计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):1−9 [doi: 10.15888/j.cnki.csa.008541]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CycleGAN 的语音可懂度关键技术①

肖　晶[1,2], 刘佳奇[1,2], 李登实[3], 赵兰馨[3], 王前瑞[3]

[1](武汉大学 计算机学院 国家多媒体软件工程技术研究中心, 武汉 430072)
[2](武汉大学 多媒体与网络通信工程湖北省重点实验室, 武汉 430072)
[3](江汉大学 人工智能学院, 武汉 430056)
通信作者: 肖　晶, E-mail: jing@whu.edu.cn

摘　要: 语音可懂度增强是一种在嘈杂环境中再现清晰语音的感知增强技术. 许多研究通过说话风格转换 (SSC) 来增强语音可懂度, 这种方法仅依靠伦巴第效应, 因此在强噪声干扰下效果不佳. SSC 还利用简单的线性变换对基频 (*F*0) 的转换进行建模, 并且只映射很少维的梅尔倒谱系数 (MCEPs). 因为 *F*0 和 MCEPs 是语音的两个重要特征, 对这些特征进行充分的建模是非常必要的. 因此本文进行了一个创新性研究即通过连续小波变换 (CWT) 将 *F*0 分解为 10 维来描述不同时间尺度的语音, 以实现 *F*0 的有效转换, 而且使用 20 维表示 MCEPs 实现 MCEPs 的转换. 除此之外, 还利用 iMetricGAN 网络来优化强噪声中的语音可懂度指标. 实验结果表明, 提出的基于 CycleGAN 使用 CWT 和 iMetricGAN 的非平行语音风格转换方法 (NS-CiC) 在客观和主观评价上均显著提高了强噪声环境下的语音可懂度.

关键词: 深度学习; 可懂度增强; 连续小波变换; iMetricGAN; CycleGAN

引用格式: 肖晶,刘佳奇,李登实,赵兰馨,王前瑞.基于 CycleGAN 的语音可懂度关键技术.计算机系统应用,2022,31(6):1–9. http://www.c-s-a.org.cn/1003-3254/8541.html

Key Technologies of Speech Intelligibility Based on CycleGAN

XIAO Jing[1,2], LIU Jia-Qi[1,2], LI Deng-Shi[3], ZHAO Lan-Xin[3], WANG Qian-Rui[3]

[1](National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China)
[2](Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China)
[3](School of Artificial Intelligence, Jianghan University, Wuhan 430056, China)

Abstract: Speech intelligibility enhancement is a perceptual enhancement technique for clean speech reproduced in noisy environments. Speaking style conversion (SSC) is used in many studies to achieve speech intelligibility, which relies solely on the Lombard effect and thus demonstrates poor performance with strong noise interference. In addition, the SSC method models the conversion of fundamental frequency (*F*0) with a straight forward linear transform and only maps Mel-frequency cepstral coefficients (MFCCs) with few dimensions. As *F*0 and MFCCs are critical aspects of hierarchical intonation, adequate modeling of these features is essential. Therefore, we use the continuous wavelet transform (CWT) to decompose *F*0 into ten dimensions to describe speech at different time scales for effective *F*0 conversion and represent MFCCs with 20 dimensions for MFCC conversion. Furthermore, we utilize an iMetricGAN to optimize speech intelligibility metrics in strong noise. The experimental results show that in objective and subjective evaluations, the proposed non-parallel speech style conversion method using CWT and iMetricGAN based on CycleGAN (NS-CiC) significantly increases speech intelligibility in robust noise environments.

Key words: deep learning; intelligibility enhancement; continuous wavelet transform (CWT); iMetricGAN; CycleGAN

① 基金项目: 国家重点研发计划 (1502-211100026)
收稿时间: 2021-09-14; 修改时间: 2021-10-14; 采用时间: 2021-10-29; csa 在线出版时间: 2022-05-26

语音可懂度增强是在嘈杂环境中再生干净语音(没有噪音)的感知增强技术, 常应用在移动设备通信时. 近几十年来, 语音可懂度增强[1]技术引起了广泛的关注. 最近的理论发展表明基于掩码原理和数字信号处理算法增强语音可懂度使语音自然度下降. 一些常见的提升语音可懂度的方法有修改谱特性[2], 非线性放大如范围压缩[3], 选择性的增强某些信号成分如文献 [4], 或语音调制[5], 以及时间尺度的修改[6].

说话风格转换 (SSC)[7] 是一种数据驱动方法, 该方法基于一种被称为伦巴第效应[8] 的特殊发声效果, 旨在改变给定语音信号的风格同时保持说话者的声学特征, 以增强语音可懂度. SSC 通过参数化方法利用数据来对正常风格语音到伦巴第风格语音转换进行学习和建模, 可以对语音转换进行更全面的处理, 同时保证转换后的语音的质量和自然度. 本文遵循同样的策略使用一种基于 WORLD 声码器[9] 的参数化方法.

目前主流的 SSC 方法分为平行 SSC 和非平行 SSC. 对于平行 SSC, 以前的方法利用基于平行数据的学习技术如高斯混合模型 (GMM)[10], 深度神经网络 (DNN) 架构[11] 和 RNN 架构[12], 该技术依赖于有效的源 (正常) 语音和目标 (伦巴第) 语音平行语音对. 然而, 在伦巴第反射下人说话的语速通常较慢. 平行 SSC 需要利用时间对齐操作对训练数据进行预处理这将会导致一些特征失真, 因此更加推荐使用非平行 SSC 来避免时间对齐操作. 最近的一些研究已经结合了循环一致的生成对抗网络 (CycleGANs) 来学习生成接近目标的分布, 该方法不需要平行的语音对. 通过使用 CycleGAN 得到了非平行 SSC, 该方法生成的语音比平行 SSC 生成的语音有更好的可懂度和自然度.

但是, 仍然存在两个主要的两个局限性:

(1) 基频 (*F*0) 是一维特征. ① *F*0 既受短时依赖关系的影响, 也受长时依赖关系的影响. ② 如图 1 所示, 由于清音和浊音的存在, 导致 *F*0 是不连续的. 另一个更重要的特征即梅尔倒谱系数 (MCEPs) 是一种高维连续的特征. 文献 [13] 将 MCEPs 和一维的 *F*0 一起映射会导致分布混合, 而且其提取了 40 维 MCEPs, 但是只用 10 维进行特征映射不能代表完整的伦巴第风格.

(2) 伦巴第效应在不同噪声环境中性能方面存在局限性, 仅具有伦巴第效应的 SSC 在信噪比非常低的强噪声干扰中效果不佳, 尤其在 SNR≤0 dB 的情况下.

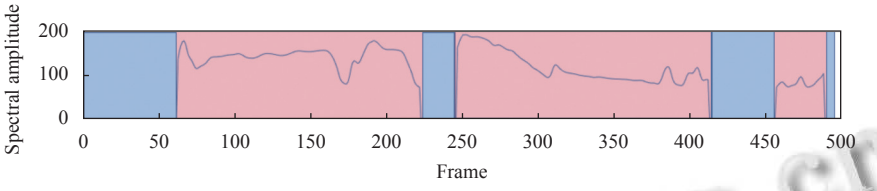


图 1 *F*0 的频谱图, *F*0 是一维不连续特征 (蓝色区域表示清音帧, 红色区域表示浊音帧)

为了克服第 1 个局限性, 提出使用连续小波变换 (CWT) 映射 *F*0 的时间相关性, 并使用 20 维 MCEPs 特征全面地表示声学特征. CWT 被用来描述多个时间尺度上的语音参数, 实际上, 其已经应用在许多领域中, 例如文献 [14].

为了克服第 2 个局限性, 提出使用 iMetricGAN[15], iMetricGAN 是一个生成对抗网络 (GAN) 系统, 由生成器和判别器组成, 生成器作为可懂度增强模块用来增强语音信号, 判别器用来学习预测生成器生成语音的可懂度分数. 与一般的 GAN 网络不同, iMetricGAN 的判别器不是用来鉴别真假, 而是作为一个学习代理尽可能地接近可懂度指标, 然后可以在这个代理的指导下适当地训练生成器.

本文的主要贡献包括: 1) 提出了一个基于 CycleGAN 使用 CWT 和 iMetricGAN 的非平行 SSC 框架, 称为 NS-CiC; 2) 针对线性 *F*0 缺乏时间相关性的问题, 使用 CWT 变换将低维的 *F*0 处理成 10 维的 CWT 系数并映射了更高维的 MCEPs 特征; 3) 利用 iMetricGAN 方法优化对抗生成网络 (GANs) 的语音可懂度指标. 实验结果表明, NS-CiC 在客观和主观评价上均显著提高了强噪声环境下的语音可懂度.

1 参照方法: 非平行 SSC 框架

1.1 基础框架

研究基准[13] 的非平行 SSC 框架如图 2 所示. 源语音 (输入) 是正常风格的语音, 而目标语音 (输出) 是内容相同的伦巴第风格语音. 非平行 SSC 框架主要由 3 部分组成: 声码器分析器、特征映射和声码器合成器. 首先, 通过声码器分析器从输入信号中提取语音特征. 然后, 利用映射系统对与 SSC 密切相关的特征进行

转换. 最后, 将映射的和未转换的特征输入声码器, 声码器合成伦巴第风格的目标语音.

1.2 CycleGAN

CycleGAN 包含 3 部分损失: 对抗损失、循环一致性损失和身份映射损失, 通过这 3 个损失函数来学习源数据和目标数据之前的正向和反向映射. 对于正向映射, 其定义为:

$$L_{\mathrm{ADV}}(G_{X\to Y},D_Y,X,Y)=\mathbb{E}_{y\sim P(y)}[D_Y(y)]+\mathbb{E}_{x\sim P(x)}[\log(1-D_Y(G_{X\to Y}(x))] \quad (1)$$

转换后的数据与目标数据的分布越接近, L_{ADV} 就越小. 为了保证 X 与 $G_{X\to Y}$ 的上下文信息一致, 定义了循环一致性损失函数如下:

$$L_{\mathrm{CYC}}(G_{X\to Y},G_{Y\to X})=\mathbb{E}_{x\sim P(x)}[\|G_{Y\to X}(G_{X\to Y}(x))-x\|_1]+\mathbb{E}_{y\sim P(y)}[\|G_{X\to Y}(G_{Y\to X}(y))-y\|_1] \quad (2)$$

该损失促使 $G_{X\to Y}$ 和 $G_{Y\to X}$ 通过循环转换找到 (x, y) 的最佳伪对. 为了在没有任何额外处理的情况下保留语言信息, 引入了如下身份映射损失函数:

$$L_{\mathrm{ID}}(G_{X\to Y},G_{Y\to X})=\mathbb{E}_{x\sim P(x)}[\|G_{Y\to X}(x)-x\|]+\mathbb{E}_{y\sim P(y)}[\|G_{X\to Y}(y)-y\|] \quad (3)$$

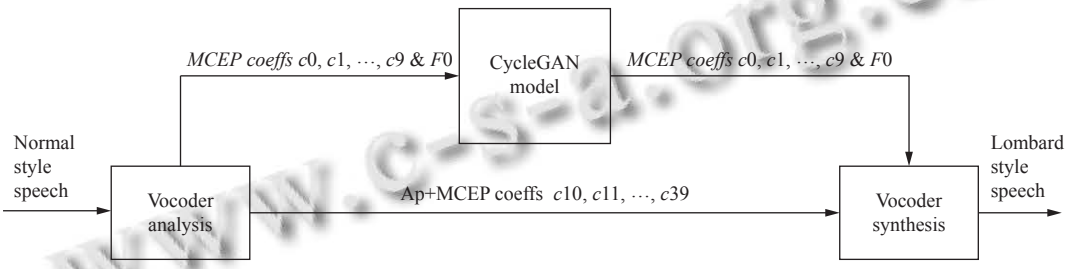


图 2 非平行 SSC 示意图

2 基于 CycleGAN 使用 CWT 和 iMetricGAN 的非平行 SSC

文献 [13] 仅映射不连续的 1 维 $F0$ 并将其与连续的 10 维 MCEPs 一起训练, 而且该系统不擅长处理强噪声的干扰. 为了克服这些局限性, 提出了 NS-CiC 方法.

2.1 基础框架

提出的框架的整个过程如图 3 所示, WORLD 声码器以 5 ms 的帧偏移执行分析和合成. 基于语音的两个重要特性, 将整个框架分为两部分即 CWT 模块和 iMetricGAN 模块. 首先用声码器提取语音信号的 $F0$ 和 MCEP 特征. 然后整个系统使用 CycleGAN 作为基础映射模型, 声码器提取的两个特征通过 CWT 模块和 iMetricGAN 模块的过程分别进行映射. 最后, 预测的特征和转变的特征一起作为声码器的输入以合成增强的语音.

在 CWT 模块中, $F0$ 是一维特征. 因为语音中存在浊音/清音, 因此从 WORLD 声码器中提取的 $F0$ 特征是不连续的. 由于 CWT 对 $F0$ 的不连续性很敏感, 对 $F0$ 进行以下预处理步骤: 1) 在清音区域上进行线性插值; 2) 将 $F0$ 从线性转换为对数; 3) 将 $F0$ 标准化使其均值为 0 方差为 1. 我们首先用 CWT 将 1 维的 $F0$ 分解为 10 维并将不连续的 $F0$ 插值为连续特征. 然后用 CycleGAN 网络来映射源训练数据和目标训练数据, 这些数据均来自于同一个人, 但是由不同的风格和能量组成. 最后, 用 CWT 逆变换将映射的 10 维 $F0$ 变换成 1 维.

在 iMetricGAN 模块中, 用 40 维的 MCEPs 表示频谱包络, 前 20 维 MCEPs ($c0$–$c19$) 作为训练数据. 在第 1 阶段, 使用 CycleGAN 通过对抗损失和循环一致性损失同时学习正向和反向映射, 经 CycleGAN 预测, 正常语音的 MCEPs 转变为伦巴第特征. 在第 2 阶段, 使用 iMetricGAN 增强在强噪声环境中的可懂度. 将预提取的噪声特征和变换后的 MCEP 一起作为 iMetricGAN 模型的输入, 从而获得增强的 MCEPs.

2.2 连续小波变换 (CWT) 模块

小波变换为信号提供了一种易于理解的可视化表示. 使用 CWT, 可以将信号分解为不同的时间尺度.

我们注意到 CWT 已经成功地应用于语音合成[16]和语音转换[17].

给定一个有界的连续信号 k_0, 其 CWT 用 $W(k_0)(\tau,t)$ 表示可以写成:

$$W(k_0)(\tau,t)=\tau^{-1/2}\int_{-\infty}^{+\infty}k_0(x)\psi\left(\frac{x-t}{\tau}\right)dx \quad (4)$$

其中, ψ 是 Mexican-hat 母小波函数. 原始信号 k_0 可以通过逆变换从小波表示 $W(k_0)$ 中恢复, 如下:

$$k_0(t)=\int_{-\infty}^{+\infty}\int_{0}^{+\infty}W(k_0)(\tau,x)\tau^{-5/2}\psi\left(\frac{t-x}{\tau}\right)dxd\tau \quad (5)$$

但是, 如果关于 $W(k_0)$ 的所有信息都不可用, 则逆变换是不完整的. 在这项研究中, 将分析器固定在

10 个离散的尺度上, 相隔一个八度. 分解如下:

$$W_i(k_0)(t) = W_i(k_0)\left(2^{i+1}\tau_0, t\right)(i+2.5)^{-5/2} \quad (6)$$

逆变换得到的 k_0 近似为:

$$k_0(t) = \sum_{i=1}^{10} W_i(k_0)(t)(i+2.5)^{-5/2} \quad (7)$$

其中, i=1, ⋯, 10, τ_0=5 ms, 这些最初是在文献 [18] 中提出的. 因为语音的韵律在不同的时间维度上的表现是不同的. 图 4(b) 给出了 10 个维度的 CWT 分量, 而图 4(a) 只能表示一维的不连续特征. 通过多维度表示, 低维度可以捕捉到短期变化, 高维度可以捕捉到长期变化. 通过这种方式, 能够对 $F0$ 从细致的韵律级到整体语音级进行风格转换.

2.3 iMetricGAN 模块

iMetricGAN 模型框架如图 5 所示. 它由一个生成器 (G) 网络和一个判别器 (D) 网络组成. G 接收语音 s 特征和噪声 w 特征, 生成增强的语音. 生成的语音表示为 $G(s, w)$. 给定 s 和 w, D 被用于预测增强语音 $G(s, w)$ 的可懂度分数. D 的输出表示为 $D(G(s, w), s, w)$, 并希望它能接近通过特定方法计算得到的真实可懂度分数. 引入了函数 $Q(\cdot)$ 来表示要建模的可懂度指标, 即以比特为单位的高斯信道语音可懂度 (SIIB G-auss)[19] 和扩展短时客观可懂度 (ESTOI)[20](已经取得了最好的表现).

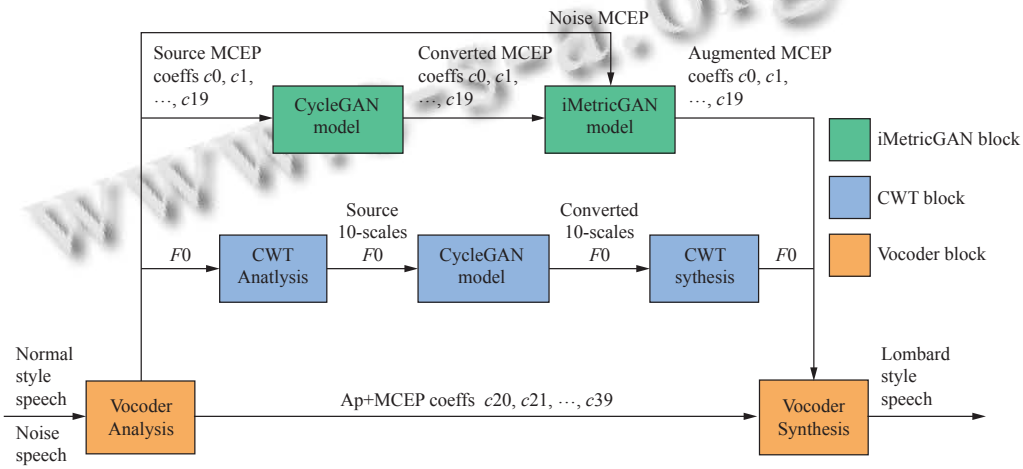
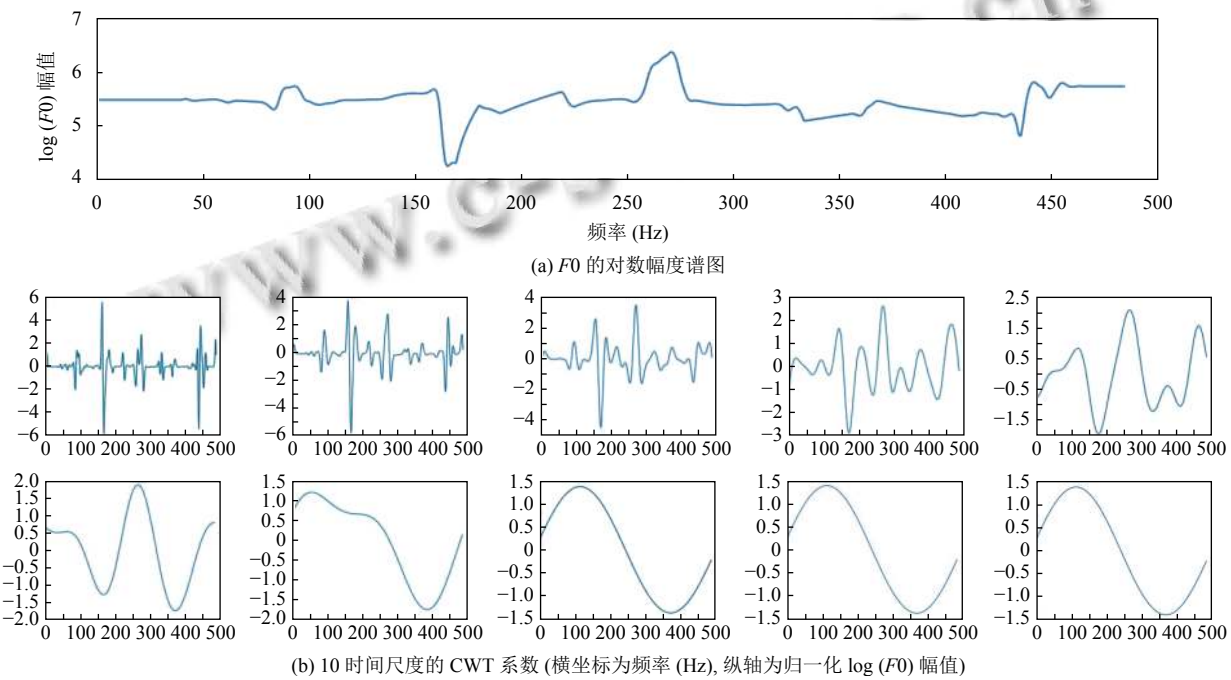


图 3 提出的 NS-CiC 方法示意图



(a) $F0$ 的对数幅度谱图

(b) 10 时间尺度的 CWT 系数 (横坐标为频率 (Hz), 纵轴为归一化 log ($F0$) 幅值)

图 4 一维 $F0$ 与多维 CWT 系数的频谱图

使用上述符号, 图 5(a) 中 D 的训练目标可以表示为最小化以下损失函数:

$$L_D = \mathbb{E}_{s,w}\left[(D(G(s,w),s,w) - Q(G(s,w),s,w))^2\right] \quad (8)$$

在 D 训练的过程中, 损失函数被扩展到式 (9):

$$L_D = \mathbb{E}_{s,w}[D(G(s,w),s,w) - Q(G(s,w),s,w))^2 + (D(\hat{s},s,w) - Q(\hat{s},s,w))^2] \quad (9)$$

可以把式 (9) 视为有辅助知识的损失函数, 式 (8) 视为有零知识的损失函数. 请注意, $\hat{s}$不应被视为真实语音特征或训练标签. G 的训练过程如图 5(b) 所示, 其中, D 的参数是固定的, 通过训练 G 以达到尽可能高的可懂度分数. 为了实现这一点, 将式 (10) 中的目标分数 t 设置为可懂度指标的最大值.

$$L_G = \mathbb{E}_{s,w}\left[(D(G(s,w),s,w) - t)^2\right] \quad (10)$$

G 和 D 迭代训练, 直到收敛. G 充当增强模块, 通过训练 G 来欺骗 D 以获得更高的可懂度分数. 另一方面, D 试图不被欺骗并准确评估转变后的语音的分数. 这个极大极小博弈最终使 G 和 D 都有效. 因此, 通过 G 可以提高输入语音的可懂度.

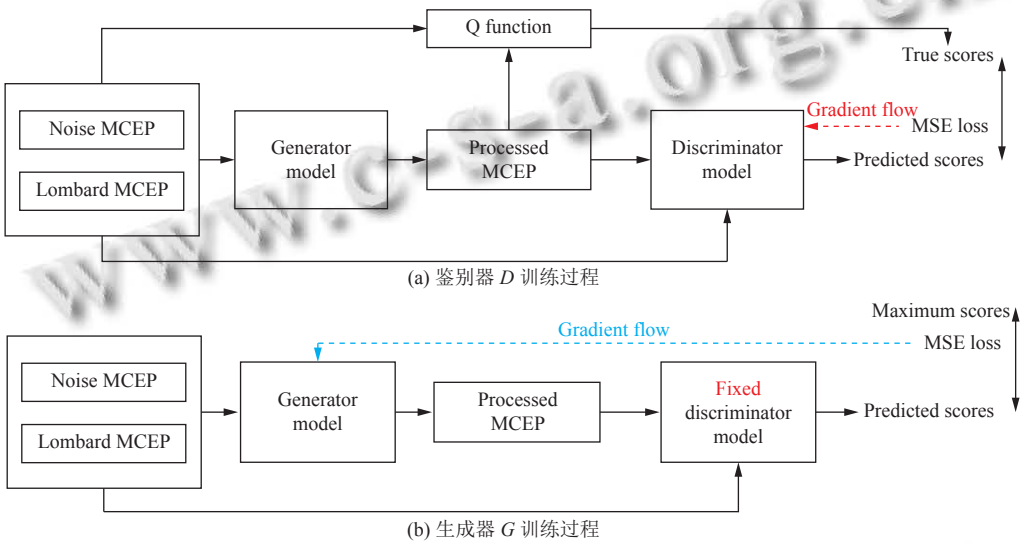


图 5 iMetricGAN 的网络框架及其训练过程

3 实验分析

3.1 实验设置

数据集: 选择了最新的开源伦巴第语料库 (不使用视频数据) Lombard Grid[21] 作为数据集, 该语料库包含 30 名女性说话者和 24 名男性说话者, 每个人以 16 kHz 采样率记录了 50 个正常风格的语音和 50 个伦巴第风格的语音. 语料库的 2/3 和 1/3 分别用于训练和评估. 我们选择了一个德语语料库[22] 作为测试集, 该语料库包含来自 8 个说话者的 40 句话, 每句话具有 3 种不同的伦巴第风格 0、55、70.

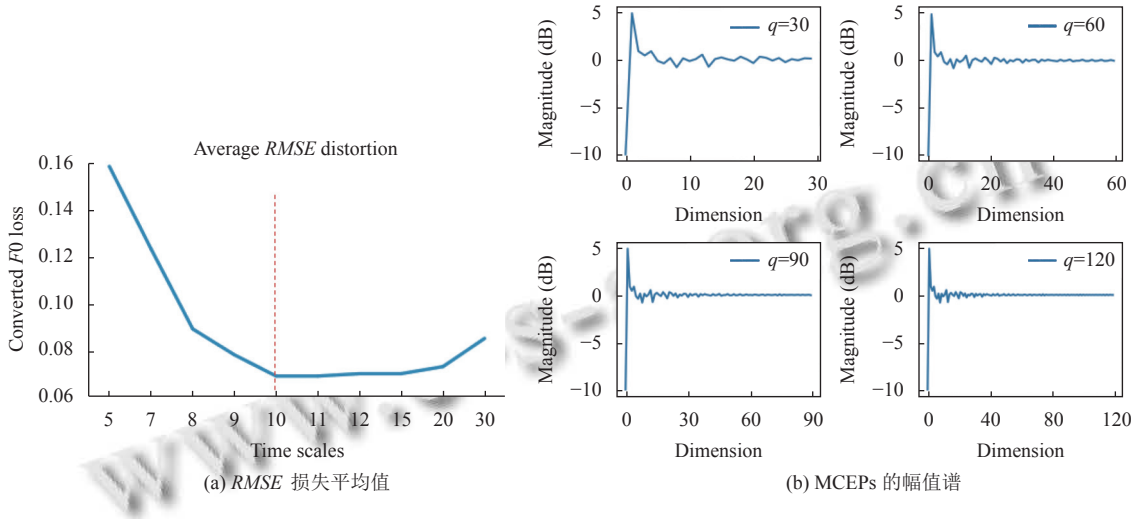
环境噪音: 参考相关研究, 从 NOISEX-92 数据库[23] 中选取 3 种类型的噪声进行实验. 这 3 种噪音分别是 Factory1 (非平稳)、Factory2 (非平稳) 和 Volvo (平稳). 我们将 Volvo 噪声的信噪比设置为−25 dB 和−15 dB, Factory1 和 Factory2 噪声的信噪比设置为−5 dB 和 0 dB.

对比方法: 我们设置了 4 组对比实验, 分别是 CycleGAN 方法, 最新的 IISPA 方法[24], 提出的 10 维 MCEPs 即 NC-CiLC 和 20 维 MCEPs 即 NS-CiHC 方法. 为了验证语音可懂度增强的效果, 在实验中还加入了正常语音.

提出的方法的实现: 在 CycleGAN 中, 参数的设置与参照方法相同. 如图 5 所示, D 的输入特征是 3 通道频谱图 (即已处理、未处理、噪声). D 由 5 层具有以下数量的过滤器和卷积核大小: [8 (5, 5)]、[16 (7, 7)]、[32 (10, 10)]、[48 (15, 15)] 和 [64 (20, 20)] 的 2 维 CNN 组成, 其中每层都带有 LeakyReLU 激活函数. 全局平均池化之后是最后一个 CNN 层, 该层产生固定的 64 维特征. 再连续添加两个带有 LeakyReLU 激活函数的全连接层, 这两层分别有 64 个节点和 10 个节点. D 的最后一层是全连接层, 它的输出代表可懂度指标的分数. 因此, 最后一层的节点数等于可懂度指标的节点数. 我们将 ESTOI 评分标准化使其范围为 0 到 1, 因此, 用 ESTOI 评分训练时, 最后一层使用 Sigmoid 激活函数.

超参数设置: CWT 时间尺度设置: 将 $F0$ 分解为不

同时间维度的 CWT 系数, 并分别比较 CWT 系数的权重、对特征变换的影响程度以及实现的复杂度. 如图 6(a) 所示, 就特征变换的质量和复杂性而言, 发现将 *F*0 分解成 10 维的 CWT 系数更合适. MCEPs 维度设置: 图 6(b) 为同一句子不同发声方式的 MCEPs 幅值谱 (*q*=40). 将不同维度的 MCEPs 特征进行比较, 在大约 $q \leqslant 20$ 时可以看到明显的数值差异, 而在 20 维后 MCEP 值都在 0 左右, 从粗粒度语音频谱图中看不出差异. 通过实验, 选择用 40 维表示频谱特征并选择前 20 维作为映射特征.



(a) *RMSE* 损失平均值 (b) MCEPs 的幅值谱

图 6 不同参数维度特征值的损失系数

客观实验设置: 通过计算 *F*0 的均方根误差 (*RMSE*)[25] 评估 *F*0 映射的性能, 并计算 SIIB Gauss 分数和 ESTOI 分数来证明在强噪声环境中通过 iMetricGAN 模型提高了可懂度, 从而进行客观评价. 最后, 比较了所有方法的整体性能.

使用 *RMSE* 代表 *F*0 预测的性能. 转换后的 *F*0 和相应的目标 *F*0 的 *RMSE* 被定义为:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F0_i^c - F0_i^t\right)^2} \tag{11}$$

其中, $F0_i^c$ 和 $F0_i^t$ 分别表示预测和目标插值 *F*0 特征, *N* 是 *F*0 序列的长度. *RMSE* 值越小代表 *F*0 预测性能越好. 主观实验设置: 主观听力评估采用比较平均意见评分标准 (CMOS)[26]. 研究人员要求听者对两种方法得到的同一段话给出一个从−3 分到 3 分的分数: –3: 很差; –2: 差; –1: 稍差; 0: 差不多; 1: 稍好; 2: 好; 3: 很好. 评分基于可懂度、自然度和舒适度. 20 名参与者在一个消声室中用音频技术公司的 ATH-M50x 耳机进行测试, 参与者听到是被处理过并与噪音混合的语音. 这些参与者的年龄在 18 到 30 岁之间且英语口语流利. 由于听众选择有限, 没有对德语测试集进行主观听力测试. 选择信噪比为−5 dB 和 0 dB 时的 Factory1 和 Factory2 以及信噪比为−15 dB 和−25 dB 时的 Volvo 作为重度和轻度噪声环境. 每位听众测试了 96 条记录: 每句话 3 种方法×(2 名女性+2 名男性)×6 种噪音情景×3 组比较.

3.2 性能测试

*F*0 转换: 如表 1 所示, 女性组的 *F*0 失真比男性组更严重, 这是因为女性的 *F*0 值更高. 最初录制的清晰语音的 *RMSE* 值小于比较不清晰语音的 *RMSE* 值. 从大约提高了 8% 的 SIIB Gauss 分数和 10% 的 ESTOI 分数的可懂度指标可以看出, 与参照方法使用 CycleGAN 联合训练相比, NS-CiC 将 *F*0 分解为 10 维的 CWT 系数具有更乐观的失真分布并且语音可懂度显著提高.

表 1 *F*0 的平均 RMSE 损失

语言	性别	清晰度	CycleGAN	NS-CiC
英语	女性	清晰	0.085	**0.078**
		不清晰	0.097	**0.093**
	男性	清晰	0.089	**0.081**
		不清晰	0.099	**0.095**
德语	女性	清晰	0.052	**0.048**
		不清晰	0.062	**0.057**
	男性	清晰	0.056	**0.050**
		不清晰	0.065	**0.059**

通过 iMetricGAN 增强可懂度: 表 2 给出了不同信噪比下的正常语音、伦巴第语音和经过 iMetricGAN 处理的语音的 SIIB Guass 分数和 ESTOI 分数, 从表中可以看出, 仅仅依靠伦巴第效应在轻微的嘈杂环境中

有很好的提升作用, 但在严重的嘈杂环境中并没有达到增强的效果. 该表显示, 在强噪声环境下, iMetricGAN 方法使 SIIB Guass 分数和 ESTOI 分数分别提高了 23.2% 和 23.6%.

表 2 用 iMetricGAN 增强伦巴第语音的分数

语言	度量指标	SNRs	Normal	Lombard	IISPA	NS-CiHC
英语	SIIB Gauss	Mild	66.0	89.5	88.3	**91.1**
		Severe	34.7	39.7	43.6	**51.7**
	ESTOI	Mild	0.297	0.445	0.433	**0.447**
		Severe	0.183	0.301	0.311	**0.394**
德语	SIIB Gauss	Mild	44.0	143.9	133.4	**144.2**
		Severe	18.6	47.2	45.1	**70.3**
	ESTOI	Mild	0.284	0.442	0.398	**0.444**
		Severe	0.157	0.307	0.355	**0.401**

3.3 整体实验

客观评价: SIIB Guass 和 ESTOI 用于估计说话者和听者之间共享的信息量, 单位为 b/s. 图 7(a)– 图 7(c) 给出了客观评估分数, 因为语音之间的分数非常接近, 所以省略了置信区间. 与正常情况相比, 在严重的噪声环境中, SIIB Guass 分数最多提高 50% 而 ESTOI 分数大概提高了 115%. 在轻度噪声环境中, 由于正常语音本身的清晰度, SIIB Guass 分数提高了 35% 左右, ESTOI 分数提高了 50%. 与 CycleGAN 相比, 我们的方法主要将 SIIB Guass 分数提高了大约 25%, ESTOI 分数提高了 30%. 与最新的 IISPA 方法相比, 我们的方法提升可懂度分平均高 15% 左右. 通过增加 MCEPs 维度, 可懂度方面的提升效果对比参照方法有稳定的提高, 将 SIIB Guass 分数和 ESTOI 分数分别提高约 17% 和 18.6%. 图 7(d)–图 7(f) 显示在德语测试集下增强效果更加明显. 总之, NS-CiHC 比其他几种方法更能提高可懂度.

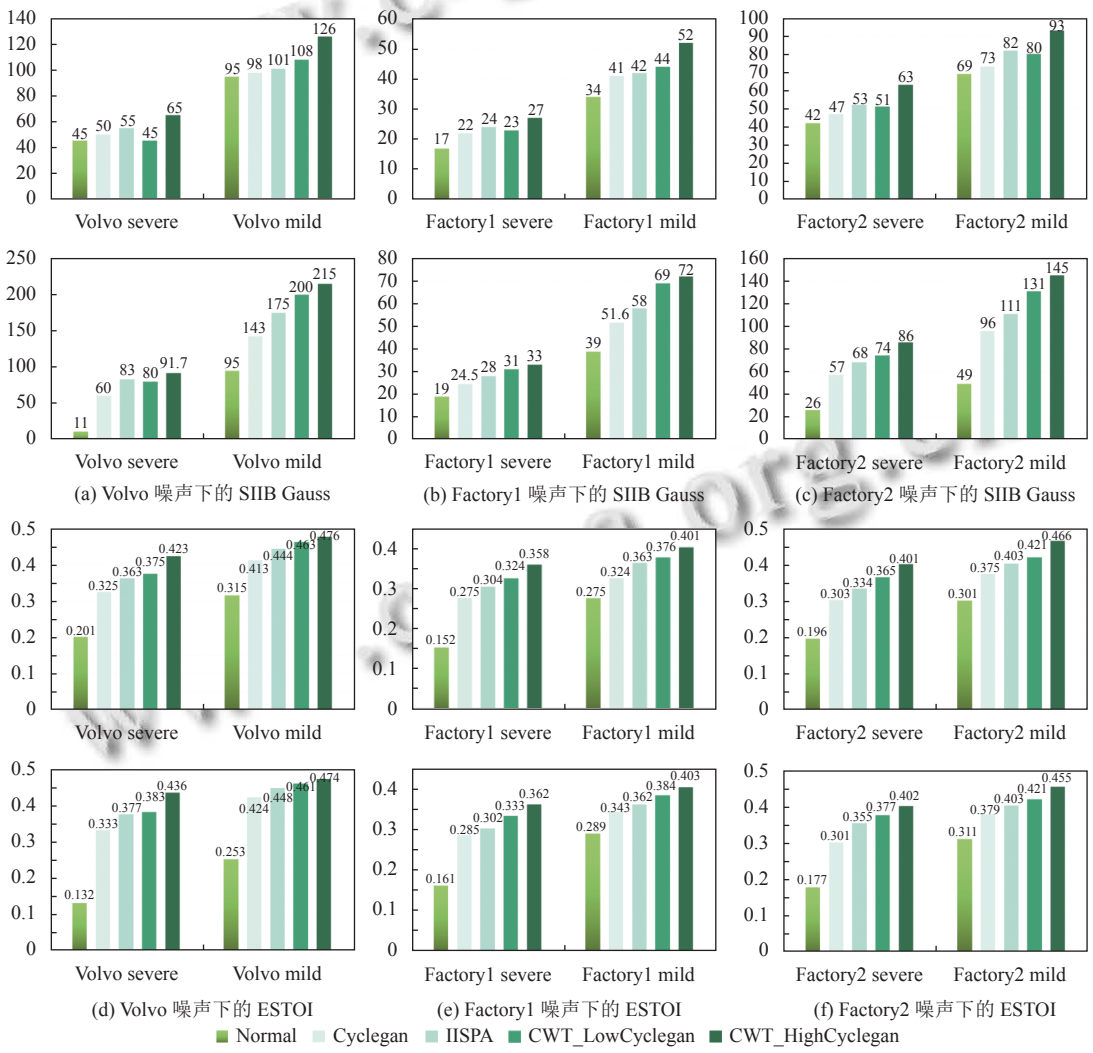


图 7 不同噪声环境下两种数据集可懂度值 (横坐标表示噪声 (dB), 纵坐标表示可懂分 (%))
各子图第 1 行表示英语集结果, 第 2 行表示德语集结果

主观评价: 如图 8 所示, 没有低于 0 的分数, 这意味着提出的方法相对于 CMOS 评分显示, NS-CiC 在质量上明显优于正常语音, 得分约为 1.8 分. 与 CycleGAN 相比, 大约达到了 0.9 分. 与轻度噪声环境相比, 我们的方法在重度噪声环境下的可懂度提升了 10%. 总之, 主观实验的结果与客观实验的表现基本一致.

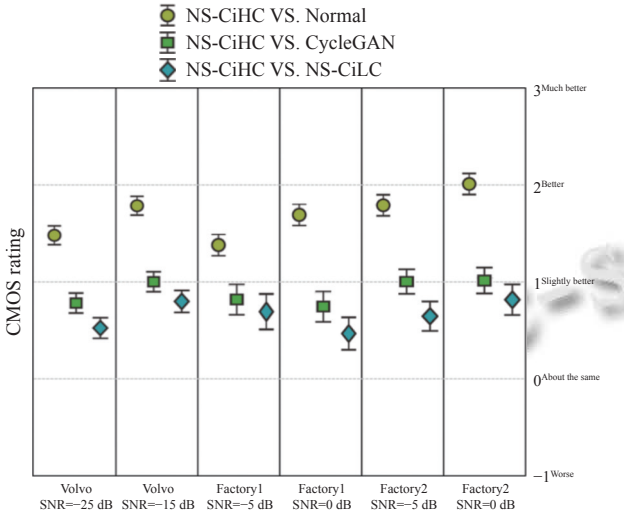


图 8 具有 95% 置信区间的平均 CMOS 评分

3.4 消融实验

我们将提出的 NS-CiC 框架进行了各个模块的实验测试, 如表 3, 仅对 *F*0 的处理能够使语言的音调提升明显, 但声音的能量仍未变化, 可懂度值也提升在 30% 左右, 但在语言的细腻自然方面与 Lombard 更接近; 对 MCEP 的处理我们用到了 iMetricGAN, 语音频谱表示了语音的能量大小, 可以看出在对频谱进行映射后的语音可懂分有大幅度的提升, 而且相较于正常 Lombard 的映射也提升了 10%. 从中我们可以得出 *F*0 对语音的音调有决定性作用, 频谱包络对语音的能量有决定性作用, 本文分别进行 *F*0 的多时间尺度变换与 MCEP 的噪声增强处理能实现语音可懂度的增强.

表 3 不同模块对语音可懂度 SIIB Gauss 分的提升

SNR	Normal	*F*0 conversion	Spectrum conversion
Mild	66.0	75.3	81.8
Severe	34.7	42.6	46.8

4 结论

本文提出了一种高质量非平行的语音风格转换框架, 基于 CycleGAN 进行频谱和韵律转换. 用一种非线性方法研究使用连续小波变换 (CWT) 将 *F*0 分解为 10 维, 从而实现 *F*0 的有效转换, 还使用 20 维 MCEPs 来更全面的表示声学特征从而实现 MCEPs 转换. 考虑到伦巴第效应在强噪声环境下表现不是特别好, 所以使用 iMetricGAN 技术对其进一步增强. 实验结果表明, 提出的框架优于参照方法, 在客观评价中将 SIIB Guass 和 ESTOI 分数分别提高了 25% 和 30%, 在主观评价中将 CMOS 分数提升了 0.9, 有效地增强了语音可懂度.

参考文献

1 Kleijn WB, Crespo JB, Hendriks RC, *et al*. Optimizing speech intelligibility in a noisy environment: A unified view. IEEE Signal Processing Magazine, 2015, 32(2): 43–54. [doi: 10.1109/MSP.2014.2365594]

2 Taal CH, Hendriks RC, Heusdens R. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. Computer Speech & Language, 2014, 28(4): 858–872.

3 Licklider JCR, Pollack I. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. The Journal of the Acoustical Society of America, 1948, 20(1): 42–51. [doi: 10.1121/1.1906346]

4 Arai T, Hodoshima N, Yasu K. Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(7): 1775–1780. [doi: 10.1109/TASL.2010.2052165]

5 Kusumoto A, Arai T, Kinoshita K, *et al*. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. Speech Communication, 2005, 45(2): 101–113. [doi: 10.1016/j.specom.2004.06.003]

6 Aubanel V, Cooke M. Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon: ISCA, 2013. 3592–3596.

7 Paul D, Shifas MPV, Pantazis Y, *et al*. Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020. 1361–1365.

8 Garnier M, Henrich N. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? Computer Speech & Language, 2014, 28(2): 580–597.

9 Morise M, Yokomori F, Ozawa K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems, 2016, E99-D(7): 1877–1884. [doi: 10.1587/transinf.2015EDP7457]

10 Kawanami H, Iwami Y, Toda T, *et al*. GMM-based voice conversion applied to emotional speech synthesis. Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva: ISCA, 2003. 208–211.

11 Seshadri S, Juvela L, Räsänen O, *et al*. Vocal effort based speaking style conversion using vocoder features and parallel learning. IEEE Access, 2019, 7: 17230–17246. [doi: 10.1109/ACCESS.2019.2895923]

12 Ming HP, Huang DY, Xie L, *et al*. Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016: 2453–2457.

13 Seshadri S, Juvela L, Yamagishi J, *et al*. Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 6835–6839.

14 Ribeiro MS, Clark RAJ. A multi-level representation of F0 using the continuous wavelet transform and the discrete cosine transform. Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015. 4909–4913.

15 Li HY, Fu SW, Tsao Y, *et al*. iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020. 1336–1340.

16 Kruschke H, Lenz M. Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis. Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva: ISCA, 2003. 2881–2884.

17 Mishra T, Van Santen J, Klabbers E. Decomposition of pitch curves in the general superpositional intonation model. Proceedings of the 3rd International Conference on Speech Prosody 2006. Dresden: ISCA, 2006.

18 Sisman B, Li HZ. Wavelet analysis of speaker dependent and independent prosody for voice conversion. Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018. 52–56.

19 van Kuyk S, Kleijn WB, Hendriks RC. An evaluation of intrusive instrumental intelligibility metrics. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2153–2166. [doi: 10.1109/TASLP.2018.2856374]

20 Alghamdi A, Chan WY. Modified ESTOI for improving speech intelligibility prediction. Proceedings of 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). London: IEEE, 2020. 1–5.

21 Alghamdi N, Maddock S, Marxer R, *et al*. A corpus of audio-visual Lombard speech with frontal and profile views. The Journal of the Acoustical Society of America, 2018, 143(6): EL523–EL529. [doi: 10.1121/1.5042758]

22 Soloducha M, Raake A, Kettler F, *et al*. Lombard speech database for German language. Proceedings of the 42nd Annual Conference on Acoustics. Aachen, 2016.

23 Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication, 1993, 12(3): 247–251.

24 Schädler MR. Optimization and evaluation of an intelligibility-improving signal processing approach (IISPA) for the Hurricane Challenge 2.0 with FADE. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020. 1331–1335.

25 Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 2014, 7(3): 1247–1250. [doi: 10.5194/gmd-7-1247-2014]

26 Rec IP. 800: Methods for subjective determination of transmission quality. Geneva: ITU, 1996. 22.