatone filterbank is proposed.

➢ Parameterized Gammatone filterbank (GTFB)

GTFB

GTFB is based on the Gammatone function, whose parameters are jointly learned with the network.

$$g(t) = \sqrt{2\frac{(4\pi b)^{2p+1}}{2p!}} t^{p-1} e^{-2\pi bt} \cos(2\pi f_c t + \phi)$$

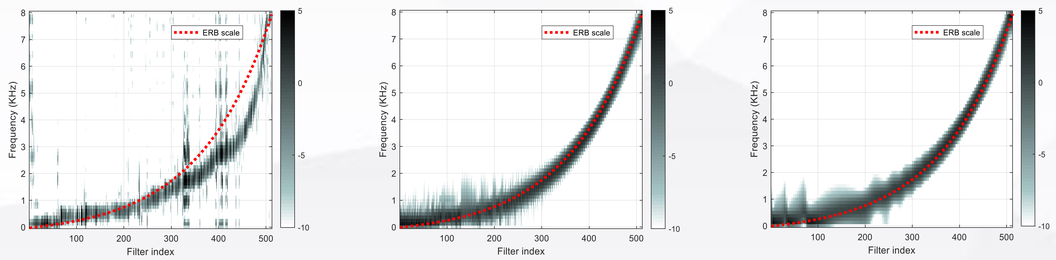
- Learnable parameter set: $\theta_i = \{p_i, f_c^i, b_i, \phi_i\}_{i=1,\dots,N}$
- Parameter set initiation mimic the mechanical response of the basilar membrane.

separation results

Filterbanks	SI-SDRi (dB)
STFT	9.21
Fixed-GTFB	8.15
Param-GTFB	**10.46**
Free	10.02

Parameterized Gammatone filterbank benefit universal source separation!

➢ Filterbank properties



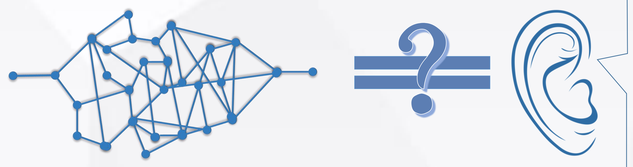
(a) free filterbank (b) parameterized GTFB (c) fixed GTFB

Fig. 1. Frequency response of different filterbanks. The red dashed line indicates the mapping from linear frequency to ERB scale.

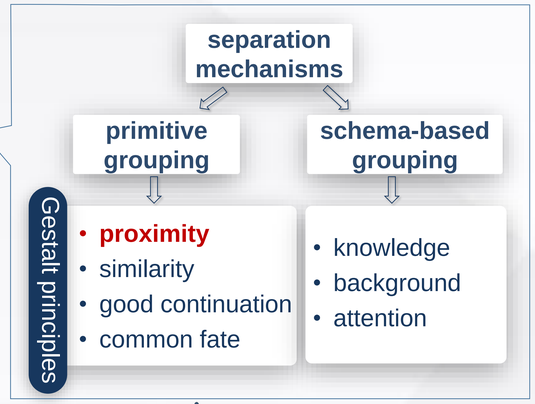
- The free learned filters are bandpass filters that are distributed on a nonlinear scale like in the auditory system.
- **Auditory-like filterbanks are suitable for the source separation system !**

Underlying separation mechanisms

➢ Does deep network mimic the human ear?

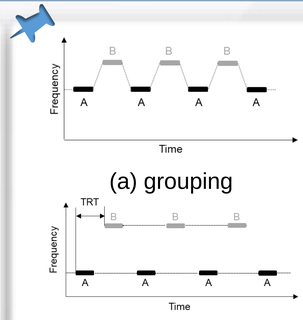


input deep network output



➢ Methods and stimuli

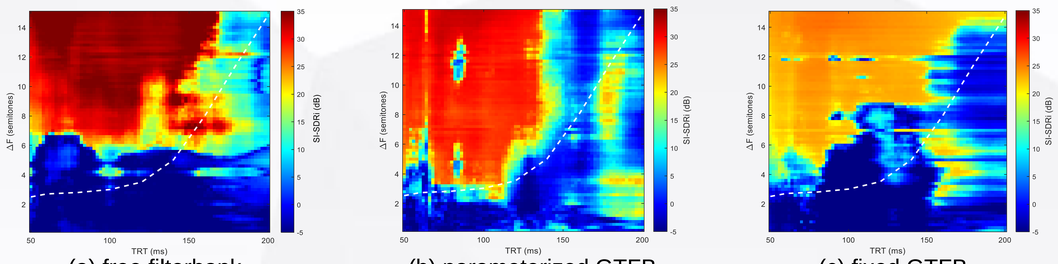
- One classic segregation experiment revealing the ***proximity principle*** in frequency and time is chosen to test the model's behavior, as shown in the right figure.
- If the frequency and temporal distance between tones A and B is small, they tend to group to one source (top panel). If the distance is large, they are easier to separate into two sources (bottom panel).
- All experiments are tested on the same model trained by the universal dataset without any other adjustment, and the experimental mixtures here are not included in the training dataset.



(a) grouping

(b) separation

➢ Results



(a) free filterbank (b) parameterized GTFB (c) fixed GTFB

Fig. 2. SI-SDR improvement (dB) for tone sequences separation with different filterbanks as a function of ΔF and TRT.

- When ΔF is large, and TRT is short (top left corner), tone sequences A and B are more likely to be separated. It is consistent with the temporal coherence boundary presented by van Noorden (the white dashed line).
- **No matter which filterbank is used, the network learned the general Gestalt principle (proximity in frequency and time) automatically from nature sound sources !**