基于语音分离的人工设计特征,参数化特征和可学习特征的比较

朱文博，王谋，张晓雷 [1]，Susanto Rahardja
（西北工业大学航海学院智能声学与临境通信研究中心，陕西西安 710072）

摘要：在语音分离任务中，声学特征的设计是十分重要的。声学特征可以大致分为三类：人工设计特征，参数化特征和可学习特征。其中，可学习特征是指将其与分离网络以端到端的方式进行联合训练，如时域卷积语音分离网络(convolutional time domain audio separation network, Conv-Tasnet)，这成为了如今语音分离研究中的一种新的趋势。然而在最近的研究中证明了人工设计特征以及参数化特征也能产生具有竞争力的结果。但是，截止目前还没有工作对这三种声学特征进行系统的比较。本文通过设置不同声学特征作为编码器和解码器，在 Conv-Tasnet 框架下对它们进行比较。我们还将人工设计的多相位 Gammatone 滤波器组(multi-phase gammatone filterbank, MPGTF)扩展为一种新的参数化多相位 gammatone 滤波器组(Parameterized MPGTF, ParaMPGTF)。在 WSJ0-2mix 数据集上的实验结果表明，(i)如果解码器是可学习特征时，将编码器设置为 STFT,MPGTF,ParaMPGTF 以及可学习特征的性能相近，(ii)如果将 STFT,MPGTF,ParaMPGTF 的逆变换作为解码器时，所提出的 ParaMPGTF 相比于其他两种人工设计特征有更好的性能。

关键词：语音分离；人工设计特征；参数化特征；可学习特征；多相位 gammatone 滤波器组
中图分类号：TN912.3　　**文献标识码：A**

A COMPARISON OF HANDCRAFTED, PARAMETERIZED, AND LEARNABLE FEATURES FOR SPEECH SEPARATION

Wenbo Zhu, Mou Wang,Xiao-Lei Zhang,Susanto Rahardja

(CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shanxi 710072, China)

Abstract: The design of acoustic features is important for speech separation. It can be roughly categorized into three classes: handcrafted, parameterized, and learnable features. Among them, learnable features, which are trained with separation networks jointly in an end-to-end fashion, become a new trend of modern speech separation research, e.g. convolutional time domain audio separation network (Conv-Tasnet), while handcrafted and parameterized features are also shown competitive in very recent studies. However, a systematic comparison across the three kinds of acoustic features has not been conducted yet. In this paper, we compare them in the framework of Conv-Tasnet by setting its encoder and decoder with different acoustic features. We also generalize the handcrafted multi-phase gammatone

[1] 通讯作者：张晓雷，1983 年 4 月生，教授，博士生导师，主研究兴趣是语音处理、机器学习， 邮箱：xiaolei.zhang@nwpu.edu.cn

filterbank (MPGTF) to a new parameterized multi-phase gammatone filterbank (ParaMPGTF). Experimental results on the WSJ0-2mix corpus show that (i) if the decoder is learnable, then setting the encoder to STFT, MPGTF, ParaMPGTF, and learnable features lead to similar performance; and (ii) when the pseudo-inverse transforms of STFT, MPGTF, and ParaMPGTF are used as the decoders, the proposed ParaMPGTF performs better than the other two handcrafted features.

Key words：Speech separation, handcrafted features, learnable features, parameterized features, multi-phase gammatone filterbank

1 引言 [2]

语音分离的目的是将多个音源的混合语音分离成其对应成分。在本文中，我们研究了基于深度学习的说话人无关情况下的语音分离，其中说话人无关的情况是指训练时所用到的说话人与测试中的说话人可以不相同[1]。Hershey等人首先提出用深度聚类的方法来解决语音分离问题[2]。在此之后，针对语音分离问题又提出了多种方法，例如置换不变训练[3-4]，深度吸引子网络[5]。在这些方法中，被广泛应用的声学特征是短时傅里叶变换的幅度谱(short – time Fourier transform,STFT)。然而，在从分离后的幅度谱恢复成时域信号的过程中，所用到的是含有噪声的相位谱，这会导致得到次优的性能。

为了克服这一缺陷，数据驱动的从时域到时频域变换的可学习特征成为了新的趋势。其中代表性的就是一维卷积滤波器(1D-conv)[6-9]。由于该变换是与分离网络联合训练的，并且不需要额外的人工操作，因此该变换相比于STFT来说使语音分离的性能得到了提升。在这些时域方法中, Conv-Tasnet在帧长设置为仅2毫秒的低时延情况下得到了杰出的分离性能，从而受到了广泛的关注。

近期有一些工作旨在研究Conv-Tasnet的声学特征。例如，Ditter和Gerkmann用人工设计特征[10]，即多相位Gammatone滤波器组(MPGTF)来代替Conv-Tasnet中编码器部分的可学习特征，并在尺度无关信噪比(scale-invariant source-to-noise，SI-SNR)上带来了提升。Pariente等人将参数化滤波器扩展为了复值的解析滤波器[11-12]，同时他们也提出了类似的一维卷积滤波器的解析版本。解析的一维卷积滤波器相比于原始的Conv-Tasnet也有性能上的提升。上述结果表明，人工设计特征和参数化特征与目前最先进的可学习特征相比也具有竞争力。

然而，目前缺少对于可学习特征，人工设计特征以及参数化特征的比较。受到用人工设计特征来代替编码器或解码器的可学习特征的启发，在这篇文章中我们将三种类型的特征在Conv-Tasnet框架下进行了比较。同时为了了解这三种特征之间的联系，我们将多相位gammatone滤波器组和参数化特征进行了结合，提出了参数化多相位gammatone滤波器组(ParaMPGTF)。其中，ParaMPGTF的中心频率和带宽将与分离网络联合训练。我们在WSJ0-2mix数据集[2]上比较了STFT, MPGTF,ParaMPGTF以及可学习特征。实验结果表明，如果解码器是可学习特征，将编码器设置为参与比较特征中的任意一种都产生了相似的性能。我们还比较了将STFT,MPGTF,ParaMPGTF作为编码器，它们的逆变换作为解码器。实验结果表明，我们所提出的ParaMPGTF比其他两种人工设计特征的性能要好。

本文将以下面所述进行组织编排。第二节介绍了比较的框架以及所提出的ParaMPGTF，第三节展示实验结果。第四节总结了我们的发现。

2 方法

2.1 问题描述

当给定 C 个声源$\mathbf{s}_c(t)_{c=1}^{C}$，其中$t$是时间索引，则它们的混合语音被定义为：

$$\mathbf{x}(t) = \sum_{c=1}^{C} \mathbf{s}_c(t) \tag{1}$$

则语音分离问题的定义可以描述为从$\mathbf{x}(t)$中得到第 c 个声源$\mathbf{s}_c(t)$的精确估计$\hat{\mathbf{s}}_c(t)$。

本文研究的基础分离框架是 Conv-Tasnet。如图

1 所示，它由三个主要部分构成：编码器，分离网络和解码器。编码器和解码器采用小帧长来显著降低系统时延。编码器和解码器是可学习的一维卷积滤波器，他的作用是在时域信号和时频特征之间进行类似的转换。分离网络是一个由一维扩张卷积块堆叠成的全卷积的分离模块 [13-14]，以 SI-SNR 为损失进行优化。其作用是为每个音源产生一个掩模。

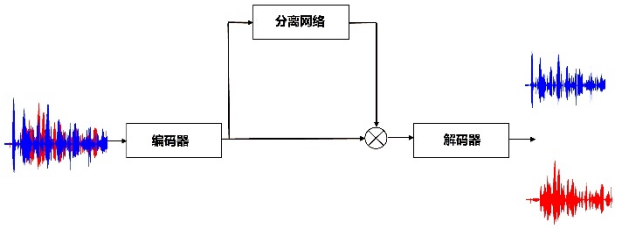


图 1 Conv-Tasnet 的框架图

2.2 比较框架

本文比较了用人工设计特征，参数化特征以及可学习特征作为编码器和解码器的实验结果。编码器可以看作是N个长度为L的滤波器的集合。编码器的输出是由输入混合语音和滤波器卷积所产生的：

$$\mathbf{X}(n,i)=\mathcal{H}(\sum_{l=0}^{L-1}\mathbf{x}(iD+l)\mathbf{h}_n^{\mathrm{Enc}}(L-l)) \quad (2)$$

其中n是滤波器的索引，i是帧数的索引，D是帧移，$\mathbf{h}_n^{\mathrm{Enc}}(\cdot)$是滤波器组中第$n$个滤波器，$l$是一帧当中采样点的索引，$\mathcal{H}(\cdot)$是修正线性单元(ReLU)，其目的是为了保证所输出的表示非负。在本文的比较中，$\mathbf{h}_n^{\mathrm{Enc}(\cdot)}$可以代表三种特征变换中的任意一种。

解码器的作用是重构第c个说话人的时域语音信号$\hat{\mathbf{s}}_c\in\mathbb{R}^T$。解码器的输出为：

$$\hat{\mathbf{s}}_c(k,i)=\sum_{n=0}^{N-1}\hat{\mathbf{S}}_c(n,i)\mathbf{h}_{N-n}^{\mathrm{Dec}}(k) \quad (3)$$

其中$\hat{\mathbf{S}}_c(n,i)$是第c个说话人的分离网络的输出，k是滤波器权重的索引，$\mathbf{h}_n^{\mathrm{Dec}(\cdot)}$是解码器中的第$n$个滤波器，$\hat{\mathbf{s}}_c(k,i)$是第$c$个说话人在第$i$帧的估计。为了对语音帧之间的帧移操作进行解码，解码器进一步计算$\hat{\mathbf{s}}_c(t)=\sum_{i=-\infty}^{\infty}\hat{\mathbf{s}}_c(t-iD,i)$。

用 STFT，MPGTF，ParaMPGTF 和可学习特征作为$\mathbf{h}_n^{\mathrm{Enc}(\cdot)}$，它们的逆变换作为$\mathbf{h}_{N-n}^{\mathrm{Dec}}(\cdot)$的比较，以及所提出的 ParaMPGTF 将会在下一小节展示。

2.3 参数化多相位 gammatone 滤波器组

Gammatone 滤波器组模拟了人类听觉系统的掩蔽效应，在语音分离任务中是一种良好的特征 [15]。Gammatone 滤波器的冲激响应函数$\gamma(t)$为：

$$\gamma(t)=\alpha t^{n-1}\exp(-2\pi bt)\cos(2\pi f_c t+\phi) \quad (4)$$

其中n是滤波器阶数，b是带宽参数，f_c是滤波器的中心频率，$t>0$是时间，α是幅度，ϕ是相移。Ditter 和 Gerkmann 将传统的 gammatone 滤波器组改进成了多相位 gammatone 滤波器组 [10]，主要有以下三个方面的改进。第一，滤波器的长度被设置成了 2 毫秒，其目的是为了使系统低时延。第二，对每个滤波器$\mathbf{h}_n^{\mathrm{Enc}(\cdot)}$，MPGTF 引入了$-\mathbf{h}_n^{\mathrm{Enc}(\cdot)}$来保证在每一个中心频率处，至少有一个滤波器含有能量。第三，相移ϕ在相同中心频率下变化。关于 MPGTF 的详见参考文章[10]。

由公式(4)，我们发现滤波器中心频率f_c和带宽参数b是两个重要的参数。他们由矩形带通滤波器的等效矩形带宽所决定 [16]：

$$\mathrm{ERB}(f_c,c_1,c_2)=c_1+\frac{f_c}{c_2} \quad (5)$$

$$f_c=c_2(\mathrm{ERB}-c_1) \quad (6)$$

$$b=\frac{\mathrm{ERB}\sqrt{(n-1)!}}{\pi((2n-2)!)2^{2-2n}} \quad (7)$$

其中，c_1和c_2是两个参数。通常情况下，根据经验公式，c_1和c_2分别被设置为 24.7 和 9.265[16]。然而，这种经验设置可能不够准确，可能会导致次优的性能。

为了克服这个问题，我们提出了参数化多相位 gammatone 滤波器组，滤波器组参数c_1和c_2将与网络联合训练。同时，对每一次迭代，我们将由公式(7)计算的参数b以及中心频率$f_{c_1},f_{c_2},\ldots,f_{c_M}$按下式计算：

$$f_{c_j}=\mathrm{ERB}_{\mathrm{scale}}^{-1}(\mathrm{ERB}_{\mathrm{scale}}(f_{c_{j-1}})+1) \quad (8)$$

其中f_{c_j}表示根据更新后的c_1和c_2计算得到第j个滤波器的中心频率，M是滤波器组中的滤波器数量。$\mathrm{ERB}_{\mathrm{scalc}}$表示的是将$1/\mathrm{ERB}(f_c)$根据频率积分得到的ERB尺度，$\mathrm{ERB}_{\mathrm{scalc}}^{-1}$是$\mathrm{ERB}_{\mathrm{scalc}}$的逆。事实上，$\mathrm{ERB}_{\mathrm{scalc}}$和$\mathrm{ERB}_{\mathrm{scalc}}^{-1}$是由以下公式计算得到的：

$$\mathrm{ERB}_{\mathrm{scale}}(f_{\mathrm{Hz}}) = c_2 \log(1 + \frac{f_{\mathrm{Hz}}}{c_1 c_2}) \quad (9)$$

$$\mathrm{ERB}_{\mathrm{scale}}^{-1}(\mathrm{ERB}_{\mathrm{scale}}) = c_1 c_2 (\mathrm{e}^{\frac{\mathrm{ERB}_{\mathrm{scale}}}{c_2}} - 1) \quad (10)$$

其中f_{Hz}表示频率变量。当得到$f_{c_1}, \ldots, f_{c_M}$和b后，我们根据公式(4)得到更新后的滤波器组。为了使ParaMPGTF 成为有实际物理意义的滤波器组，$f_{c_1}, f_{c_2}, \ldots, f_{c_M}$应该被限制在 100Hz 和 4000Hz 之间。为了满足这一限制，我们在整个训练过程中将f_{c_1}固定为 100Hz。综上所述，ParaMPGTF 将数据驱动方式和 MPGTF 进行了结合，它同时也继承了 MPGTF 的性质。

3 实验及结果

3.1 数据集

我们使用 WSJ0-2mix 数据集对双说话人语音分离性能进行比较[2]。它包含了 30 个小时的训练数据，10 小时的验证数据以及 5 小时的测试数据。WSJ0-2mix 中的混合语音是通过在 Wall Street Journal(WSJ0)训练集 si_tr_s 中随机选择不同的说话者和句子产生的，并将它们以-5 分贝到 5 分贝范围中的随机信噪比混合。测试集中的句子来自于 WSJ0 数据集中 si_dt_05 和 si_et_05 中 16 个训练中未用到的说话人。WSJ0-2mix 中所有的语音均被重采样至 8000 赫兹。

3.2 实验设置

该网络在 4 秒长的片段上进行了 200 个周期的训练。优化器采用 Adam 优化器，初始学习率为 0.001。如果在验证集上连续 5 个周期性能没有提升则学习率减半。同时，当验证集上的性能在过去的 10 个周期内都没有提升时，网络训练将会被停止。网络的超参数设置遵循 Conv-Tasnet 中的网络超参数[10]，其中滤波器数目N为 512。时序卷积网络(Temporal Convolutional Networks，TCN)的掩模函数分别被设置为 sigmoid 函数和修正线性单元(rectified linear unit，ReLU)。对于 ParaMPGTF,我们将阶数n设置为 2，幅度α设置为 1。我们将c_1和c_2的初始值设置为其经验值，即$c_1 = 24.7$, $c_2 = 9.265$。我们将 SI-SNR 作为评价指标。所报告的结果均是 3000 句测试混合语音的平均结果。

表 1 当解码器为可学习特征时，不同特征作为编码器的比较

编码器	解码器	掩模函数	SI-SNR(dB)	
			Dev	Test
Learned	Learned	Sigmoid	17.61	16.92
Learned	Learned	ReLU	17.45	16.89
MPGTF	Learned	ReLU	17.66	17.20
ParaMPGTF	Learned	ReLU	17.71	17.06
STFT	Learned	ReLU	**17.96**	**17.28**

表 2 当解码器为可学习特征时，MPGT 和 ParaMPGTF 中c_1和c_2的比较。

	MPGTF	**ParaMPGTF**
c_1	24.7	25.09
c_2	9.265	9.198

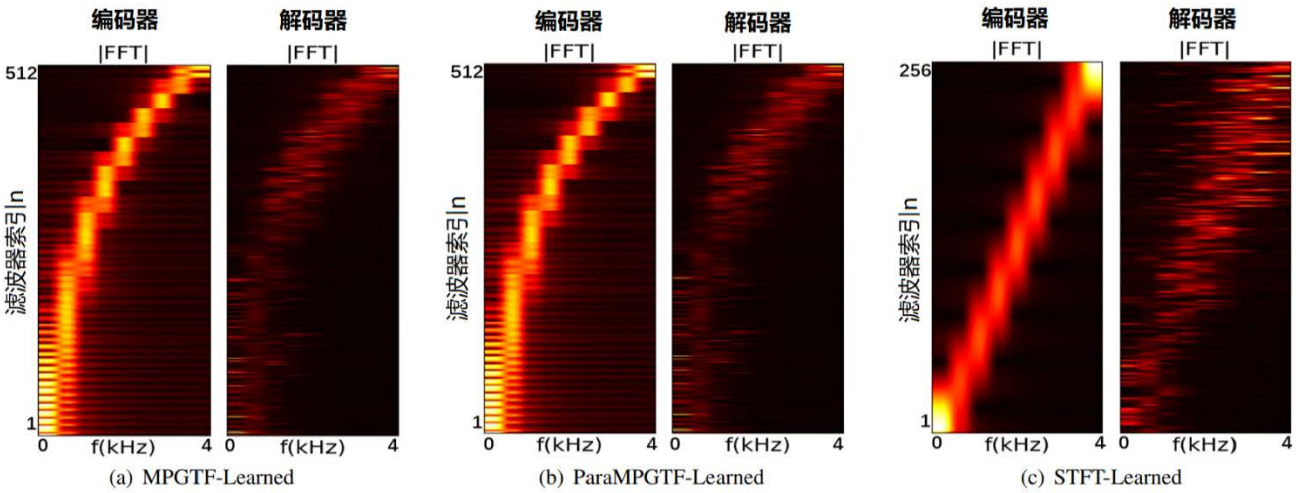
3.3 解码器为可学习特征时的结果

我们首先比较了解码器为可学习特征，编码器为 STFT,MPGTF,ParaMPGTF 和可学习特征时的情况，表 1 列出了比较结果。从表 1 中可以看出，这四种特征并没有产生很大的性能差异。如果我们仔细比较，我们发现 STFT 特征在测试集和验证集都达到最高的性能。MPGTF 和 ParaMPGTF 性能比较接近，ParaMPGTF 在验证集上略好于 MPGTF，而在测试集上略差于 MPGTF。

3.4 解码器为编码器逆变换时的结果

图 2 所示的是用 MPGTF,ParaMPGTF,STFT 和可学习特征作为编码器，解码器为可学习特征的幅度谱图，由于 STFT 的实部部分和虚部部分有相似的形状[17-18]，因此我们这里只绘制了从 1 到 256 频点的 STFT。滤波器在 0 到 4000 赫兹的范围内均匀分布。从图中可以看出，ParaMPGTF 和 MPGTF 的幅度谱图是相似的。这一现象不仅说明了它们的性能相似，而且也说明了参数化特征能够被成功地优化。不仅如此，图 2 也表明了(i)MPGTF 是一个良好的人工设计特征,(ii)可学习的解码器能够有效的

学习到编码器的反变换。表 2 列出了人工设计特征 MPGTF 的C_1和C_2以及 ParaMPGTF 中优化得到的C_1和C_2。从表中我们可以看出两组参数十分接近，这也进一步解释了 MPGTF 和 ParaMPGTF 相似的性能。



(a) MPGTF-Learned (b) ParaMPGTF-Learned (c) STFT-Learned

图 2 不同设置的编码器和解码器的幅度谱图的可视化。左边为基于 MPGTF 的编码器，中间为基于 ParaMPGTF 的编码器，右边为基于 STFT 的编码器。

在该实验中，我们将分别将编码器设置为 STFT，MPGTF，ParaMPGTF，并将解码器设置为其对应的逆变换。表 3 列出了 STFT，MPGTF，ParaMPGTF 以及它们逆变换分别作为编码器和解码器的实验结果。从表中我们可以看出，这三种比较方法的性能大体上是相似的。

表 3 编码器和解码器为不同特征及其逆变换时的比较。

编码器	**解码器**	SI-SNR(dB	
		Dev	Test
MPGTF	MPGTF Pseudo Inv.	16.32	15.73
ParaMPGTF	ParaMPGTF Pseudo Inv.	**16.64**	**16.04**
STFT	ISTFT	16.31	15.82

如果我们仔细研究细节，我们发现在测试集和验证集上，我们所提出的 ParaMPGTF 都达到了最好的性能，这也表明了参数化训练的策略有改进传统人工设计特征的潜力。图 3 展示的是将解码器为编码器的逆变换时所训练的模型在验证集上的收敛曲线。图中我们可以发现可学习特征比人工设计特征和参数化特征收敛的更快。尽管人工设计特征和 ParaMPGTF 在前期以相似的速度收敛，然而 ParaMPGTF 收敛的更快。

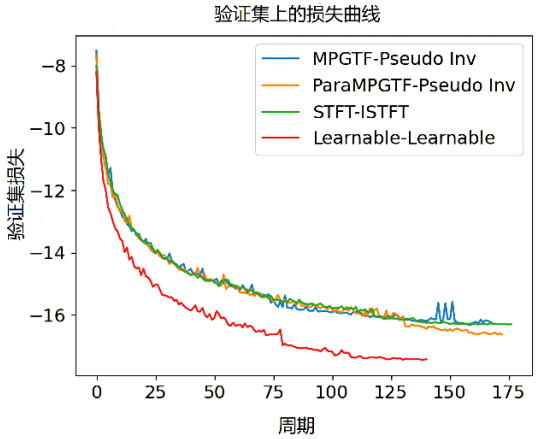


图 3 不同编码器-解码器的收敛曲线

4 结论

在本文中，我们提出了一种参数化的多相位 gammatone 滤波器组。ParaMPGTF 将 MPGTF 中的核心参数与网络进行联合训练。我们还在同一个实验框架中比较了人工设计特征，参数化特征和可学习特征。据我们所知，这是第一个将三种特征放在一起比较。所比较的特征有 STFT，MPGTF，ParaMPGTF 和可学习特征。实验结果表明，当解码器设置为可学习特征时，这四种特征的表现相似。STFT 比其他特征的性能稍好。当解码器设置为编码器的逆变换时，ParaMPGTF 比其他人工设计特征的性能好。

参考文献

[1] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018

[2] J. R. Hershey, Z.Chen, J.Le Roux, and S. Watanabe,“Deep clustering: Discriminative embeddings for segmentation and separation,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 31–35

[3] D. Yu, M. Kolbk, Z. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 241–245

[4] M. Kolbk, D. Yu, Z. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901–1913, 2017

[5] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,”in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 246–250CaminosL,Garcia-Gonzalez A, Gonzalez-Herrera A, et al. Numerical Analysis of the Influence of the Auditory External Canal Geometry on the Human Hearing Response[C]. AIP Conference Proceedings-American Institute of Physics. 2011, 1403(1): 515.

[6] Yi Luo and Nima Mesgarani, “Tasnet: time-domain audio separation network for real-time, single-channel speech separation,” 2017.

[7] Yi Luo and Nima Mesgarani, “Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation,” 2018.

[8] A. Pandey and D. Wang, “Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6875–6879.

[9] Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu, Jiqing Han, and Anyan Shi, “Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks,” 2019.

[10] D. Ditter and T. Gerkmann, “A multi-phase gammatone filterbank for speech separation via tasnet,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 36–40

[11] Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, “Filterbank design for end-toend speech separation,” 2019

[12] Mirco Ravanelli and Yoshua Bengio, “Speaker recognition from raw waveform with sincnet,” 2018

[13] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. ´Hager, “Temporal convolutional networks: A unified approach to action segmentation,” in Computer Vision–ECCV 2016 Workshops, Gang Hua and Herve J ´ egou, ´ Eds., Cham, 2016, pp. 47–54, Springer International Publishing.

[14] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1003–1012.

[15] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, “Complex sounds and auditory images,” Auditory Physiology and Perception, pp. 429–446, 1992

[16] V Hohmann, “Frequency analysis and synthesis using a gammatone filterbank,” Acta Acustica United with Acustica, vol. 88, no. 3, pp. 433–442, 2002.

[17] A. Pandey and D. Wang, “A new framework for cnn-based speech enhancement in the time domain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1179–1188, 2019

[18] Ashutosh Pandey and DeLiang Wang, “A new framework for cnn-based speech enhancement in the time domain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1179–1188, 2019