E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TN912.3　　**文献标识码**:A　　**文章编号**:1006-8961(2025)05-1272-14

论文引用格式:Wang L B, Jiang Y, Wang T R, Wang X B and Dang J W. 2025. Information disentanglement-based self-supervised learning speech pretrained large model. Journal of Image and Graphics, 30(5):1272-1285(王龙标, 江宇, 王天锐, 王晓宝, 党建武. 2025. 信息解耦式自监督预训练语音大模型. 中国图象图形学报, 30(5):1272-1285)[DOI:10.11834/jig.240607]

信息解耦式自监督预训练语音大模型

王龙标[1*],江宇[1],王天锐[1],王晓宝[1],党建武[2]

1. 天津大学智能与计算学部认知计算与应用重点实验室,天津　300072; 2. 中国科学院深圳先进技术研究院,深圳　518067

摘　要:**目的**　探讨了一种基于语音信息解耦策略的语音预训练大模型,利用海量无标注语音数据提取独立的语言信息、副语言信息和非语言信息,为下游的大语言模型和生成模型提供完备且可控的语音信息,推动言语交互系统的发展。**方法**　提出了一种基于信息解耦的自监督语音表征学习大模型,以高效解耦韵律、说话人及内容特征。在编码器风格的自监督预训练策略基础上,引入两个轻量化模块,增强韵律和说话人特征提取。同时为避免已提取的信息干扰内容信息的学习,模型通过残差机制将其从主分支中剔除,并采用语音掩码预测机制训练主分支,以优化深层特征在语言处理任务中的表现。通过结合多层特征并调整权重,模型能够获取适用于各类下游任务的特定特征。此外,提出的渐进式解码器优化了预训练大模型在语音生成任务中的适应性。**结果**　实验结果表明,本文方法针对不同数量音频训练的两个版本模型(Base和Large)在多项任务中均表现优越。与HuBERT(speech processing universal performance benchmark)模型相比,Base版本在语音识别、说话人验证和情感识别任务中的准确率分别提升5.65%、13.02%和2.43%;Large版本分别提升2.53%、5.76%和1.78%。在情感音色转换任务中,相较于基线模型ConsistencyVC和wav2vec-vc,本文模型在说话人相似度、情感相似度、词错率和感知质量评分等指标上均有所提升,进一步验证了模型的有效性。**结论**　通过将信息解耦思路融入自监督预训练特征提取大模型,有效提升了模型对语音信息的解析与重构能力,为言语交互大模型提供了新的研究视角与实用工具。本文开源代码地址:https://github.com/wangtianrui/ProgRE。

关键词:信息解耦;自监督学习(SSL);语音编解码;言语交互大模型;语音合成

Information disentanglement-based self-supervised learning speech pretrained large model

Wang Longbiao[1*], Jiang Yu[1], Wang Tianrui[1], Wang Xiaobao[1], Dang Jianwu[2]

1. Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin　300072, China; 2. Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen　518067, China

Abstract: **Objective** With the advancement of technology, the demand for large models in speech interaction has increasingly grown in recent years. This paper investigates a self-supervised pretrained large model based on the speech information disentanglement technique, aiming to train model that extracts linguistic, paralinguistic, and nonlinguistic information from speech. This approach leverages vast amounts of unannotated data, enabling the model to learn effectively even in the absence of labels. By achieving the independence of the extracted speech representations, downstream models can clearly

收稿日期:2024-12-31;**修回日期**:2025-02-23;**预印本日期**:2025-03-02

*通信作者:王龙标　longbiao_wang@tju.edu.cn

基金项目:国家自然科学基金联合基金项目(U23B2053);国家自然科学基金面上项目(62176182)

Supported by: National Natural Science Foundation of China (Joint Fund Program) (U23B2053); National Natural Science Foundation of China(General Program)(62176182)

distinguish between different types of information, which is a crucial step in enhancing the accuracy and controllability of speech processing. Specifically, the core of the disentanglement technique lies in effectively extracting and separating different layers of information within speech signals. Linguistic information conveys specific content, while paralinguistic information includes the speaker's emotions, intonation, and other nuances. Nonlinguistic information may encompass the speaker's physiological state, environmental sounds, or background noise. Therefore, by modularizing these types of information, the model not only gains a better understanding of speech content but also allows for flexible adjustments or replacements of these elements as needed. This process provides comprehensive and detailed speech information for downstream language and generative models, considerably enhancing their support for complex verbal interactions. Furthermore, this technique facilitates multitask learning. In speech interaction tasks, different application scenarios have varying demands for speech information. Through disentanglement, the model can adapt to these diverse requirements, achieving higher performance across multiple tasks, such as speech recognition, emotion recognition, and speech synthesis. By leveraging the speech information disentanglement technique, the self-supervised pretrained large model not only offers flexibility for practical applications but also opens up new avenues for further research. **Method** To address this challenge, we propose an information disentanglement-based self-supervised speech representation learning model that effectively leverages vast amounts of unannotated data, achieving high-quality speech information disentanglement. Specifically, we build upon an encoder-style self-supervised learning (SSL) framework and introduce two lightweight specialized modules. Both modules enhance the model's capacity to extract pitch variation and speaker identity from speech signals, which are crucial for achieving expressive and contextually rich speech generation. We employ a residual removal approach that systematically disentangles the extracted pitch variation and speaker information from the main processing branch, thus ensuring that these components do not interfere with the learning of content information. The main branch is then trained using HuBERT's speech masking prediction mechanism, which optimizes the deep layers of the encoder for superior performance in linguistic tasks. This method allows for the progressive extraction and refinement of the representations of pitch variation, speaker identity, and content from the input speech, thereby fostering a more nuanced understanding of the speech signal. Furthermore, we combine the diverse representations obtained from different layers, strategically adjusting their weights to generate task-specific representations tailored for various downstream speech processing applications. Such flexibility is essential for effectively addressing the distinct requirements of tasks such as speech recognition, emotion recognition, and voice conversion. Additionally, we introduce a progressive generator that builds upon these representations, resulting in the seamless execution of downstream speech generation tasks. This comprehensive approach not only enhances the model's adaptability and performance across multiple tasks but also paves the way for more sophisticated and context-aware speech interaction systems. **Result** Experimental results indicate that our proposed method demonstrates significant advantages across various tasks, including speech recognition, speaker verification, speech enhancement, emotion recognition, and voice conversion tasks. Notably, in the disentanglement-based emotion and voice conversion task, our model achieves substantial improvements in emotional similarity, speaker similarity, word accuracy, and audio quality ratings compared to the second best model. These enhancements reflect the model's capability to effectively disentangle and manipulate various speech information components, which, in turn, contributes to a natural and expressive speech output. This result underscores the effectiveness of our approach in not only improving the quality of synthesized speech but also enhancing the controllability of emotional and speaker characteristics, ultimately paving the way for more sophisticated applications in human-computer interaction. **Conclusion** Incorporating information disentanglement into the pretrained extraction model enhances the analysis and synthesis capabilities of speech information. This feature enables the model to clearly identify and manipulate aspects of speech, such as linguistic content, emotional state, and speaker identity. Improved clarity of speech features is crucial for advancing large models focused on verbal interaction. Furthermore, this approach offers insights and practical tools that are applicable in various fields, from enhancing speech recognition to improving emotional expressiveness in generated speech. This method also fosters more nuanced human-computer communication and encourages research into complex speech dynamics, potentially leading to innovations in personalized virtual assistants and adaptive language learning tools and ultimately enriching user experiences in human-computer interactions.

Key words: information disentanglement; self-supervised learning(SSL); speech codec; speech interaction large model; speech synthesis

0 引 言

ChatGPT的出现引发了全球对大模型研究的热潮，大幅改变了自然语言处理和人机交互领域的研究方式，同时也对多模态研究方向产生了影响（严昊等，2023；刘婷婷 等，2021；陶建华 等，2024）。在语音信号处理领域，语音大模型已成为研究热点（许裕雄 等，2024），旨在用一个系统实现语音感知、知识推理和语音回复等功能。例如SpeechGPT（Zhang等，2023）、VioLA（Wang 等，2024b）和AudioPaLM（Rubenstein 等，2023）等模型通过将语音信号转换为离散字符，并训练具有大量参数的GPT（generative pre-trained Transformer）模型来建模文本到语音的转换，进而实现语音识别、翻译和合成。然而，这些模型往往未能充分利用预训练大语言模型的语言知识，甚至在微调后可能会遗忘已有知识。为解决这一问题，计算机视觉领域的模态映射方法（Li等，2023）逐步被应用于语音大模型任务，以便大语言模型能直接理解语音信息，但其性能仍受到预训练语音感知模块的限制。

“海河·谛听”大模型是一款感知—生成全链路言语意图理解模型，如图1所示。该模型通过模态映射实现语音交互任务，其主要由3个部分构成：自监督语音预训练大模型、大语言模型和个性化语音生成模型。自监督预训练模型作为语音编码器，从语音中提取语言信息、副语言信息和非语言信息，并将这些信息传递给大语言模型和语音生成模型，以生成相应的语音回复。其中，自监督预训练大模型是大模型理解语音模态的关键，旨在利用大量未标记数据提取通用表示，以支持各种下游任务。这一方法通过语音生成监督信号进行预训练，降低了对人工标注的需求，提升了数据利用效率。预训练完成后，模型使用标注的监督语音数据进行微调，优化特定任务性能，如语音识别或情感分析。现有的语音自监督学习方法主要分为生成性和对比性两类。生成性方法如Encodec（Défossez 等，2022）、TERA（Transformer encoder representations from alteration）（Liu 等，2021）和SoundStream（Zeghidour 等，2022）通过编码器将语音转换为表示并重建语音，表现优于声学任务，但在内容相关任务上效果较弱。对比性方法如wav2vec（Schneider 等，2019）、HuBERT（Hsu 等，2021）、WavLM（Chen 等，2022）和Data2vec（Baevski 等，2022）则通过测量不同输入间表示的相似性进行训练，适合内容任务，但在声学任务上性能较低。在言语交互大模型中，预训练大模型的普适性愈加重要，强大的下游模型需从预训练大模型中获取全面的语音信息。为此，SUPERB（speech processing universal performance benchmark）（Yang 等，2021）和SUPERB-SG（speech processing universal performance benchmark with semantic and generative capabilities）（Tsai 等，2022）基准测试汇集了多项下

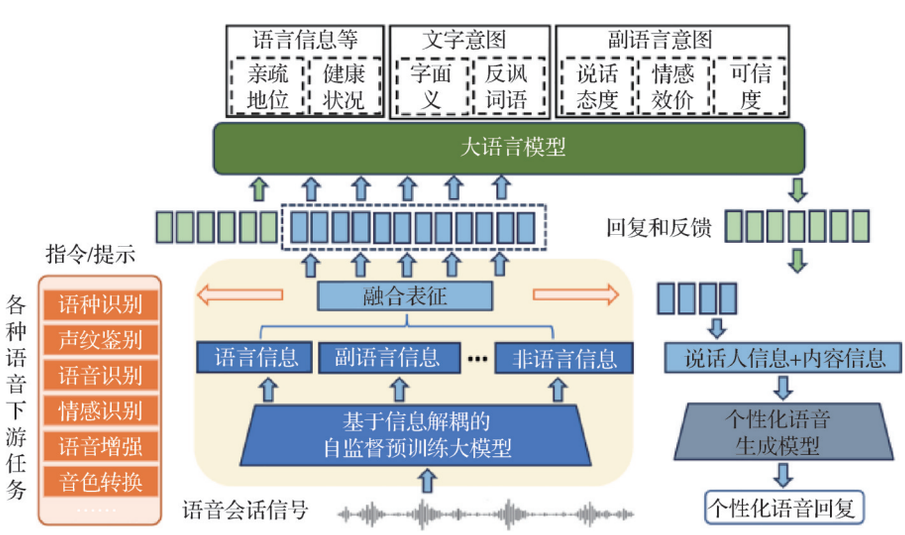


图1 “海河·谛听”言语交互大模型示意图

Fig. 1 Framework of the “Haihe·Diting” speech interaction large model

游任务，评估预训练模型在各领域的表现。

然而，针对特定任务（如说话人识别和情感识别）的自监督学习（self-supervised learning，SSL）策略在提升某一任务能力时，常常导致其他任务的性能下降。这引发了一个挑战性的研究问题，即语音预训练能否同时提升多个任务的性能？解决这一问题将推动多任务学习系统的智能化和高效化。初步研究已集中在结合多个任务特定的预训练模型，或采用基于适配器的多任务预训练。这些方法虽然在并行提取语音表示方面取得了一定成功，但基于专家混合（mixture of experts，MOE）的原理增加了资源需求，并未根本解决通用语音自监督学习的挑战。任务间的不兼容性使模型难以找到共同的收敛方向，这主要源于不同任务对语音信息的不同需求。

现有自监督学习模型呈现任务模块化趋势：Transformer 的浅层表示关注声学信息，而深层表示则强调上下文和语义信息。理论上，语音可以解耦为非语言、副语言和语言信息。实际应用中，语音被解耦为说话人、内容和韵律变化，这三者理论上相互独立，且可以自由组合，以适应不同任务（Wang 等，2024a）。相关研究表明，去除内容信息可以增强说话人识别性能，而去除与内容无关的说话人信息则可改善内容相关任务的表现。基于此，本文提出了一种基于信息解耦的预训练方法，逐步提取语音中的韵律变化、说话人和内容的表征，所提取的表征能够更好地适应不同需求的下游任务，从而实现兼容性效果的提升。

具体而言，本文在自监督模型 HuBERT 的两个中间层中增强了韵律变化和说话人信息的提取。鉴于韵律变化、说话人信息与内容信息在理论上的独立性，本文逐步从主分支中去除这些增强的表征，以减轻模型负担，防止增强信息对其他无关信息（尤其是内容信息）的学习产生干扰。为此，在模型中插入两个轻量级提取器，以渐进的残差方式建模语音的韵律变化和说话人信息，最终通过 HuBERT 的自监督策略训练残差主分支，以预测被掩蔽的单元。在此基础上，针对生成式下游任务，本文构建了一个渐进式语音生成器，以实现语音到语音的全链路转换。实验结果表明，该方法显著提升了各类任务的性能，且可视化结果显示特定层对相应任务的贡献更加明显。这种增强不同层对语音信息的作用的方式结合加权求和机制，也为分析下游任务对各种类型语音信息的需求提供了新的视角。

本文的主要贡献总结如下：1）指出并通过实验验证，预训练策略促进的不同层任务特征及缓解任务不兼容性是实现通用预训练模型的关键；2）提出了一种基于渐进残差提取的语音表示学习预训练方法，使模型能够平衡韵律变化、说话人信息和内容信息的提取，从而在各类下游任务中实现共同性能提升；3）引入了韵律变化和说话人信息的提取器，大大增强了模型提取语调和非语言信息的能力，并在多项下游任务中实现了最先进的性能，尤其是情感音色转换任务上；4）除了在 960 h 数据集上评估所提方法的性能外，还在 84 500 h 的英汉双语大规模预训练任务中验证了其有效性。本文在 https://github.com/wangtianrui/ProgRE 开源了代码。

1 研究方法

本文深入探讨了基于信息解耦的语音自监督预训练大模型，并验证了其在多种下游任务中的可拓展性和有效性。该预训练大模型采用自监督学习方法，在海量无标注语音数据上进行训练，充分挖掘了大规模数据集的潜力。模型设计中融入了渐进式解耦的思想，逐步从语音中提取出包含丰富韵律、说话人及内容信息的语音表征。

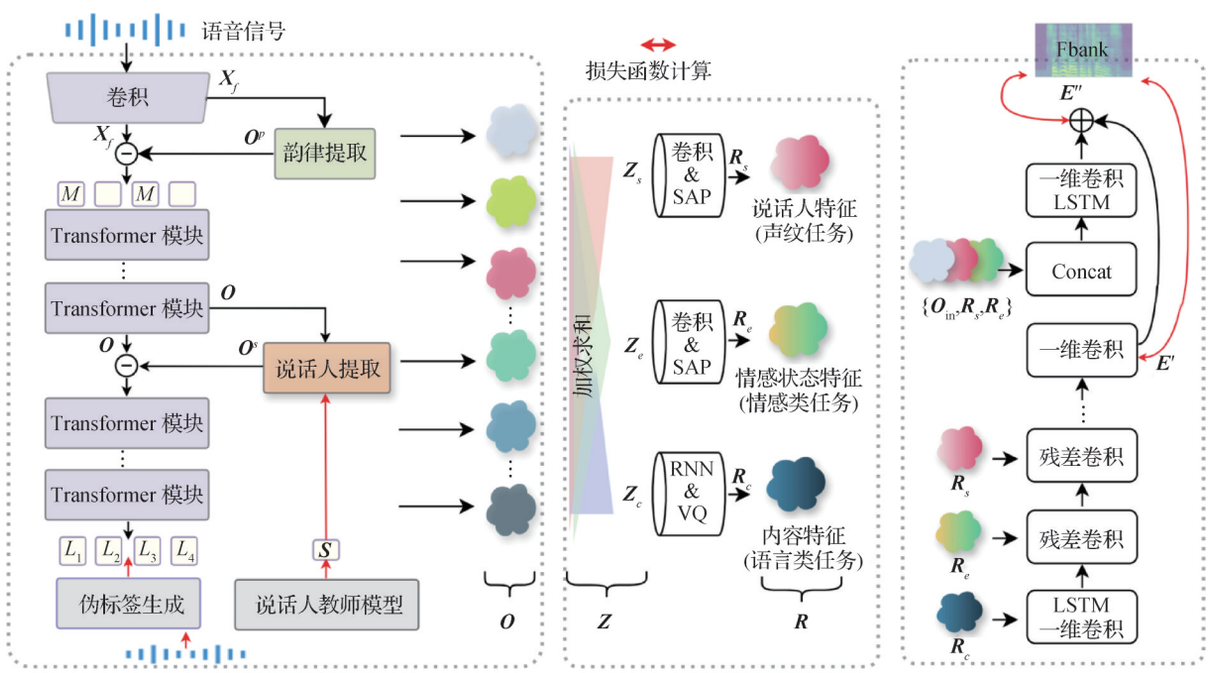
而后，在基于解耦的多任务特征优化模块中引入了可学习的权重，通过加权求和对预训练提取模型的层级输出进行整合，并串接少量参数的下游模块，以提取可针对特定下游任务的说话人信息、情感状态及内容表征。对于生成类任务，如情感音色转换任务，框架中设计了一个渐进式语音生成器，实现了语音生成。具体而言，首先利用渐进式生成器根据提取的 3 种表征重构出一个粗糙的结果，随后，声学补偿器基于编码器浅层的输出对该粗糙结果进行声学细节的补偿，结合声码器实现高质量的语音输出。

1.1 基于信息解耦的语音自监督预训练大模型

本节介绍的用于特征提取的基于信息解耦的语音大模型是本文框架的核心内容。首先，模型基于 HuBERT 的预训练策略，设计旨在使 Transformer 的深层结构专注于内容的深层表征提取，而浅层则保留丰富的声学细节。为了进一步增强模型捕捉韵律波动和说话人特性的能力，本文引入了渐进式残差

提取机制，并将其融入HuBERT框架中。

图2(a)所示的流程展示了语音信号的处理流程。输入的语音信号首先通过一个由7层一维卷积构成的卷积模块，该模块将信号转换为帧级表示$\boldsymbol{X}_{\mathrm{f}}$，每帧的步长为20 ms。随后利用韵律提取器提取并移除了韵律信息$\boldsymbol{O}^{\mathrm{p}}$，得到更加纯净的语音表示$\boldsymbol{X}$。接下来，多层Transformer对这一表示进行编码，得到表示$\boldsymbol{O}$，逐层深入地捕捉语音的语义特征和上下文信息。在Transformer的中间第4层插入了一个说话人提取器，该提取器负责识别并剥离说话人信息$\boldsymbol{O}^{\mathrm{s}}$，使主分支能够更专注于语音内容本身。预训练阶段，对输入Transformer的数据X进行了随机掩码处理，并利用无监督或自监督学习策略为主分支和说话人教师网络生成伪标签。进入微调阶段，为了满足不同下游任务（如语音识别、说话人识别等）的特定需求，本文采用了一种学习权重的加权求和机制，能够根据任务需求智能地调整不同特征（如内容、韵律、说话人特征等）的权重，从而生成最适合的特征表示。



(a) 基于信息解耦的语音自监督预训练大模型　(b) 多任务特征优化模块　(c) 渐进式语音生成器

图2　基于信息解耦的语音自监督预训练大模型示意图及其在下游任务的使用示意

Fig. 2　Schematic diagram of a speech self-supervised pre-training large model based on information disentanglement and its application in downstream tasks((a)information-disentangled self-supervised pre-trained large model for speech; (b)multi-task feature optimization module; (c)progressive speech generator)

为了在提取模型中实现有效的解耦机制，本文将残差矢量量化(residual vector quantization, RVQ)技术拓展至连续表征，并实施一种称为渐进式残差精炼的策略。该策略的核心在于层层剥离表征中的残差细节，其框架设计融入了HuBERT体系之中。在此框架内，本文特别引入了韵律与说话人特征提取器，并将它们从主信息流中剥离，以此促进主路径上的信息经历更为精细的渐进式处理，最终交由HuBERT的自监督学习机制进行调控，这一过程旨在强化模型对内容相关聚类信息的捕捉能力。此外，渐进式残差精炼中的信息解耦效果，只有在所有组件协同训练时方能充分展现，因此，本文采用了HuBERT的自监督框架来统一预训练所有模块。此外，为了确保说话人特征提取器的专业性，本文引入了专注于说话人信息捕获的教师模型进行辅助监督，以进一步提升其性能。而对于韵律提取器，本文则通过输入归一化的韵律数据，引导其专注于提取韵律变化的本质信息，即

$$L = \lambda_{\mathrm{f}} \cdot L_{\mathrm{f}} + \lambda_{\mathrm{s}} \cdot L_{\mathrm{s}} + \lambda_{\mathrm{c}} \cdot L_{\mathrm{c}} \tag{1}$$

式中，L_{f}表示与韵律相关的输出X_{f}的均方误差(mean squared error, MSE)，λ_{f}、λ_{s}和λ_{c}为超参数，L_{s}和L_{c}为另两个损失函数项，分别聚焦于说话人信息

建模与内容信息建模的损失函数。为了在训练过程中实现3个不同损失部分之间的有效平衡，本文设定了超参数λ_{f}、λ_{s}和λ_{c}，其值分别设定为10.0、1.0和1.0。这样的设置旨在根据各损失项的重要性给予不同的权重，确保模型在优化过程中能够全面而均衡地提升其在韵律识别、说话人区分以及内容理解3个方面的贡献。

1.1.1 副语言信息建模

受音色转换中解耦策略的启发，模型聚焦于波形分析，从中提取副语言信息，特别是音高$\boldsymbol{F}_0$，作为语调变动的锚点。韵律提取机制旨在捕捉语调动态信息，同时排除内容与说话人信息的干扰。为此，对每个波形的音高实施对数归一化处理，以确保副语言特征的有效性和可靠性，即

$$\boldsymbol{P}=\frac{\log \boldsymbol{F}_0-\operatorname{mean}\left(\log \boldsymbol{F}_0\right)}{\operatorname{std}\left(\log \boldsymbol{F}_0\right)} \tag{2}$$

式中，$\boldsymbol{F}_0$代表音高，归一化后的音高作为输入，确保副语言信息提取模块的输出准确反映韵律的变化。为此，本文采用轻量级卷积递归架构，如图2所示，其中卷积块（convolution，Conv）包括1D卷积、批量归一化和ReLU激活函数。后通过全连接（fully connected，FC）层进一步精炼输出，精准捕捉韵律变化特征$\boldsymbol{O}^{p}$。由于韵律是从波形中提取的，为避免其对主信号的干扰，通过下式在卷积块之后从主分支中移除韵律变化信息

$$\boldsymbol{X}=layernorm\left(\operatorname{Conv}(\boldsymbol{x})-\boldsymbol{O}^{p}\right) \tag{3}$$

式中，$\boldsymbol{x}$是输入信号，$\boldsymbol{O}^{p}$为韵律变化特征，移除后执行层归一化（$layernorm$），以加速模型的收敛。通过这种方式，模型能够有效地提取和建模副语言信息。

1.1.2 非语言信息建模

在非语言信息建模方面，本文聚焦于说话人信息的提取，采用了帧级说话人表示提取方法，该方法显著区别于传统的语句级处理方式。此帧级提取模块通过自监督学习的掩码预测策略进行预训练，其核心在于让编码器学会预测输入序列中被随机掩盖的信息片段。这一过程极大地增强了模型捕捉双向及全局说话人特征的能力。

说话人提取器由一个FC层、帧级注意力统计（frame-level attention statistics，FAS）和一个输出FC层后的层归一化组成。FAS是帧级注意力统计池化方法，计算每帧的均值和方差。此提取器被嵌入到Transformer架构的第4个模块之后，以精准捕捉并生成说话人表示$\boldsymbol{O}^{s}=\left[\boldsymbol{o}_1^{s},\boldsymbol{o}_2^{s},\cdots,\boldsymbol{o}_T^{s}\right]\in\mathbf{R}^{T\times D}$。除了利用HuBERT的自监督学习策略与主分支同步训练外，本文还引入了一种教师—学生式的指导机制，进一步促使说话人提取器专注于说话人信息的提取。这一机制类似于HuBERT的K-means语音单元概念，利用一个离线自监督预训练模型EMA-DINO（Han等，2023）获得了一个语句级的说话人提取目标$\boldsymbol{s}\in\mathbf{R}^{K}$。该提取器基于ECAPA-TDNN（Desplanques等，2020）模型进行预训练，并且在整个过程中不依赖任何标签，采用了知识蒸馏技术。然后，通过掩码回归训练说话人提取器，计算为

$$L_{\mathrm{s}}=-\frac{1}{T_{\mathrm{m}}}\sum_{t\in T_{\mathrm{m}}}\log\sigma\left(sim\left(\boldsymbol{A}\boldsymbol{o}_t^{s},\boldsymbol{s}\right)\right) \tag{4}$$

式中，$\boldsymbol{A}^{s}\in\mathbf{R}^{D\times K}$是一个投影矩阵，$\boldsymbol{o}^{s}$为说话人表示，$\boldsymbol{s}$为说话人提取目标，$T_{\mathrm{m}}$为掩码帧，$sim(\cdot,\cdot)$表示余弦相似度，$\sigma(\cdot)$表示sigmoid。说话人损失$L_{\mathrm{s}}$仅在掩码帧$T_{\mathrm{m}}$上计算。

1.1.3 语言信息建模

在语言信息建模方面，本文模型主分支的训练过程遵循HuBERT的框架。具体而言，本文采用了基于K-means聚类的类似于BERT（bidirectional encoder representations from transformers）的掩码伪标签预测任务。这一训练目标促使编码器深入学习内容表示，同时允许靠近输入的浅层表示保留更多的声学细节，从而确保语言信息的准确提取。

在预训练之前，模型会对已有预训练模型的梅尔频率倒谱系数（Mel-scale frequency cepstral coefficients，MFCC）或隐藏层表示进行离线聚类，以生成伪标签$\boldsymbol{L}=\left[l_1,l_2,\cdots,l_T\right]$，每个标签$l\in[U]$对应一个$U$类变量，即K-means聚类中的类别变量。正如图2中所示，在预训练阶段，卷积模块的帧级输出会被随机且连续地掩码处理，然后输入至Transformer编码器。在移除韵律变化和说话人信息后，主分支会被训练用于预测被掩码帧的伪标签，具体为

$$L_{\mathrm{c}}=\frac{1}{T_{\mathrm{m}}}\sum_{t\in T_{\mathrm{m}}}\log\frac{\exp\left(sim\left(\boldsymbol{A}^{\mathrm{c}}\boldsymbol{o}_t^{\mathrm{c}},\boldsymbol{e}_u\right)/\tau\right)}{\sum_{u'}^{U}\exp\left(sim\left(\boldsymbol{A}^{\mathrm{c}}\boldsymbol{o}^{\mathrm{c}}t,\boldsymbol{e}_{u'}\right)/\tau\right)} \tag{5}$$

式中，投影矩阵$\boldsymbol{A}^{\mathrm{c}}$用于最后一层Transformer的输出$\boldsymbol{O}^{\mathrm{c}}=\left[\boldsymbol{o}_1^{\mathrm{c}},\boldsymbol{o}_2^{\mathrm{c}},\cdots,\boldsymbol{o}_T^{\mathrm{c}}\right]$，$\boldsymbol{e}_u$是K-means聚类单元$u$的嵌入，$\tau$是对数比例值，设定为0.1。类似于说话人信

息建模中的说话人损失，内容损失L_c同样只在掩码帧上应用。

1.2 基于预训练大模型的下游任务

在基于预训练语音解耦的大模型中实现了逐步提取语音的语言信息、副语言信息和非语言信息，使得各类特征相互独立，从而满足不同下游任务的需求，如语音识别、声纹识别、情感识别和语音生成等。具体而言，引入了多任务特征优化模块，通过加权求和的方式处理来自不同层次的标准化输出，提取出情感分类、说话人识别和内容识别所需的特征。针对生成类任务，本文构建了渐进式语音生成器，利用提取的特征逐步重构语音，实现了从粗糙结果到精细输出的转变。

1.2.1 多任务特征优化模块

预训练的特征提取模型在不同层次上展示了任务特性，靠近输入的浅层保留了丰富的声学细节，而深层则更多地捕捉到语义信息。基于这一特性，模块中选择冻结预训练的语音特征提取模型，并引入可学习的权重$\boldsymbol{W}=\{\omega_1,\omega_2,\cdots,\omega_N\}$。通过类似于SUPERB的加权求和方式，对来自$N$层编码器的标准化输出$\boldsymbol{O}\in\mathbf{R}^{N\times T\times D}=\{\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots,\boldsymbol{O}_N\}$进行处理

$$\boldsymbol{Z}=\sum_{i=1}^{N}\omega_i\cdot layernorm(\boldsymbol{O}_i) \tag{6}$$

式中，$\omega_i\in\boldsymbol{W},\boldsymbol{O}_i\in\boldsymbol{O}$，通过3个不同的权重$\{\boldsymbol{W}_e,\boldsymbol{W}_s,\boldsymbol{W}_c\}$来提取用于情感分类、说话人识别和内容识别任务的特定特征$\{\boldsymbol{Z}_e,\boldsymbol{Z}_s,\boldsymbol{Z}_c\}$。对于语句级别的情感状态和说话人信息，模块结合了卷积层和自注意力池化(self-attention pooling, SAP)方法来提取相关特征，使用交叉熵损失作为监督信号进行优化。至于内容识别任务，模块采用长短期记忆网络(long short-term memory, LSTM)进行处理，并通过连接时序分类(connectionist temporal classification, CTC)损失进行训练。这样，轻量化的下游监督模块能够有效地将任务相关的表征转换为不同的情感状态、说话人身份和内容表示，即$\boldsymbol{R}_e\in\mathbf{R}^D,\boldsymbol{R}_s\in\mathbf{R}^D,\boldsymbol{R}_c\in\mathbf{R}^{T\times D}$，从而实现精准的特征提取与分类。

1.2.2 渐进式语音生成器

针对生成类任务，本文在框架中设计了一个语音生成器，以实现语音生成的完整流程。具体而言，渐进式生成器根据提取的3种特征重构出初步的语音结果。声学补偿器利用编码器的浅层输出，对该初步结果进行声学细节的补充。根据Fujisaki理论(Fujisaki, 2004)，说话者在表达过程中通过语言信息、语言外信息和非语言信息逐步组织语言单元，使其成为富有意义和表现力的语句。在本文框架的渐进式生成器中，首先利用一维卷积和长短期记忆网络(LSTM)来对语音内容信息进行建模。随后将重复T次的情感状态和说话人表示，逐步级联到前两层残差卷积块(residual convolution, ResConv)的输入部分。最后通过一维卷积网络预测得到粗略的结果$\boldsymbol{E}'\in\mathbf{R}^{T\times F}$，并使用L1损失函数对生成的输出进行监督训练。预训练提取器的Transformer输入$\boldsymbol{O}_{\text{in}}\in\mathbf{R}^{T\times D}$包含了丰富的声学细节和完整的语音信息。直接将其与渐进生成器结合使用，可能会导致模型输出被$\boldsymbol{O}_{\text{in}}$所主导，从而削弱对期望的情感、说话人和内容表示$\{\boldsymbol{R}_e,\boldsymbol{R}_s,\boldsymbol{R}_c\}$的控制效果。为了解决这一问题，本文提出了一个声学补偿器，用于细化渐进生成器生成的粗略结果。具体而言，首先将Transformer输入$\boldsymbol{O}_{\text{in}}$与重复$T$次的情感状态和说话人表示$\boldsymbol{R}_e,\boldsymbol{R}_s\in\mathbf{R}^{T\times D}$级联，并将它们送入多层LSTM模型中处理。该补偿器通过一维卷积来预测补偿频谱，然后将该补偿谱与渐进生成器的输出相加，得到精细化的最终预测结果$\boldsymbol{E}''\in\mathbf{R}^{T\times F}$。并使用L1损失函数对补偿后的输出进行监督。

2 实验与设置

2.1 实验数据

在实验中，本文针对基于信息解耦的自监督预训练大模型和不同下游任务阶段采用了不同的数据集策略。所有音频样本的采样率均为16 kHz。

2.1.1 自监督语音大模型的预训练

对于基于信息解耦的自监督大模型的预训练，实验中在多样化的数据集上进行了预训练，包含LibriSpeech(Panayotov等，2015)、多语言语音(multilingual LibriSpeech, MLS)(Pratap等，2020)及WenetSpeech(Zhang等，2022)。在实验中部署了两种规模的预训练大模型：Base与Large。

1) Base模型。基于960 h LibriSpeech数据训练，确保了基础性能。

2) Large模型。通过整合总计84 500 h的中英文无标签数据预训练。

Large模型中使用的双语数据集包括44 500 h的MLS英语数据、10 000 h的Wenetspeech中文数据以及30 000 h从互联网收集的中文语音数据。

2.1.2　多种下游任务的训练及微调

为了评估预训练大模型在内容、说话人、语调和声学学习方面的表现，以及验证其多任务普适性，在语音识别（automatic speech recognition，ASR）、说话人识别（speaker identification，SID）、情感识别（emotion recognition，ER）和情感音色转换（emotional voice conversion，EVC）4项关键任务上进行了微调实验。

1）语音识别微调。实验中使用了LibriSpeech的train-clean-100和dev-clean子集作为ASR的训练和开发数据集。模型性能在LibriSpeech的test-clean和test-other子集上进行评估。

2）说话人识别微调。在VoxCeleb1（Nagrani等，2017）数据集上进行了模型的微调和评估，以执行SID任务。VoxCeleb1包含来自1 251位名人的超过100 000条话语，提取自视频。

3）语音情感识别微调。在IEMOCAP数据集的第2—第5部分上对模型进行了微调，并在第1部分上评估其性能。IEMOCAP（Busso等，2008）数据集包含大约12 h的录音，涵盖各种情感表达。值得注意的是，IEMOCAP强调了在对话中自然表达情感，其中语音的情感内容与所说内容的上下文密切相关。

4）情感音色转换微调。EVC综合了ASR、SID和ER 3种任务的特点，因此本文在实验中特别针对这项任务设计了训练策略。在多任务特征优化模块和渐进式解码器的训练中，使用了来自多个渠道的共计12个英文情感语音数据集，包括CREMA-D（Cao等，2014），EMNS（Noriy等，2023），EmoV-DB（Adigwe等，2018），eNTERFACE（Martin等，2006），IEMOCAP，JL-Corpus（James等，2018），MEAD（Wang等，2020），MELD（Poria等，2019），RAVDESS（Livingstone和Russo，2018），TESS（Dupuis和Pichora-Fuller，2010），ESD（Zhou等，2021）和MSP（Martinez-Lucas等，2020）。数据集包含8个情感标签：愤怒、快乐、悲伤、恐惧、厌恶、中性、惊讶以及轻蔑，共计2 080位不同说话者、总时长达到267 h。

2.2　基线模型

实验中，本文选择在PyTorch框架上训练的HuBERT作为对比模型，以评估特征提取器在各项任务中的表现。与本文的预训练模型相同，HuBERT同样训练了Base和Large两个版本，训练采用了完全一致的配置。

在EVC任务中，实验使用了ConsistencyVC（Guo等，2023）和wav2vec-vc（Lim和Kim，2024）作为基线。ConsistencyVC通过ASR微调的wav2vec2.0-large来提取内容，并将从Mel频谱图中提取的说话人或情感信息整合到情感音色转换中。wav2vec-vc模型基于Wav2vec2.0框架，结合了门控循环单元（gated recurrent unit，GRU）和AdaIN-VC（adaptive instance normalization for voice conversion）的转换方法，实现音色转换。GRU用于建模序列数据，确保语音内容的连贯性，而AdaIN-VC通过自适应实例归一化增强了说话人和情感特征的迁移能力。

2.3　模型结构及参数设置

基于信息解耦的自监督预训练大模型如图2（a）所示。在编码器风格的自监督学习（SSL）主干中插入了两个提取器。韵律提取器由3个具有512个通道、卷积核大小为5的卷积层、1个具有256个单元的单层门控循环单元（GRU）和1个具有隐藏特征维度单元的全连接（FC）层组成。说话人提取器由1个具有256个单元的全连接层、1个具有隐藏特征维度单元的FAS层和1个具有隐藏特征维度单元的全连接层组成。

训练的模型分为两个版本：Base和Large。二者均基于昇腾910与MindSpore框架进行训练，由于框架的精度问题，在昇腾910与MindSpore框架上的模型性能略逊于基于Pytorch框架的模型（Wang等，2024a）。在Base模型中，隐藏特征维度为768，在Large模型中为1 024。Base和Large主要结构的参数配置也存在一些调整。对于Base模型，卷积模块由7层组成，每层具有512个通道，步长为{5，2，2，2，2，2，2}，卷积核大小为{10，3，3，3，3，2，2}。Transformer包含12层，维度为768，内部维度为3 072，注意力头数为12。相比之下，对于Large模型，卷积模块的配置与基础版相同，但其Transformer包含24层，维度为1 024，内部维度为4 096，注意力头数为16。

预训练模型的下游任务具体设置如下所示。对于ASR，本文使用了一个包含1 024个单元的2层双向LSTM，采用字符单元的CTC损失进行优化。对于SID，本文首先对语音片段进行均值池化，随后连接

一个包含1 251个类别的全连接(FC)层,并使用交叉熵损失(cross-entropy loss)进行优化。对于情感识别任务,采用基于语句的均值池化进行特征提取,并进一步引入卷积注意力机制(convolutional attention mechanism)进行处理,其中卷积核大小设为5,损失函数同样为交叉熵损失。如图2(b)所示,在多任务特征优化模块中,为了联合建模情感状态和说话人表示,本文使用带有SAP的4层卷积模块,每个模块的核大小为5,内部维度为256。内容表示采用具有1 024个隐藏单元的2层双向LSTM。在生成式解码器中,渐进式生成器包含具有7个核大小的conv1D层、一个中间单向LSTM(维度为1 024)和4个ResConv层(核大小为3,扩张率为2),这些层将通道维度从1 280减少到160。框架中使用了HiFi-GAN1(16 k,80维Mel频谱图)作为声码器。

2.4 预训练过程及微调

在语音大模型的预训练设置中,进行了两次迭代,主要区别在于伪标签的生成方式。第1次迭代中提取13维的MFCC及其一阶和二阶差分特征,使用K-means算法生成100类聚类中心作为伪标签;第2次迭代则使用第1次迭代模型的中间层输出作为特征,生成500类聚类中心。聚类采用MiniBatchK-Means并多次K-means++初始化。自监督说话人教师模型使用Wespeaker工具包预训练了EMA-DINO模型提取说话人嵌入,Base版使用LibriSpeech数据集,Large版使用MLS数据集,分别进行两次40万步和1 750 000步的迭代,通过卷积和Transformer生成帧级别的语音表示。

本文在语音识别(ASR)、说话人识别(SID)、情感识别(ER)和语音重构任务进行了实验,微调使用的数据如2.1节所示。在微调过程中冻结预训练模型的参数,仅微调下游模型的权重以及加权和机制的权重。所有微调都使用Adam(adaptive moment estimation)优化器。由于不同下游任务的数据规模不同,因此每个下游任务的训练步数、学习率和批量大小也不同。详细的配置如表1所示。

2.5 评价指标

在模型性能评估过程中,使用了多种指标,包括准确率(accuracy)、说话人相似度(speaker similarity, SS)、情感表示相似度(emotion representation similarity, ERS)、词错率(word error rate, WER)、感知质量评分(DNS mean opinion score, DNSMos)(Reddy等,2021)。其中,SS通过Resemblyzer模型(Wan等,2018)计算目标语音与生成语音的说话人表示之间的余弦相似度,从而衡量生成语音在保持说话人特征上的准确性。ERS则通过使用Emotion2vec模型(Ma等,2023)计算情感表示之间的余弦相似度,以评估生成语音在传达情感特征方面的有效性。对于词错率,则是通过将生成语音中的识别输出与目标文本进行对比,利用Whisper Large模型(Radford等,2022)来识别语音,进而评估语音生成的准确度。而DNSMos提供了一种非侵入式的语音感知质量评估方法,通过对生成语音的质量进行感知评分,衡量其整体听觉效果。

表1 大模型下游任务微调的参数设置

Table 1 Parameter settings of the pre-trained large model during fine-tuning

任务	更新步数/k	学习率	批次数据量
语音识别	40	1E-4	32
声纹识别	100	1E-3	64
语音情感识别	50	1E-4	16
语音重构	150	5E-4	64

3 结果与分析

3.1 语音信息感知能力的验证

语音预训练提取模型是本文框架的核心部分,本节首先对其特征提取的有效性和普适性进行了验证。首先在语音识别(ASR)、说话人识别(SID)和情感识别(ER)任务上进行实验,实验中针对不同任务都进行了微调。在微调过程中冻结预训练模型的参数,仅微调了下游模型的权重以及加权和机制中的权重。通过这些实验可以重点验证模型理解语言信息、副语言信息和非语言信息的能力,实验结果如表2所示。

在语音识别任务上,为了评估模型的内容理解能力,实验中对其进行了细致的微调,并将结果与HuBERT进行对比。如表2所示,本文的提取模型在Base和Large版本中均在词错率(WER)上优于HuBERT,显示了显著的性能提升。这一改进表明,残差提取策略中的韵律变化和说话人信息有效地增强了模型对内容信息的学习能力,从而降低了词错率。在Base版本中,相较于HuBERT,本文方法在

表 2　基于信息解耦的预训练大模型的多任务性能评测

Table 2　Evaluation of multi-task performance of our information disentanglement based speech pre-trained large model

/%

任务和测试集		本文 Base	HuBERT Base	本文 Large	HuBERT Large
语音识别 (WER)	test-clean	**6.52**	6.72	**3.86**	3.96
	test-other	**15.20**	16.11	**8.64**	8.82
声纹识别 (Acc)	dev	**90.95**	81.42	**97.97**	92.41
	test	**90.61**	80.17	**97.61**	92.29
语音情感识别 (Acc)		**63.96**	62.44	**70.73**	69.49

注：加粗字体为每组结果的最优值。

WER 上实现了 5.65% 的提升，证明了其在捕捉和利用韵律变化及说话人信息方面的有效性，而在 Large 版本中，提升幅度相对较小（2.53%）。

本文通过说话人识别任务评估了模型提取说话人信息的能力。结果表明，无论在 Base 版本还是 Large 版本中，本文的提取模型均实现了最优性能，而且相比 HuBERT 有显著改进。这一提升归功于在最佳位置插入说话人提取器进行残差提取，这不仅增强了主分支的训练效果，也显著提高了提取说话人信息的能力。

由于 IEMOCAP 数据集中情感表达和内容是相互关联的，因此语音情感识别的性能反映了模型对内容理解和副语言信息提取的能力。结果表明，在基础配置和大型配置下，该模型均取得了最佳性能。这一改进可归因于韵律变化信息的残差提取，这使得模型能够捕捉更多的语调和音调细节，同时提取出精准且精炼的内容信息。

3.2　基于情感音色转换的模型信息解耦能力验证

在基于信息解耦的情感音色转换任务上验证本文方法的信息解耦能力。该任务不仅需要有效地提取这些信息，还需要保证特征信息之间的相互独立性，从而实现特征的替换和语音属性的转换。

在实验中，本文选择的基线模型为 ConsistencyVC 和 wav2vec-vc，与本文两个版本的预训练大模型在多个任务中的性能进行了对比，评估了其在情感转换（EC）、音色转换（VC）、情感音色转换（EVC）和直接重构等任务上的性能，并通过 4 个任务的均值量化评价模型的均衡性。如表 3 所示，尽管 ConsistencyVC 在情感和说话人相似性方面表现良好，但其词错率（WER）显著较高。这种差异可归因于使用梅尔频谱图作为转换的参考特征，该特征包含内容信息，并导致内容发生非预期的更改。相比之下，wav2vec-vc 实现了更稳定的结果。在 4 个任务上，基于本文框架在实验中的词错率明显低于参考方法。

表 3　情感转换、音色转换、情感音色转换和语音重构任务上的评估结果

Table 3　Evaluation on emotion conversion，voice conversion，emotion and voice conversion，and speech reconstruction tasks

模型/任务	情感转换				音色转换				情感音色转换				语音重构				均值			
	ERS/%	SS/%	WER/%	DNS	ERS/%	SS/%	WER/%	DNS	ERS/%	SS/%	WER/%	DNS	ERS/%	SS/%	WER/%	DNS	ERS/%	SS/%	WER/%	DNS
ConsistencyVC	69.01	72.07	12.40	3.08	93.45	69.18	11.20	3.04	69.29	69.46	13.70	2.90	93.06	84.84	12.30	2.91	81.20	73.88	12.40	2.98
wav2vec-vc	67.54	69.37	6.44	3.01	93.97	67.17	5.93	3.06	68.75	68.42	10.72	2.97	94.64	84.91	5.14	3.01	81.22	72.46	7.06	3.01
Fbank	67.44	71.28	9.48	2.96	**94.05**	58.76	7.56	2.98	66.96	60.24	9.90	2.74	94.16	86.20	8.32	2.87	80.65	69.12	8.81	2.88
HuBERT base	68.61	72.11	3.50	**3.05**	93.85	**71.03**	3.67	**3.09**	69.03	71.47	3.75	**2.98**	96.00	84.12	3.35	2.96	81.87	74.68	3.57	**3.02**
本文 Base 版本	67.64	71.46	2.78	2.95	93.81	68.71	3.14	2.94	67.06	68.49	3.35	2.68	**97.06**	85.21	2.60	2.99	81.39	73.47	2.97	2.89
HuBERT Large	67.11	71.90	2.32	2.97	93.71	63.24	**2.46**	2.97	66.13	63.55	2.48	2.75	96.51	**87.62**	**1.99**	2.96	80.86	71.57	2.31	2.91
本文 Large 版本	**69.23**	**72.61**	**2.20**	2.98	93.72	70.68	2.51	2.80	**69.24**	**72.28**	**2.39**	2.58	96.24	87.06	2.00	**3.01**	**82.11**	**75.66**	**2.28**	2.84

注：加粗字体为每部分最优值。

在本文模型中，使用24层的Large版本的预训练提取模型时，其在整体性能上超越了参考模型。对应模块的解耦能力可以通过评价指标之间的对比来呈现。24层的表征提取模型在情感和说话人相似性指标上显著优于其他预训练模型，特别是在EVC任务中，情感表示相似度（ERS）为69.24%，说话人相似度（SS）为72.28%，词错率为2.39%。这表明其提取的3个表示之间的耦合度极低。重构任务能有效证明本文方法减少了内容表示中的情感和说话人泄漏，使情感和说话人转换更加可控。同样地，对于12层Base版本的预训练模型，表3展示了其平均结果。对于12层版本在情感和说话人相似性得分上优于其24层模型，尽管词准确性显著降低。这表明虽然24层模型在特定任务（如情感识别、说话人识别和自动语音识别）上表现出色，但其解耦信息的能力较差，导致内容表示泄露较为显著。因此，生成的语音在情感和说话人表示的控制上表现不够理想。

3.3 消融实验与模块分析

在全面分析了实验结果后，本文对框架中的关键模块进行了深入探讨。特征提取模型通过逐步解耦语音中的韵律、说话人和内容信息，结合多任务特征优化模块，生成包含多层次信息的表征。这些模块协同作用，确保系统在情感音色转换任务中的高效表现。为验证这些模块的有效性，本文设计了消融实验，涉及残差提取、说话人提取器和韵律提取器。实验中重点评估了模型在两个任务上的表现：语音识别和说话人识别。这两个任务分别评估了模型理解语音内容和非语言信息的能力。获取的语音表征因包含丰富的说话人身份特征和韵律细节，为后续特征优化模块在识别情感、说话人身份及内容特征方面提供了强有力的支持。

3.3.1 残差提取

残差提取在提取模型中起到了至关重要的作用。基于基础版本的模型，本文将残差提取与多任务提取进行了对比，多任务提取将残差提取中的解耦策略由“减法”（substitutions，Sub）替换为“加法”（additions，Add），实验结果如表4所示。由于说话人提取器的插入位置也会影响模型性能，通过实验将其放置在基础版本表现最佳的位置，即Transformer的第4层之后。

尽管多任务提取（加韵律加说话人）显著提升了预训练模型提取说话人信息的能力，但在语音识别任务中的表现却有所下降。这是因为多任务提取增强了韵律和说话人信息的捕获，但与内容无关的冗余信息干扰了模型深层次的内容提取。当移除韵律信息（减韵律加说话人）时，语音识别性能有所提升，同时说话人识别的改进更加显著，说明韵律变化与内容和说话人信息相关性较低，去除冗余信息有助于提升模型的多任务能力。而当仅移除说话人信息时（加韵律减说话人），尽管说话人识别性能与单任务提取相当，但语音识别性能优于仅去除韵律的情况，这表明说话人信息对内容提取的干扰更为显著。最终，实验结果表明，渐进式残差提取（减韵律减说话人）能够在语音识别和说话人识别任务中实现性能的联合提升。通过逐步精炼主分支中的内容信息，残差提取有效提高了模型在多任务场景中的整体表现。

表4 残差提取与多任务提取策略的对比

Table 4 Evaluation of residual extraction by compared with multi-task extraction

/%

方法	语音识别(WER)		声纹识别(Acc)	
	test-clean	test-other	dev	test
基础版本	6.85	16.77	81.01	79.94
加韵律加说话人	8.06	19.11	88.75	87.58
减韵律加说话人	7.87	18.55	89.69	87.64
加韵律减说话人	6.71	16.36	88.77	87.51
减韵律减说话人	**6.52**	**15.20**	**90.95**	**90.61**

注：加粗字体为每部分最优值。

3.3.2 说话人和韵律提取

本节探讨了韵律提取和说话人提取在特征提取模块中的作用，并将结果汇总于表5。可以看出，引入说话人提取模块，即去除说话人信息后，在说话人识别任务中的表现显著提升，提升幅度为9.63%。在去除韵律信息的情况下，说话人识别的准确率同样有一定的提升。同时，词错误率整体下降，表明其在提高说话人区分能力的同时，也对自动语音识别（ASR）性能有所贡献。

研究表明，强化韵律提取并随之移除韵律变化信息，有助于提升提取模型在语音识别和说话人识别任务上的性能。韵律变化主要包含语调信息，与说话人和内容信息的相关性较低，因此去除这些信息可以更好地分离出内容和说话人信息，从而提升

表 5　说话人提取和韵律提取模块的性能评估

Table 5　Evaluation of speaker and prosody extractions

/%

方法	语音识别(WER)		声纹识别 (Acc)	
	test-clean	test-other	dev	test
基础版本	6.85	16.77	81.01	79.94
减韵律提取	6.74	16.38	81.66	80.03
减说话人提取	6.67	16.04	88.89	87.64
减韵律减说话人	**6.52**	**15.20**	**90.95**	**90.61**

注：加粗字体为每部分最优值。

提取模型在多任务中的表现。类似地，去除说话人信息也可以更好地分离出韵律和内容信息。这表明，强化说话人提取对捕捉说话人信息至关重要，而移除韵律信息则有助于减少干扰，进一步提高提取模型在内容提取任务中的综合性能。

3　讨　论

本文聚焦于海河·谛听大模型中的自监督语音预训练大模型模块，以通用语音感知模型为目标探究了残差提取在语音自监督预训练中的有效性，强调了预训练策略所促成的不同层次任务特征在实现多个下游任务共同提升中的关键作用。实验结果显示，主动从主分支中去除内容无关的语音信息（如说话者信息或音高变化）能够显著提高模型提取内容信息的能力。这一发现验证了残差提取在单一自监督学习（SSL）模型中适应不同类型语音信息需求的必要性，为构建多任务学习模型提供了实证依据。此外，本文的研究表明，预训练任务的增强必须与主分支特定层的任务特征相匹配。偏离这些特征会显著影响训练效果，进而导致模型性能下降。通过将数据集扩展至 84 500 h 的中英双语数据，实验证明了所提方法的可扩展性，特别是在说话者信息提取和语音信息解耦方面取得了显著成果。这些结果不仅为语音处理领域的研究提供了新的视角，也为后续的技术应用奠定了基础。

4　结　论

本文探讨了一种基于信息解耦的自监督语音预训练大模型，旨在利用海量无标注数据提取语音的语言信息、副语言信息和非语言信息，并使这些表征相互独立，以支持下游的大语言模型和生成模型。在此过程中，预训练大模型通过引入轻量化模块增强韵律变化和说话人信息的提取，同时采用残差方法去除其对内容学习的干扰。

实验结果表明，本文提出的预训练模型在多个任务中表现出色，特别是在语音识别、说话人验证、情感识别以及音色转换等任务中，相比现有方法有显著提升。在语音识别、说话人验证和情感识别任务中，Base 版本较 HuBERT 模型分别提升了 5. 65%、13. 02% 和 2. 43%；而 Large 版本则分别提高了 2. 53%、5. 76% 和 1. 78%。在情感音色转换任务上，本文模型超越了 ConsistencyVC 和 wav2vec-vc，在说话人相似度、情感相似度、词错率和感知质量等指标上均有所改善，验证了模型在情感和音色解耦方面的优势。

尽管本研究取得了显著进展，但仍存在一些可以继续完善的方面。模型的实时推理能力还有提升空间，且其多语种跨语言的适应性仍可进一步验证。未来的研究将重点关注以下几个方向：1）多语种与多采样率扩展：进一步拓展至多语种和多采样率数据集，以增强模型在跨语言、跨域场景中的适应性。2）多模态融合：结合视觉与听觉等多模态信息，以提升情感识别与语音转换的自然度和准确性。3）实时推理与效率优化：通过模型压缩与高效推理算法，优化大规模数据集的实时推理能力，提升系统响应速度与实用性。

参考文献（References）

Adigwe A, Tits N, El Haddad K, Ostadabbas S and Dutoit T. 2018. The emotional voices database: towards controlling the emotion dimension in voice generation systems [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/1806.09514.pdf

Baevski A, Hsu W N, Xu Q T, Babu A, Gu J T and Auli M. 2022. data2Vec: a general framework for self-supervised learning in speech, vision and language//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 1298-1312

Busso C, Bulut M, Lee C C, Kazemzadeh A, Mower E, Kim S, Chang J N, Lee S and Narayanan S S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4): 335-359 [DOI: 10.1007/s10579-008-9076-6]

Cao H W, Cooper D G, Keutmann M K, Gur R C, Nenkova A and Verma R. 2014. CREMA-D: crowd-sourced emotional multimodal

actors dataset. IEEE Transactions on Affective Computing, 5(4): 377-390 [DOI: 10.1109/TAFFC.2014.2336244]

Chen S Y, Wang C Y, Chen Z Y, Wu Y, Liu S J, Chen Z, Li J Y, Kanda N, Yoshioka T, Xiao X, Wu J, Zhou L, Ren S, Qian Y M, Qian Y, Wu J, Zeng M, Yu X Z and Wei F R. 2022. WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505-1518 [DOI: 10.1109/JSTSP.2022.3188113]

Défossez A, Copet J, Synnaeve G and Adi Y. 2022. High fidelity neural audio compression [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2210.13438.pdf

Desplanques B, Thienpondt J and Demuynck K. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2005.07143.pdf

Dupuis K and Pichora-Fuller M K. 2010. Toronto emotional speech set (TESS)-younger talker happy [EB/OL]. [2024-09-15]. https://hdl.handle.net/1807/24487

Fujisaki H. Information, prosody, and modeling//Proceedings of Speech Prosody 2004. Nara, Japan: [s.n.], 2004: 1-10 [DOI: 10.21437/speechprosody.2004-1]

Guo H J, Liu C R, Ishi C T and Ishiguro H. 2023. Using joint training speaker encoder with consistency loss to achieve cross-lingual voice conversion and expressive voice conversion//Proceedings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei, China: IEEE: 1-8 [DOI: 10.1109/ASRU57964.2023.10389651]

Han B, Huang W, Chen Z Y and Qian Y M. 2023. Improving DINO-based self-supervised speaker verification with progressive cluster-aware training//Proceedings of 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSPW59220.2023.10192957]

Hsu W N, Bolte B, Tsai Y H H, Lakhotia K, Salakhutdinov R and Mohamed A. 2021. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 3451-3460 [DOI: 10.1109/TASLP.2021.3122291]

James J, Tian L and Watson C. 2018. An open source emotional speech corpus for human robot interaction applications//Proceedings of Interspeech 2018. Hyderabad, India: [s.n.]: 2768-2772 [DOI: 10.21437/Interspeech.2018-1349]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 19730-19742

Lim J and Kim K. 2024. Wav2vec-VC: voice conversion via hidden representations of wav2vec 2.0//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea(South): IEEE: 10326-10330 [DOI: 10.1109/ICASSP48485.2024.10447984]

Liu A T, Li S W and Lee H Y. 2021. TERA: self-supervised learning of transformer encoder representation for speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 2351-2366 [DOI: 10.1109/TASLP.2021.3095662]

Liu T T, Liu Z, Chai Y J, Wang J and Wang Y Y. 2021. Agent affective computing in human-computer interaction. Journal of Image and Graphics, 26(12): 2767-2777 (刘婷婷, 刘箴, 柴艳杰, 王瑾, 王媛怡. 2021. 人机交互中的智能体情感计算研究. 中国图象图形学报, 26(12): 2767-2777) [DOI: 10.11834/jig.200498]

Livingstone S R and Russo F A. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5): #e0196391 [DOI: 10.1371/journal.pone.0196391]

Ma Z Y, Zheng Z S, Ye J X, Li J C, Gao Z F, Zhang S L and Chen X. 2023. emotion2vec: self-supervised pre-training for speech emotion representation [EB/OL] [2024-09-15]. https://arxiv.org/pdf/2312.15185.pdf

Martin O, Kotsia I, Macq B and Pitas I. 2006. The eNTERFACE'05 audio-visual emotion database//Proceedings of the 22nd International Conference on Data Engineering Workshops. Atlanta, USA: IEEE: 8-8 [DOI: 10.1109/ICDEW.2006.145]

Martinez-Lucas L, Abdelwahab M and Busso C. 2020. The MSP-conversation corpus//Proceedings of Interspeech 2020. Shanghai, China: [s.n.]: 1823-1827 [DOI: 10.21437/Interspeech.2020-2444]

Nagrani A, Chung J S and Zisserman A. 2017. VoxCeleb: a large-scale speaker identification dataset//Proceedings of Interspeech 2017. Stockholm, Sweden: [s.n.]: 2616-2620 [DOI: 10.21437/Interspeech.2017-950]

Noriy K A, Yang X S and Zhang J J. 2023. EMNS/Imz/Corpus: an emotive single-speaker dataset for narrative storytelling in games, television and graphic novels [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2305.13137.pdf

Panayotov V, Chen G G, Povey D and Khudanpur S. 2015. Librispeech: an ASR corpus based on public domain audio books//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Australia: IEEE: 5206-5210 [DOI: 10.1109/ICASSP.2015.7178964]

Poria S, Hazarika D, Majumder N, Naik G, Cambria E and Mihalcea R. 2019. MELD: a multimodal multi-party dataset for emotion recognition in conversations [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/1810.02508.pdf

Pratap V, Xu Q T, Sriram A, Synnaeve G and Collobert R. 2020. MLS: a large-scale multilingual dataset for speech research [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2012.03411.pdf

Radford A, Kim J W, Xu T, Brockman G, McLeavey C and Sutskever I. 2022. Robust speech recognition via large-scale weak supervision

[EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2212.04356.pdf

Reddy C K A, Gopal V and Cutler R. 2021. DNSMOS: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6493-6497 [DOI: 10.1109/ICASSP39728.2021.9414878]

Rubenstein P K, Asawaroengchai C, Nguyen D D, Bapna A, Borsos Z, de Chaumont Quitry F, Chen P, El Badawy D, Han W, Kharitonov E, Muckenhirn H, Padfield D, Qin J, Rozenberg D, Sainath T, Schalkwyk J, Sharifi M, Ramanovich M T, Tagliasacchi M, Tudor A, Velimirović M, Vincent D, Yu J H, Wang Y Q, Zayats V, Zeghidour N, Zhang Y, Zhang Z S, Zilka L and Frank C. 2023. AudioPaLM: a large language model that can speak and listen [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2306.12925.pdf

Schneider S, Baevski A, Collobert R and Auli M. 2019. wav2vec: unsupervised pre-training for speech recognition [EB/OL]. [2024-09-15]. https://arxiv.org/abs/1904.05862.pdf

Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(6): 1607-1627) [DOI: 10.11834/jig.240017]

Tsai H S, Chang H J, Huang W C, Huang Z L, Lakhotia K, Yang S W, Dong S Y, Liu A T, Lai C I J, Shi J T, Chang X K, Hall P, Chen H J, Li S W, Watanabe S, Mohamed A and Lee H Y. 2022. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2203.06849.pdf

Wan L, Wang Q, Papir A and Moreno I L. 2018. Generalized end-to-end loss for speaker verification//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 4879-4883 [DOI: 10.1109/ICASSP.2018.8462665]

Wang K S Y, Wu Q Y, Song L S, Yang Z Q, Wu W, Qian C, He R, Qiao Y and Loy C C. 2020. MEAD: a large-scale audio-visual dataset for emotional talking-face generation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 700-717 [DOI: 10.1007/978-3-030-58589-1_42]

Wang T R, Li J, Ma Z Y, Cao R, Chen X, Wang L B, Ge M, Wang X B, Wang Y G, Dang J W and Tashi N. 2024a. Progressive residual extraction based pre-training for speech representation learning [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2409.00387.pdf

Wang T R, Zhou L, Zhang Z Q, Wu Y, Liu S J, Gaur Y, Chen Z, Li J Y and Wei F R. 2024b. VioLA: conditional language models for speech recognition, synthesis, and translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 3709-3716 [DOI: 10.1109/TASLP.2024.3434425]

Xu Y X, Li B, Tan S Q and Huang J W. 2024. Research progress on speech deepfake and its detection techniques. Journal of Image and Graphics, 29(8): 2236-2268 (许裕雄, 李斌, 谭舜泉, 黄继武. 2024. 语音深度伪造及其检测技术研究进展. 中国图象图形学报, 29(8): 2236-2268) [DOI: 10.11834/jig.230476]

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类 ChatGPT 大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Yang S W, Chi P H, Chuang Y S, Lai C I J, Lakhotia K, Lin Y Y, Liu A T, Shi T T, Chang X K, Lin G T, Huang T H, Tseng W C, Lee K T, Liu D R, Huang Z L, Dong S Y, Li S W, Watanabe S, Mohamed A and Lee H Y. 2021. SUPERB: speech processing universal performance benchmark [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2105.01051.pdf

Zeghidour N, Luebs A, Omran A, Skoglund J and Tagliasacchi M. 2022. SoundStream: an end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 495-507 [DOI: 10.1109/TASLP.2021.3129994]

Zhang B B, Lv H, Guo P C, Shao Q J, Yang C, Xie L, Xu X, Bu H, Chen X Y, Zeng C C, Wu D and Peng Z D. 2022. Wenetspeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition//Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 6182-6186 [DOI: 10.1109/ICASSP43922.2022.9746682]

Zhang D, Li S M, Zhang X, Zhan J, Wang P Y, Zhou Y Q and Qiu X P. 2023. SpeechGPT: empowering large language models with intrinsic cross-modal conversational abilities [EB/OL]. [2024-09-15]. https://arxiv.org/pdf/2305.11000.pdf

Zhou K, Sisman B, Liu R and Li H. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE: 920-924 [DOI: 10.1109/ICASSP39728.2021.9413391]

作者简介

王龙标，男，教授，主要研究方向为语音信号处理、言语人机交互和人工智能。E-mail: longbiao_wang@tju.edu.cn

江宇，女，硕士研究生，主要研究方向为语音自监督预训练和语音合成。E-mail: jiang_y@tju.edu.cn

王天锐，男，博士研究生，主要研究方向为语音自监督预训练与语音合成。E-mail: wangtianrui@tju.edu.cn.

王晓宝，男，助理研究员，主要研究方向为自然语言处理和言语人机交互。E-mail: wangxiaobao@tju.edu.cn

党建武，男，教授，主要研究方向为语音信号处理、自然语言处理、言语人机交互和人工智能。E-mail: jdang@jaist.ac.jp