Computer Science and Application 计算机科学与应用, 2023, 13(4), 689-697
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134068

Hans汉斯

基于图傅里叶变换的语音增强算法研究

刘山民，徐珑婷

东华大学信息科学与技术学院，上海

收稿日期：2023年3月10日；录用日期：2023年4月7日；发布日期：2023年4月14日

摘　要

在语音增强过程中，人们往往采用语音信号的频谱信息作为特征输入，再进行进一步的训练增强处理。最为常见的便是对语音信号进行短时傅里叶变换后取其幅度频谱作为特征输入，在语音恢复阶段，则将含有噪声语音的相位信息作为增强语音的相位信息进行语音的重构。但是，这一做法必然导致相位信息的缺失。本文提出将图傅里叶变换(GFT)分别与非负矩阵分解(NMF)算法以及全卷积神经网络(FCNN)模型相结合来实现含有噪声语音的增强，实验表明，图傅里叶变换－非负矩阵分解算法在语音增强上与短时傅里叶变换－非负矩阵分解算法表现相当，基于图傅里叶变换－全卷积神经网络的语音增强相较于基于短时傅里叶变换－全卷积神经网络的语音增强有更为优异的性能。

关键词

语音增强，短时傅里叶变换，图傅里叶变换，非负矩阵分解，全卷积神经网络

Research on Speech Enhancement Algorithm Based on Graph Fourier Transform

Shanmin Liu, Longting Xu

College of Information Science and Technology, Donghua University, Shanghai

Received: Mar. 10th, 2023; accepted: Apr. 7th, 2023; published: Apr. 14th, 2023

Abstract

In the process of speech enhancement, people often use the spectral information of the speech signal as the feature input, and then carry out further training enhancement processing. The most used is to perform a short-term Fourier transform on the speech signal and take its amplitude spectrum as feature input, and in the speech recovery stage, the phase information of the noisy speech is used as the phase information of the enhanced speech for speech reconstruction. However, this practice

文章引用: 刘山民, 徐珑婷. 基于图傅里叶变换的语音增强算法研究[J]. 计算机科学与应用, 2023, 13(4): 689-697.
DOI: 10.12677/csa.2023.134068

inevitably leads to the absence of phase information. In this paper, it is proposed to combine the graph Fourier transform with the non-negative matrix factorization algorithm and the fully convolutional neural network model to realize the enhancement of noisy speech, and the experimental results show that the performance of graph Fourier transform-non-negative matrix factorization algorithm is comparable to that of the short-term Fourier transform-non-negative matrix factorization algorithm in speech enhancement, and the speech enhancement based on the graph Fourier transform-fully convolutional neural network has better performance than the speech enhancement based on the short-time Fourier transform-fully convolutional neural network.

Keywords

Voice Enhancement, Short-Time Fourier Transform, Graph Fourier Transform, Non-Negative Matrix Factorization, Fully Convolutional Neural Network

1. 引言

随着信息技术的快速发展以及人工智能技术的不断地落地推进，以语音为基础的信息科技不断涌现。例如：语音识别、声纹识别、音视频会议等。在语音技术领域，语音增强技术一直起着举足轻重的作用。随着语音识别、语音通信应用场景中环境向着复杂化、多样化发展，语音质量也将面临着各式各样的挑战，语音增强将作为众多语音应用的前端处理模块为后续模块提供高质量的语音。语音增强效果如何重点在于训练目标的选择[1] [2] [3]以及模型的构建[4] [5]。

在特征模块，研究人员大多采用时序语音信号经时频分解后得到时频特征作为模型的输入，时频特征又可分为：幅度频谱与复频谱。基于幅度频谱的语音增强，在训练阶段只将幅度频谱喂入模型中，最后得到一个收敛的模型，在语音增强阶段将带有噪声语音的相位信息与增强后的幅度频谱结合后重构语音时序信号[6] [7] [8]。虽然以幅度谱为特征进行语音增强时取得了非常优异的效果，但是其忽略了相位信息在语音增强中的作用，因为在早期研究中，人们认为相位信息对于语音的增强作用微乎其微[9]。但是，新的研究表明，相位信息对语音的质量好坏起着重要作用[10] [11]。由于相位信息具有缠绕性[12]，深度神经网络不善于处理非结构化的数据。因此，人们采用复频谱作为语音增强模型的输入特征构建一个多通道的语音增强模型，相较于幅度谱语音增强模型取得更好的语音增强效果。但是，采用复频谱为特征的双通道语音增强模型相比于幅度谱为特征的模型需要有更大的算力要求以及更多的训练时间。以DCCRN [13]为例，相较于 DCRN 增加了 4 倍的训练时间以及 2 倍的模型参数。

因此，本文拟采用图傅里叶变换作为语音时序信号的时频分解方法，分别结合传统的非负矩阵分解算法以及全卷积神经网络算法进行语音增强的研究。

2. 基于图傅里叶变换－非负矩阵分解的语音增强

2.1. 语音时序信号的图傅里叶变换

对语音信号进行图傅里叶变换，首先需要构造语音信号的图信号表示。一段语音时序信号可表示为 $S=\left[S_1,S_2,\cdots,S_n\right]$。为了将语音时序信号转变为图信号，需要先对时序语音信号进行加窗和分帧的操作，

于是有 $S=\left[S_0^f,S_1^f,\cdots,S_m^f\right]_{N\times m}$ 表示将语音时序信号进行分帧后共有 m 帧，每帧的信号点数为 N。对任意一帧有 $S_i^f=\left[s_h,s_{h+1},\cdots,s_{h+N-1}\right]$，$S_i^f$ 便是构造图所需的节点集合，进一步地表示节点之间的关系集合即为边。在此用邻接矩阵表示节点之间的关系，如下式：

$$A_{speech}=\begin{pmatrix}1&1&1&0&0&0\\0&1&\cdots&0&0&0\\0&0&1&1&\cdots&0\\0&0&0&1&\cdots&0\\\vdots&\vdots&\ddots&\ddots&\ddots&0\\1&1&\cdots&0&0&1\end{pmatrix}_{256\times256} \tag{1}$$

由此，语音时序信号的图便有 $S_i^f=\left[s_h,s_{h+1},\cdots,s_{h+N-1}\right]$ 和 A_{speech} 表征构成，基于此便可依据图傅里叶变换[14]将语音时序信号 $S=\left[S_0^f,S_1^f,\cdots,S_m^f\right]_{N\times m}$ 映射到图频域上。

2.2. 图傅里叶变换－非负矩阵分解算法

基于图傅里叶变换－非负矩阵分解的语音增强算法可分为训练和测试两个部分，其流程图如图 1 所示。

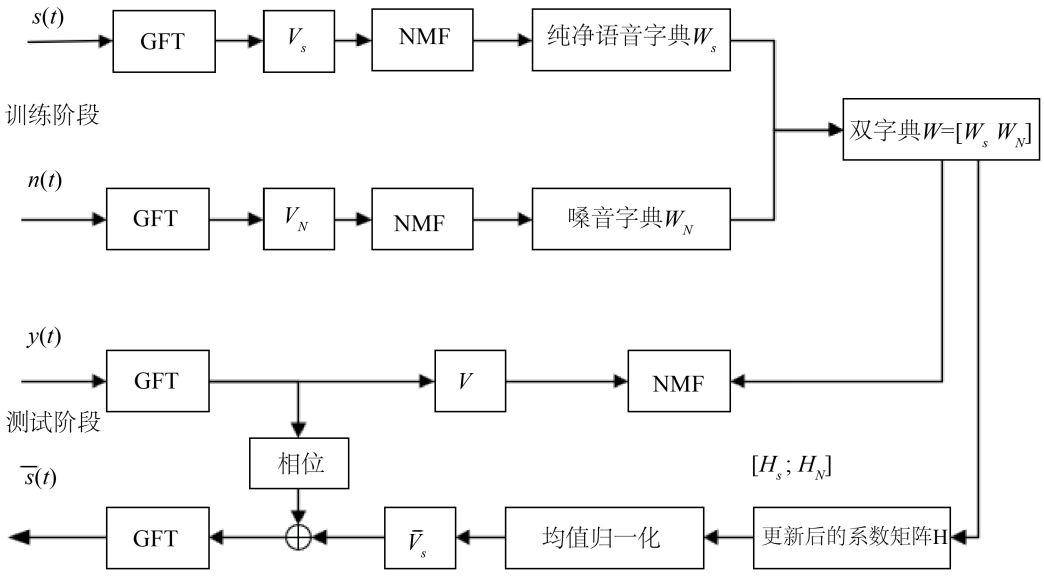

Figure 1. Flow chart of the graph Fourier transform-non-negative matrix factorization algorithm
图 1. 图傅里叶变换－非负矩阵分解算法流程图

训练阶段：

首先，通过 GFT 分别将干净语音和噪声的时域信号 $s(t)$ 和 $n(t)$ 转换为图频域信号。在 GFT 的基础上，可以获得干净语音图频域特征 V_S 和噪声图频域特征矩阵 V_n。然后，基于 NMF 算法，分别将图频域矩阵 V_S 和 V_n 分解为基矩阵 W 和参数矩阵 H。

使用 NMF 算法训练大量干净语音和噪声，以获得语音字典 W_s 和噪声字典 W_n。通过 W_s 和 W_n 的串联

连接，建立了语音增强的整个基矩阵。

测试阶段：

为了对带有噪声的语音进行增强，首先使用 GFT 获得了带有噪声的语音 $y(t)$ 的图频域谱 V 和符号。根据 NMF 算法的更新规则，更新系数矩阵 H_s'。然后，将更新的矩阵 H_s' 与先前的矩阵 W_s 相乘，以获得重构的图频域矩阵 $\tilde{V}_s$。通过增强后的语音图频域矩阵和保存的符号，使用图傅里叶逆变换(IGFT)来生成增强后的语音时域信号 $\tilde{s}(t)$。

3. 基于图傅里叶变换－全卷积神经网络的语音增强

与传统的语音增强算法相比，深度学习算法依托大数据的训练表现出更为优异的语音增强性能。

3.1. 全卷积神经网络结构

FCNN 算法模型由输入层、编码器、解码器、全连接层、输出层构成[15]，可以综合考虑特征中的时间和频率两个维度上的相关性更深层次地挖掘特征信息，从而达到更好的训练效果。在此模型中，编码器和解码器是基于特征图数量对称设立的，从编码器到解码器的特征图数量依次为 16、8、16、1。FCNN 模型结构如下图 2 所示：在输入时采用 7 帧 GFT 特征作为模型的输入，在输出时得到 1 帧的 GFT 特征输出。模型中的编码器和解码器主要负责 GFT 图频域信息的提取，实现基于频域上的信息提取。而在全连接层主要将目标帧相邻帧的特征信息汇聚起来得到目标帧，实现时域上的信息提取。

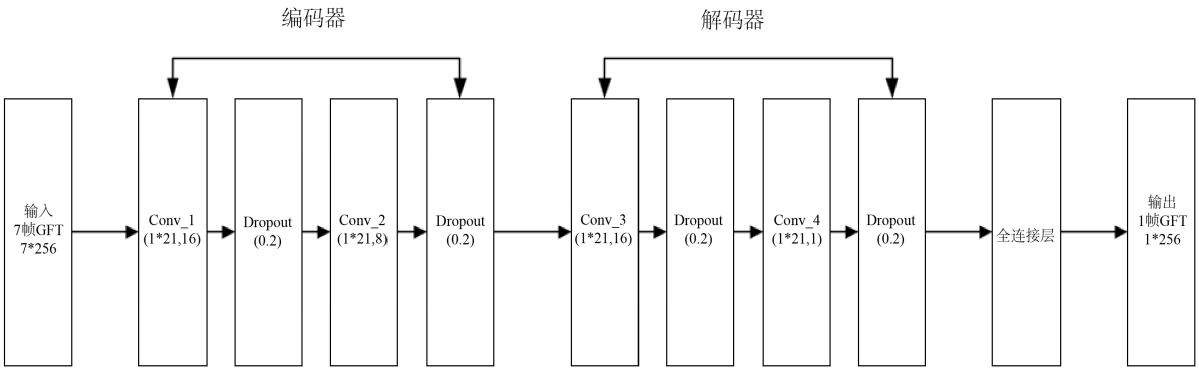

Figure 2. Flow chart of the fully convolutional neural network
图 2. 全卷积神经网络流程图

3.2. 基于全卷积神经网络的训练和测试过程

在训练阶段，将加了噪声的语音特征功率谱作为模型的输入喂入模型中，同时将与该语音特征功率谱对应的比值掩蔽(IRM) [16]作为训练目标，实现语音增强。在测试阶段，将需要增强的语音喂入到训练好的模型中，得到目标函数 IRM。基于 IRM 得到增强后的语音的特征，然后进行 IGFT 得到增强后的时序语音信号。IRM 定义如下：

$$IRM_{t,f} = \left(\frac{Px_{t,f}}{Px_{t,f} + Pn_{t,f}} \right)^{\beta} \tag{2}$$

式中，$Px_{t,f}$、$Pn_{t,f}$ 分别表示干净语音的频谱单元和噪声的频谱单元。此处 β 取值为 0.5。

输入模型的语音特征处理过程以及 FCNN 模型训练与测试过程分别如图 3、图 4 所示。

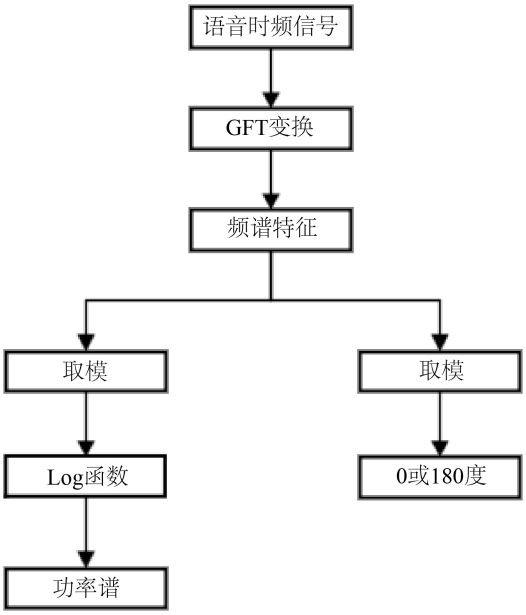

Figure 3. Diagram of the feature extraction process
图 3. 特征提取过程图

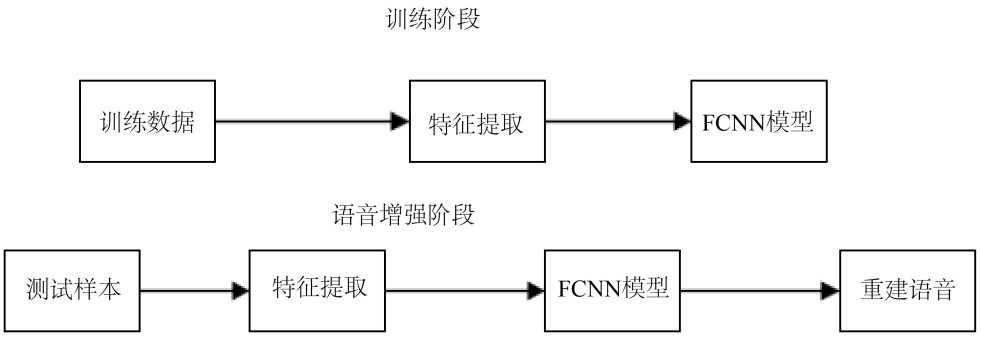

Figure 4. Diagram of the training and test process
图 4. 训练测试过程图

4. 实验结果与分析

在基于 GFT-NMF 算法以及基于 GFT-FCNN 算法的语音增强均采用 LibriTTS [17]语音语料库作为实验中干净语音的数据来源。有所不同的是：在基于 GFT-NMF 的语音增强中，在训练阶段，分别随机选择了 120 句男声的语音用于字典训练。在实验中，有七种类型的噪声被应用：驱逐舰机舱噪声(Destroyer)、座舱噪声(F16)、工厂车间噪声 1 (Factory 1)、工厂车间噪声 2 (Factory 2)、军车噪声(M109)、沃尔沃卡车内部噪声(Volvo)和白噪声(White)。在本实验中，设置了六个不同的信噪比，分别为−5 dB、0 dB、5 dB、10 dB、15 dB 与 20 dB，这些噪声人为地添加到清晰 120 句男声的语音中。在 GFT-FCNN 算法的语音增强模型中从 LibriTTS 语音语料库随机选取 4662 条语音作为训练集，同时选择 babble 与 white 作为加入干净语音中的噪声。在两个算法模型的实验中均将短时傅里叶变换(STFT)作为 baseline 作为对比。

4.1. 基于 GFT-NMF 的实验结果分析

在该实验中，采用语音质量感知评估(PESQ)以及短时客观可懂度(STOI)作为语音质量的评价标准。

以基于 STFT-NMF 算法的语音增强作为 baseline，将本文提出的基于 GFT-NMF 算法的语音增强与之进行对比参照。表 1 以 PESQ 为性能指标对基于 STFT-NMF 与基于 GFT-NMF 算法的语音增强进行了对比，表 2 则以 STOI 为评价指标对基于 STFT-NMF 与基于 GFT-NMF 算法的语音增强性能进行比较。对表 1 以及表 2 进行分析，可以得出如下结论：在语音质量感知方面，在高信噪比的情况下，基于 GFT-NMF 算法的语音增强要优于基于 STFT-NMF 算法的语音增强。在语音可懂度方面，基于 GFT-NMF 算法的语音增强则与基于 STFT-NMF 算法的语音增强表现相当。

Table 1. Comparison of male speech enhancement performance based on STFT-NMF and GFT-NMF algorithms (PESQ)
表 1. 基于 STFT-NMF 和 GFT-NMF 算法的男性语音增强性能比较(PESQ)

SNR	方法	N1	N2	N3	N4	N5	N6	N7
−5	STFT	1.3246	1.1186	1.007	1.5551	1.6102	2.5401	1.7322
	GFT	1.2415	1.0671	1.0258	1.4989	1.5057	2.4739	1.6362
0	STFT	1.7787	1.5979	1.4635	1.9823	2.0225	2.7083	2.0827
	GFT	1.7098	1.5493	1.4477	1.921	1.932	2.6787	2.0297
5	STFT	2.1521	2.0328	1.9093	2.3017	2.3359	2.8241	2.3187
	GFT	2.1138	1.9992	1.8925	2.2646	2.2831	2.841	2.3178
10	STFT	2.409	2.3447	2.2497	2.5176	2.5525	2.8974	2.4852
	GFT	2.408	2.3458	2.2479	2.5171	2.5439	2.9607	2.5204
15	STFT	2.5777	2.55	2.4768	2.662	2.6971	2.9415	2.6046
	GFT	2.6072	2.5901	2.4989	2.6941	2.7259	3.0336	2.6659
20	STFT	2.6903	2.6847	2.6243	2.7588	2.7948	2.9662	2.6928
	GFT	2.7403	2.7571	2.6696	2.8155	2.8479	3.0739	2.7739

Table 2. Comparison of male speech enhancement performance based on STFT-NMF and GFT-NMF algorithms (STOI)
表 2. 基于 STFT-NMF 和 GFT-NMF 算法的男性语音增强性能比较(STOI)

SNR	方法	N1	N2	N3	N4	N5	N6	N7
−5	STFT	0.6186	0.5492	0.4894	0.6675	0.6902	0.9271	0.6955
	GFT	0.5949	0.5308	0.4715	0.6445	0.6572	0.9075	0.6817
0	STFT	0.7503	0.6996	0.652	0.7927	0.8108	0.9461	0.7813
	GFT	0.7309	0.685	0.6291	0.7724	0.7837	0.9317	0.7757
5	STFT	0.8377	0.8117	0.7759	0.8662	0.8814	0.9554	0.8407
	GFT	0.8241	0.8045	0.7586	0.8507	0.8635	0.945	0.8406
10	STFT	0.8879	0.8787	0.8507	0.9034	0.9182	0.9599	0.8823
	GFT	0.8789	0.8768	0.8384	0.8908	0.9054	0.9515	0.8845
15	STFT	0.9157	0.915	0.8908	0.9216	0.9367	0.9622	0.911
	GFT	0.9094	0.9143	0.8802	0.9092	0.9256	0.9543	0.9128
20	STFT	0.9311	0.9338	0.9119	0.9307	0.9461	0.9634	0.9298
	GFT	0.926	0.9329	0.9009	0.917	0.935	0.9553	0.9305

为了更加直观地反映基于 STFT-NMF 算法以及基于 GFT-NMF 算法的语音增强提升效果，在图 5 中展示干净语音、带噪语音以及基于 STFT-NMF 和 GFT-NMF 算法增强后的语音的时序波形图。从波形图来看，经过噪声污染的语音相比于干净语音存在着明显的差异。结合图 5 可以看出，经 GFT-NMF 算法增强的语音以及经 STFF-NMF 算法增强的语音与干净语音有着极为相似的波形图像，相比于数据可以更加直观地表征：基于 GFT-NMF 算法的语音增强同基于 STFT-NMF 算法的语音增强具有较好的增强性能，均能较好地去除带有噪声语音中的噪声，此外还比较完整地保留带有噪声语音中的干净语音成分，在评价指标上则表现为具有较好的语音质量和可懂度。

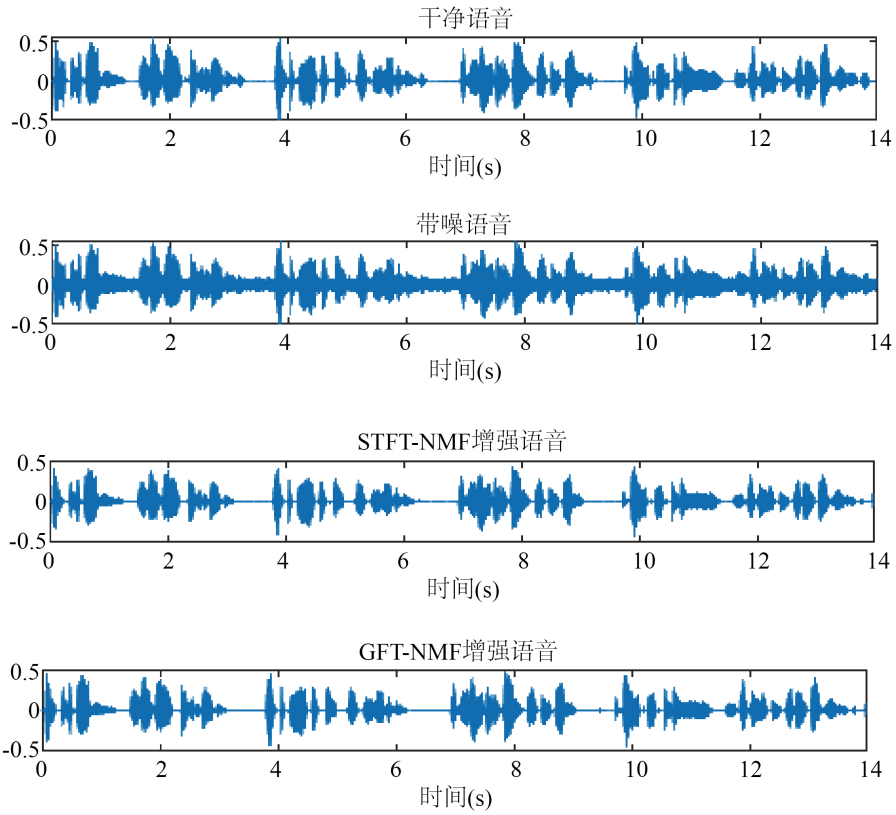

Figure 5. Diagram of waveform of speech signal under different enhancement algorithms
图 5. 不同增强算法下的语音信号波形图

4.2. 基于 GFT-FCNN 的实验结果分析

本实验是基于全卷积神经网络的实验，不同于基于 GFT-NMF 算法的语音增强实验。在本实验中，在训练时中采用了来自 LibriTTS 语音语料库的 4662 条语音作为训练集的干净语音，采用 babble 以及 white 两种噪声根据不同信噪比与干净语音混合添加后分别构成信噪比−5 db、0 db、5 db 的噪声语音。在测试阶段则选择 800 条干净语音根据信噪比加入不同噪声后构成含有噪声语音，对算法模型进行测试。

在表 3、表 4 中分别展示了含有 babble 噪声的语音以及含有 white 噪声的语音在增强前后的 PESQ 以及 STOI 指标评分。经过分析可以得出如下结论：对于受 babble 噪声污染的语音，基于 GFT-FCNN 算法的语音增强的相较于 STFT-FCNN 在 PESQ 指标上有 13.7%的性能提升，在 STOI 指标上则有 3.8%的性能提升；对于受 white 噪声污染的语音，基于 GFT-FCNN 算法的语言增强相较于 STFT-FCNN 在 PESQ 指标上性能持平，在 STOI 指标上则有 4.0%的性能提升。

Table 3. Comparison of the enhancement effects of FFT-FCNN and GFT-FCNN on language containing babble noise
表 3. FFT-FCNN 算法以及 GFT-FCNN 算法对含 babble 噪声语言的增强效果比较

SNR	−5		0		5	
评价指标	PESQ	STOI	PESQ	STOI	PESQ	STOI
未处理	1.3006	0.4841	1.5228	0.5990	1.8096	0.7178
GFT-FCNN	1.2849	**0.5800**	**1.6050**	**0.6487**	**1.9098**	0.7085
STFT-FCNN	1.2435	0.5355	1.3804	0.6226	1.5961	0.7081

Table 4. Comparison of the enhancement effects of FFT-FCNN and GFT-FCNN on language containing white noise
表 4. FFT-FCNN 算法以及 GFT-FCNN 算法对含 babble 噪声语言的增强效果比较

SNR	−5		0		5	
评价指标	PESQ	STOI	PESQ	STOI	PESQ	STOI
未处理	1.0370	0.5843	1.2699	0.6783	1.5804	0.7657
GFT-FCNN	**1.9411**	**0.7751**	**2.2311**	**0.8379**	2.4415	**0.8829**
STFT-FCNN	1.8418	0.7243	2.2444	0.8092	2.5547	0.8667

5. 总结与展望

在本文中，分别研究了基于 GFT-NMF 算法以及 GFT-FCNN 算法的语音增强。实验结果表明，GFT-NMF 算法与 STFT-NMF 算法在语音增强上的效果表现相当，GFT-FCNN 算法相较于 STFT-FCNN 算法在语音增强上则表现出更加优异的性能。未来可以研究更能够表征时序语音节点间关系的邻接矩阵，为基于图傅里叶变换的语音增强算法提高性能。

基金项目

国家自然科学基金；青年科学基金项目；录音回放攻击下说话人识别系统的模型研究 62001100。

参考文献

[1] Wang, Y., Narayanan, A. and Wang, D. (2014) On Training Targets for Supervised Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**, 1849-1858. https://doi.org/10.1109/TASLP.2014.2352935

[2] Williamson, D.S., Wang, Y. and Wang, D.L. (2016) Complex Ratio Masking for Monaural Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**, 483-492. https://doi.org/10.1109/TASLP.2015.2512042

[3] Xu, Y., Du, J., Dai, L.-R. and Lee, C.-H. (2014) An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE Signal Processing Letters*, **21**, 65-68. https://doi.org/10.1109/LSP.2013.2291240

[4] Weninger, F., *et al*. (2015) Speech Enhancement with LSTM Recurrent Neural Networks and Its Application to Noise-Robust ASR. *International Conference on Latent Variable Analysis and Signal Separation*, Liberec, 25-28 August 2015, 91-99. https://doi.org/10.1007/978-3-319-22482-4_11

[5] Zhang, X.-L. and Wang, D.L. (2016) A Deep Ensemble Learning Method for Monaural Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**, 967-977. https://doi.org/10.1109/TASLP.2016.2536478

[6] Tan, K. and Wang, D. (2018) A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. *Proceedings Interspeech*, Hyderabad, 2-6 September 2018, 3229-3233. https://doi.org/10.21437/Interspeech.2018-1405

[7] Takahashi, N., Goswami, N. and Mitsufuji, Y. (2018) MMDenseLSTM: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation. 2018 16*th International Workshop on Acoustic Signal Enhancement* (*IWAENC*), Tokyo, 17-20 September 2018, 106-110. https://doi.org/10.1109/IWAENC.2018.8521383

[8] Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y. and Takeuchi, D. (2020) Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*),

Barcelona, 4-8 May 2020, 181-185. https://doi.org/10.1109/ICASSP40776.2020.9053214

[9] Wang, D.L. and Lim, J.S. (1982) The Unimportance of Phase in Speech Enhancements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **30**, 679-681. https://doi.org/10.1109/TASSP.1982.1163920

[10] Paliwal, K., Wjcicki, K. and Shannon, B. (2011) The Importance of Phase in Speech Enhancement. *Speech Communication*, **53**, 465-494. https://doi.org/10.1016/j.specom.2010.12.003

[11] Mowlaee, P., Saeidi, R. and Stylianou, Y. (2016) Advances in Phase-Aware Signal Processing in Speech Communication. *Speech Communication*, **81**, 1-29. https://doi.org/10.1016/j.specom.2016.04.002

[12] Zheng, N.J. and Zhang, X.-L. (2019) Phase-Aware Speech Enhancement Based on Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**, 63-76. https://doi.org/10.1109/TASLP.2018.2870742

[13] Hu, Y., Liu, Y., Lv, S., *et al*. (2020) DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. *Interspeech* 2020, 21*st Annual Conference of the International Speech Communication Association*, Shanghai, 25-29 October 2020, 2472-2476. https://doi.org/10.21437/Interspeech.2020-2537

[14] 季华忠. 基于图傅里叶变换和神经网络的声信标信号识别方法研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2021. https://doi.org/10.27461/d.cnki.gzjdx.2021.002063

[15] 徐琪. 基于全卷积神经网络和 DenseNet 的语音增强算法研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2022. https://doi.org/10.27251/d.cnki.gnjdc.2022.000133

[16] 柏浩钧, 张天骐, 刘鉴兴, 叶绍鹏. 联合精确比值掩蔽与深度神经网络的单通道语音增强方法[J]. 声学学报, 2022, 47(3): 394-404. https://doi.org/10.15949/j.cnki.0371-0025.2022.03.009

[17] Zen, H.G., *et al*. (2019) LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *Interspeech* 2019, Graz, 15-19 September 2019, 1526-1530. https://doi.org/10.21437/Interspeech.2019-2441