目　录

第 1 章

第 2 章

第 3 章　人声敏感的增强损失函数设计

3.1　引言

深度学习技术的进步带动了语音识别（Automatic Speech Recognition，ASR）的发展，近场环境下的语音识别准确率已经逐渐接近人类的水平[1,2]。然而在远场场景下，麦克风除了接收到目标语音的音频之外，还会受到噪声、混响以及干扰说话人声音的影响，在此场景下，ASR 识别的准确率会急剧下降。因此设计一个良好的前端降噪模块对于后端 ASR 系统的应用非常重要。

远场语音识别问题近些年来得到了广泛的关注和研究，当收音设备为麦克风阵列时，多通道语音增强算法能有效减少噪声，提升语音识别的准确率[3]。然而，现实场景中存在很多只有单个麦克风可用的情况，该场景下的语音识别效果会落后于麦克风阵列，因此本章聚焦于单通道语音增强算法的改进。

语音增强（Speech Enhancement，SE）的目标是从混合语音中提取出目标人声，同时抑制其他干扰声音，这里的干扰主要指噪声和混响。然而大多数语音增强系统在消除噪声的同时，往往也会损坏目标人声引入失真，从而造成语音识别系统准确率的下降[4]，因此前端增强系统的设计不仅仅要考虑到噪声消除，而且要尽量减少人声的损失。

为了提升语音识别系统的鲁棒性，相关的工作可以分为一下几类：

1、在语音识别系统前添加语音增强前端，如图3-1所示。其中常见的增强前端包括谱减法[5]，维纳滤波[6] 和深度神经网络（Deep Neural Network，DNN)[7]等。然而，该类方法中语音增强的目标为估计目标语音，和语音识别系统的优化目标不同，因而可能会导致局部优化问题[8]。此外，语音增强系统往往会导致语音过度平滑，因而常常造成语音失真。因此，语音识别的性能非常依赖于增强前端的性能。

2、多条件训练（Multi-condition training，MCT），MCT 采用不同类型的数据训练（干净语音和带噪语音）语音识别模块，然而该方法会带来复杂度和训练时间的提升。不仅如此，当在未见的数据上进行评估时[9]，识别准确率也会下降，模型泛化性较差，而且仍然会受到语音失真的影响[10]。为了缓解语音失真问题，增强前端首先对训练集和测试集进行增强，并用增强后的数据对 ASR 模型进行训练。它能在一定程度上改善 ASR 性能，但仍然高度依赖于增强前端的性能。与 MCT 方法不同，铺增强方法[11] 直接将数据增强应用于神经网络的输入特征，铺

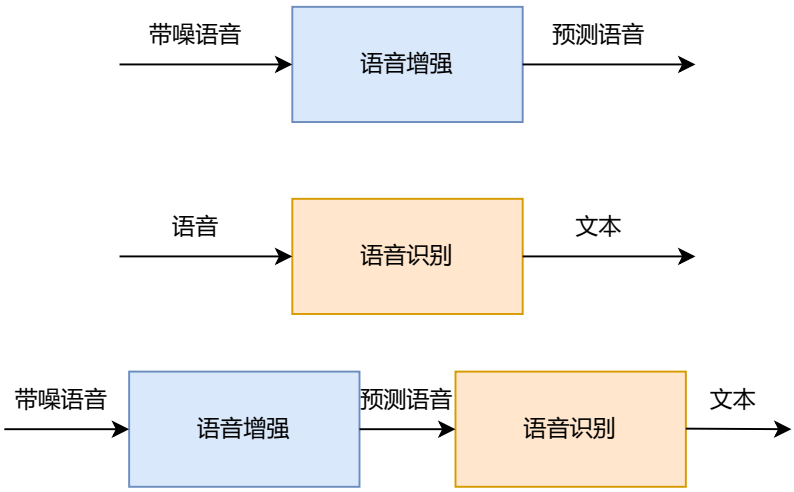


图 3-1 前端和后端任务单独优化。

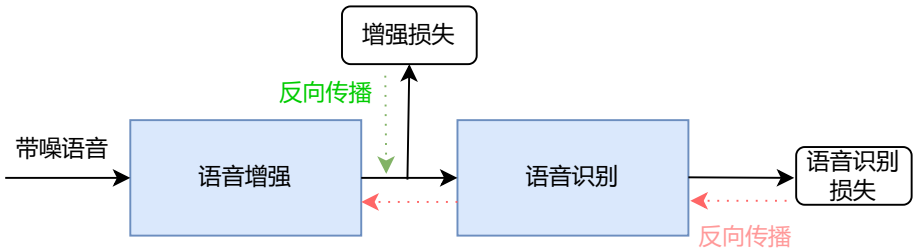


图 3-2 联合训练示意图，模型在增强和识别两类损失下优化。

增强只在训练过程中使用，包括三种谱图的增强方法: 时间扭曲、时间和频率掩蔽。虽然铺增强可以提高端到端 ASR 的性能，但在嘈杂环境中仍然需要改进。

3、联合训练[12,13]，如图3-2。这些方法采用联合训练框架，同时优化语音增强和语音识别。该方案使得语音增强和语音识别不是两个独立的任务，它们可以相互影响，从而提升性能。但该方法仅将增强后的特征作为语音识别的输入，仍会受到语音失真问题的影响。此外，在噪声较大的 AISHELL-1 数据集中，该方法的字符错误率 (CER) 仍在 50% 以上，有待改进。

4、混合语音和增强语音的融合方法[14,15]。Iwamoto 等人[16] 的工作指出，影响语音识别系统准确率的因素主要为语音增强系统引入的人工噪声（Artifacts），论文指出通过将混合语音和增强语音混合相加的方法可以有效减少语音的失真，提升语音识别系统的准确率。

上述工作大多数工作在面对混合语音时，无论是否添加增强前端，都需要对语音识别系统进行重新训练。然而在现实场景下语音识别系统的模型参数比较大，训练耗时比较长，重训的成本很高，不符合实际。因而本章根据该问题，提出对于人声敏感的损失函数，来有效减少目标语音失真。

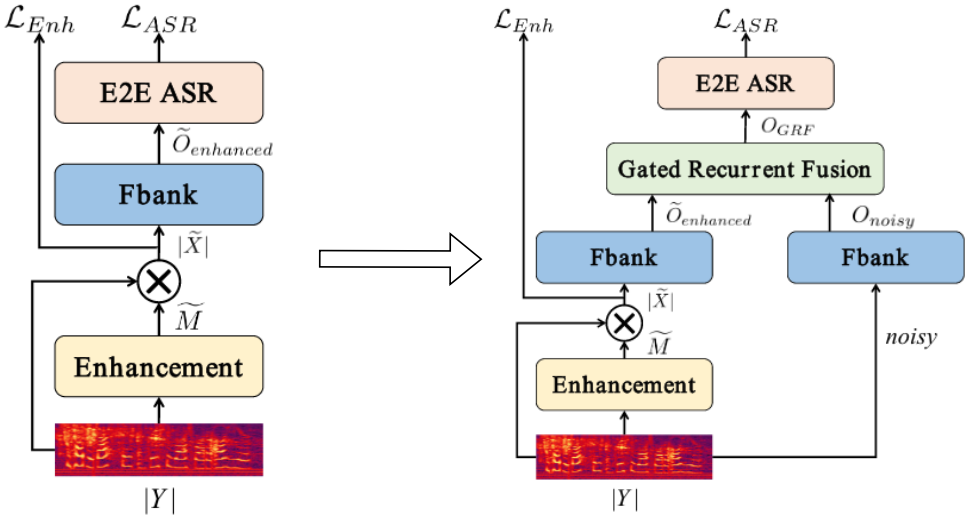


图 3-3 左图为传统联合训练的方法，右图为融合混合语音和增强语音的方法[15]。

3.2 语音增强模型介绍

本小节先介绍问题的定义，然后介绍时域和频域的语音增强模型，最后介绍训练所用的损失函数。

3.2.1 问题定义

现实场景的语音交互系统伴随着复杂的声学场景，包括混响、背景噪声以及其他说话人干扰等等，这会严重影响语音交互的性能。如图3-4所示，这一般也被称为鸡尾酒会问题。

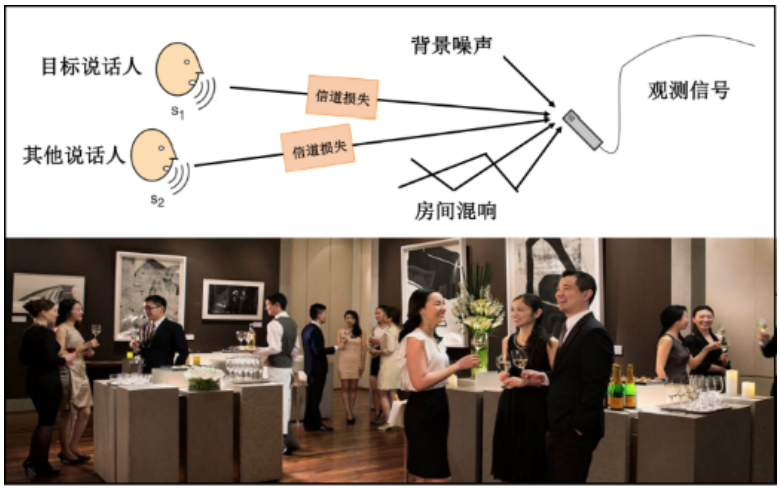


图 3-4 鸡尾酒会问题。

为了简化问题，本文只考虑混响和背景噪声的干扰，设 $y(t)$ 为麦克风接收的

信号，$y(t)$ 可以表示为：

$$y(t) = x(t) * h(t) + n(t) \tag{3-1}$$

其中 $x(t)$，$n(t)$，$h(t)$ 分别代表目标语音，加性背景噪声和混响，t 代表时间步。语音增强的目标为从混合信号 $y(t)$ 中恢复出目标语音 $x(t)$，可以表示为：

$$\mathcal{F}[y;\Theta] \rightarrow x \tag{3-2}$$

针对此问题，目前深度学习的语音增强方法基本如图3-5所示，深度神经网络通过数据驱动的方法学得人声和噪声之间在性质上的区别，从而从混合语音中将目标人声剥离。该方案中常用的网络结构一般可分为语音编码器，增强网络和语音解码器。语音编码器是根据语音信号 y（为了简化，省略了下标 t）到语音类谱表征的特征编码操作，语音解码器则是从提取的类谱表征重构还原成语音信号波形的操作。增强网络是整个方案的核心模块，它负责从编码得到的混合语音类谱表征估计掩码或者直接估计目标语音的类谱表征。

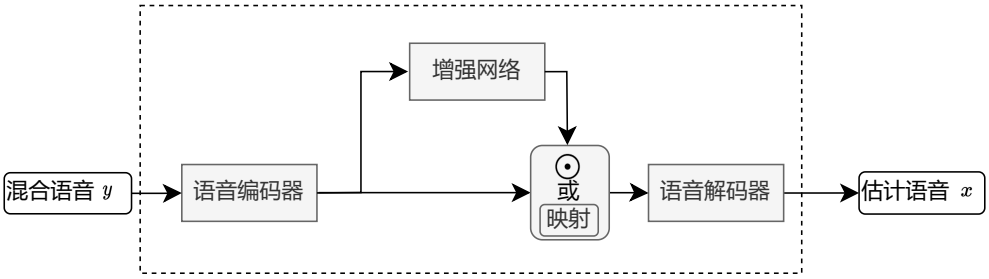


图 3-5　语音增强方案基本框架图。

以上即为数据驱动方法对于语音增强任务的解剖和定义，现存的数据驱动的工作大多数是基于这种框架的改进。因此本章接下来将对非线性滤波器 **F** 网络参数进行逐一阐述。

3.2.2　语音编解码器

表 3-1　时频域特征编码的不同特点

模型类型	语音编码器	语音解码器
频域方法	短时傅里叶变换（STFT）	逆短时傅里叶变换
时域方法	一维卷积（Conv-1D）	一维转置卷积（ConvTranspose-1D）

根据语音特征编解码方式的不同，通常将模型方法分为频域方法和时域方法。如图表3-1所示，频域方法常用短时傅里叶变换（STFT）和逆短时傅里叶变

换（iSTFT）做语音特征编解码，具体计算公式如下：

$$Y(t,f) = STFT[y(\tau)] = \sum_{\tau=tH}^{tH+N-1} y(\tau)\omega(\tau - tH)e^{-2\pi jf(\tau-tH)/N} \tag{3-3}$$

其中，ω 是窗函数，N 和 H 代表窗长和窗移，t 和 f 分别代表所得到频谱的帧和频率的索引。

$$x(\tau) = iSTFT[S(t,f)] = \begin{cases} \sum_t \frac{1}{N}\sum_{f=0}^{N-1} S(t,f)e^{2\pi jf(\tau-tH)/N}\omega(\tau - tH), tH \le \tau < tH + N \\ 0, otherwise \end{cases} \tag{3-4}$$

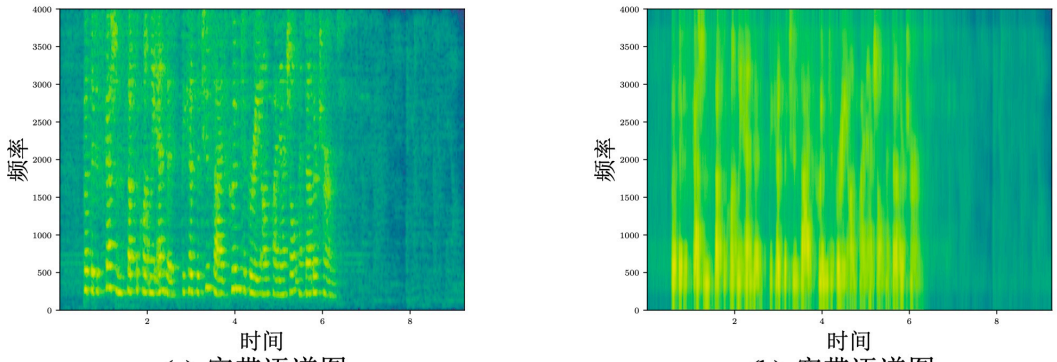
时域方法则是利用一维卷积（Conv-1D）和一维转置卷积（ConvTrasnpose-1D）操作，是可学习的编码器和解码器。能按照分离任务进行优化，可以表示为：

$$Y(t,f) = Conv1D[y(\tau)] = \sigma(y(\tau) * U_L) \tag{3-5}$$

$$x(\tau) = ConvTrans1D[S(t,f)] = X(t,f) * V_L \tag{3-6}$$

其中 U_L 和 V_L 分别代表卷积核长度为 L 的卷积参数。

频域和时域方法的编解码器由于窗长的不同存在些许区别，频域 STFT 操作的特征具有明显的声纹结构，而时域一维卷积学习得到的特征没有明显的规律。在相同域的数据集上，时域的方法往往效果更好[17]，但是泛化效果不如频域的方法。因而本章中，选择频域的方法作为编解码器。图3-6展示了两种 STFT 设置之后不同的语谱图，即为宽带语谱图和窄带语谱图。时域信号在进行短时傅里叶变换时，选择不同的帧长，绘制的语谱图特点不同。以 20 到 30ms 为帧长绘制的语谱图为窄带语谱图，该语谱图每帧时长较长，可以观察到谐波的存在，谐波体现了声音的纹理特征，但是无法观察到共振峰的频率。以 2ms 左右为帧长绘制的语谱图为宽带语谱图，由于每帧的时长较短，因此可以看到连续的共振峰频率，但无法观察到谐波。



(a) 宽带语谱图　　(b) 窄带语谱图

图 3-6　　语谱图实例。

3.2.3 增强网络

经过语音编码器之后得到的语音表征 $Y(t,f)$，需要经过增强网络提取目标语音的信号表征 $X(t,f)$，才能得到最后语音解码器之后所输出的目标语音信号，因此增强网络是整个模型的核心模块。增强网络的核心希望对更多的语音上下文信息进行跟踪与建模。

现有的增强网络总共可以分为两类：基于频谱映射的算法和基于时频掩码的算法。基于频谱映射的算法，是利用神经网络直接估计目标语音信号，建立混合语音表征 $Y(t,f)$ 和目标语音表征 $X(t,f)$ 之间的非线性联系，定义如下：

$$Y(t,f) \xrightarrow{Net} X(t,f) \tag{3-7}$$

而基于时频掩码的算法则是通过先估计掩码信息 $M(t,f)$，即预测目标语音信号在混合语音信号之间的占比，建立相关的规律性，然后利用元素乘积的方式来获得 $X(t,f)$，其定义如下：

$$Y(t,f) \xrightarrow{Net} M(t,f) \tag{3-8}$$

$$Y(t,f) \odot M(t,f) \rightarrow X(t,f) \tag{3-9}$$

这两类算法有其独特的优势，基于频谱映射的方法对数据的信噪比变化更加不敏感，在低信噪比情况下表现更优。基于视频掩码的算法则能够很好地利用数据驱动的方式挖掘目标语音语其他干扰信号之间的互信息。在现有数据驱动的背景下，基于时频掩码的方法占比更高。

深度神经网络（DNN）是最早被应用来做语音增强任务的，DNN 包含多个隐藏层的神经网络，包含输入层，隐藏层和输出层，每层之间的节点互相连接。深度自编码网络（DAE）是一种无监督学习的网络结构，它以输入数据本身作为监督，来指导神经网络学习一个非线性的映射关系。循环神经网络（RNN）则是专门用于处理时序数据的网络，通过隐藏状态以及门控机制来存储历史信息，对语音数据的时序进行上下文的建模。卷积网络（CNN）则是由卷积层、池化层和全连接层组成的神经网络，拥有局部连接，权重共享以及池化三个重要的特征，能够减少网络的参数量，允许并行输出等。

本章中采用 Luo 等人[18] 提出的 Conv-TasNet 模型中的分离网络，作为增强网络。该模型使用时间卷积网络（TCN）作为基本结构。TCN 网络由具有相同输入和输出长度的扩张一维卷积层组成。首先，其卷积网络层层之间是有因果关系的，这就意味着 TCN 能实现和 RNN 类似的效果，对历史信息能很好的保留。其次，TCN 网络可以通过扩张一维卷积堆叠，用尽可能小的参数扩张获取时间上下文信息的距离长度。最后，ConvTasNet 的分离性能也验证了其有效性和高效性，只有 5.1MB 的模型大小就实现了公开数据集 WSJ0-2Mix 上 15.3 分贝的 SI-SDR

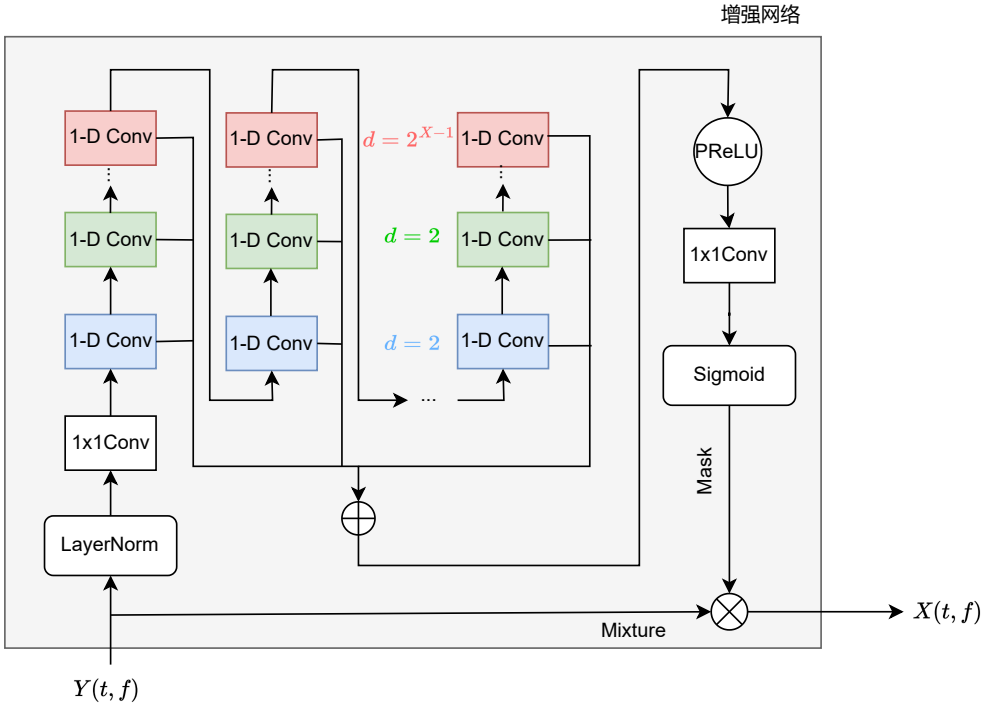


图 3-7　　增强网络中的 TCN 模型示意图。

性能提升。

图3-7为本章采用的增强网络，TCN 中一维卷积块的不同颜色表示不同的扩张因子。TCN 的每一层由膨胀系数逐渐增加的一维卷积块组成。膨胀因子呈指数增长，以确保有足够大的时间上下文窗口，以利用语音信号的长距离上下文依赖关系，如图中不同颜色所示。在该增强网络中，膨胀系数分别为 $1, 2, 4, 2^{M-1}$ 的 M 个卷积块重复了 R 次。每个卷积块的输入都有相应的零填充，以确保输出长度与输入相同。TCN 的输出先经过 PReLU 模块，然后被传递到一个核大小为 1 的卷积块 (1×1−conv 块，也称为 pointwise 卷积) 用于掩码估计。1×1 - conv 块与非线性激活函数一起估计目标语音的掩码。

图3-8为 1-D Conv 的结构示意图，该结构设计参考了 WaveNet[19]。每个卷积块的 residual path 作为下一个块的输入，所有块的 skip-connection 路径相加作为 TCN 的输出。为了进一步减少参数的数量，在每个卷积块中使用深度可分离卷积 (S-conv(·)) 来代替标准卷积。深度可分离卷积 (也称为可分离卷积) 已被证明在图像处理任务和机器翻译任务中有效。深度可分离卷积算子将标准卷积运算解耦为两个连续运算，一个是深度卷积 (D-conv(·))，然后是点卷积 (1×1−conv(·)):

$$D - conv(Y, K) = concat(y_j * k_j), j = 1, 2, ..., N \tag{3-10}$$

$$S - conv(Y, K, L) = D - conv(Y, K) * L \tag{3-11}$$

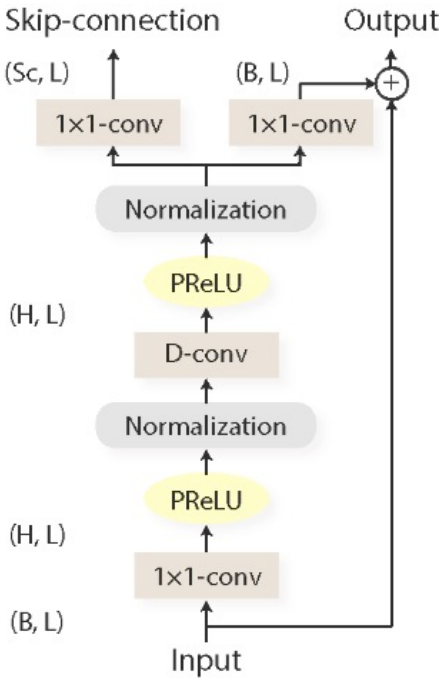


图 3-8　　1-D Conv 结构示意图。

其中 $Y \in R^{G\times M}$ 为 S-conv 的输入，$K \in R^{G\times P}$ 为大小为 P 的卷积核，$y_j \in R^{1\times M}$ 和 $k_j \in R^{1\times P}$ 为矩阵 Y 和 K 的其中一行。$L \in RG \times H \times 1$ 为大小为 1 的卷积核，* 为卷积操作。也就是，D-conv(·) 操作将输入 Y 的每一行与矩阵 K 的相应行进行卷积，1 × 1−conv 块对特征空间进行线性变换。与核大小为 $K \in RGHP$ 的标准卷积相比，深度可分离卷积只包含 $G \times P + G \times H$ 参数，当 $H >> P$ 时，模型大小减小了 $\frac{H\times P}{H+P} \approx P$ 倍。

在 1 × 1-conv 块和 D-conv 块后分别增加一个非线性激活函数和一个归一化操作。非线性激活函数是 PReLU：

$$PReLU = \begin{cases} x, if\, x \geq 0 \\ ax, otherwise \end{cases} \tag{3-12}$$

其中 $a \in R$ 是一个可训练的标量，控制该激活函数的负斜率。网络中归一化方法的选择取决于因果关系要求。对于非因果配置，全局层归一化 (gLN) 优于所有其他归一化方法。gLN 中特征会按照通道维度和时间维度一起进行归一化：

$$gLN(\mathbf{F}) = \frac{F - E[\mathbf{F}]}{\sqrt{Var[\mathbf{F}] + \epsilon}} \odot \gamma + \beta \tag{3-13}$$

$$E[\mathbf{F}] = \frac{1}{NT} \sum_{NT} \mathbf{F} \tag{3-14}$$

$$Var[\mathbf{F}] = \frac{1}{NT} \sum_{NT} (\mathbf{F} - E[\mathbf{F}])^2 \tag{3-15}$$

其中 $F \in RN \times T$ 为输入特征，$\gamma,\beta \in RN \times 1$ 为可训练参数，ϵ 为增加数值稳定性的小常数。这与计算机视觉模型中应用的标准层归一化相同。在因果配置中，gLN 不能应用，因为它依赖于信号在任何时间步长的未来值。因此，设计了一个累积层归一化 (cLN) 操作，来代替 gLN 完成归一化操作:

$$cLN(\mathbf{f}_k) = \frac{\mathbf{f}_k - E[\mathbf{f}_{t\le k}]}{\sqrt{Var[\mathbf{f}_{t\le k}]}} \odot \gamma + \beta \tag{3-16}$$

$$E[\mathbf{f}_{t\le k}] = \frac{1}{Nk}\sum_{Nk}\mathbf{f}_{t\le k} \tag{3-17}$$

$$Var[\mathbf{f}_{t\le k}] = \frac{1}{Nk}\sum_{Nk}(\mathbf{f}_{t\le k} - E[\mathbf{f}_{t\le k}])^2 \tag{3-18}$$

其中，$\mathbf{f}_k \in R^{N\times 1}$ 为输入特征 F 的第 k 帧，$\mathbf{f}_{t\le k} \in R^{N\times k}$ 对应着前 k 个帧的特征 $[\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, ..., \mathbf{f}_k]$,，$\beta \in R^{N\times 1}$ 是应用于所有框架的可训练参数。为了确保分离模块对输入的缩放是不变的，利用所选的归一化方法对编码器的输出 $Y(t,f)$ 进行缩放，然后在传递给后续的增强网络。

在增强模块之前添加线性 1x1-conv 卷积模块作为瓶颈层，该块决定了后续卷积块的输入通道数和残差路径。例如，线性瓶颈层有 B 个通道，那么对于一个有 H 个通道，核大小为 P 的 1-D Conv，第一个 1 ×1-conv 块和第一个 D-conv 块中的核大小分别为 $O \in R^{B\times H\times 1}$ 和 $K \in R^{H\times P}$，residual path 中的核大小为 $L_{Rs} \in R^{H\times B\times 1}$。

3.2.4 训练目标

在数据驱动的有监督语音增强框架下，定义一个合适的训练目标对于网络的学习和泛化能力至关重要。语音增强的优化目标是让预测的信号和目标语音信号更加接近，即让网络预测的表征 $\hat{X}(t,f)$ 通过语音解码器输出预测的语音 $\hat{x}(t)$，并使得预测语音和目标语音之间的距离最小化：

$$\hat{\Theta} = \min_{\Theta} \mathcal{D}ist\{x(t), \hat{x}(t)\} \tag{3-19}$$

其中，在语音增强中常用的距离函数可以表示为归类为和欧式距离相关的距离函数（MAE，MSE）和目标指标相关的距离函数（SDR，SA-SDR，SI-SDR，wSDR）等等。

如图3-9所示，根据训练目标在网络中设置的不同位置，语音增强模型的训练目标主要分为频域目标 L_{Freq} 和时域目标 L_{Time}。频域目标的目的是重构目标语音表征，然后在推理阶段通过解码器 Decoder 还原目标语音信号。而时域目标的目的是直接重构目标语音信号的波形，训练过程和推理过程完全一致。其中需要说明的是，频域目标架构下的 $S(t,f)$ 是泛指，此处可以是振幅特征，对数谱特征，实部和虚部拼接的复数谱特征，振幅和相位拼接的复数谱特征等，都可适用。

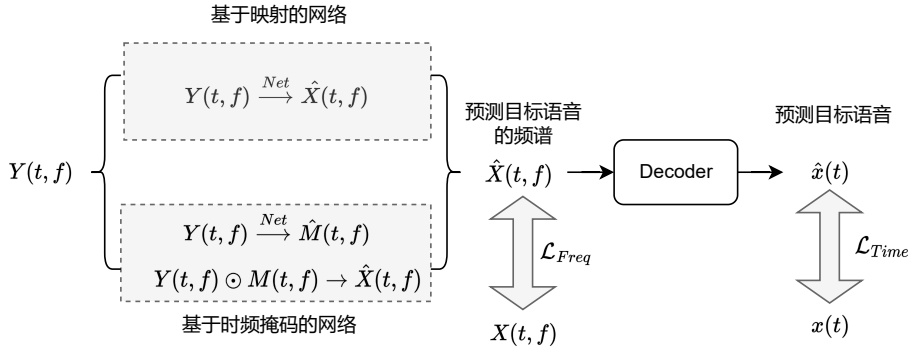


图 3-9　　时域和频域的损失函数对比。

本章采用的是频域的 MSE 损失函数，可以表示为如下：

$$\mathcal{L}_{MSE} = \|X_r - \hat{X}_r\|^2 + \|X_i - \hat{X}_i\|^2 \tag{3-20}$$

其中下标 r 和 i 代表实部和虚部。

3.3　系数压缩的损失函数

根据经验，频谱的能量分布通常在频率上是不平衡的。例如，对于 0~ 2000 Hz 的频谱区域，其频谱值较大，而对于无声或无声音区域则相反。因此，在以最小平方误差 (MSE) 等准则训练网络时，如果不对频谱进行特征压缩，频谱值较大的区域往往会优先优化，因为对这些区域的优化会带来更明显的损失减少。相反，数值较小的区域不能很好地优化，对最终损失计算的贡献变得微不足道，导致弱能量区域的频谱结构模糊。因此，如果采用合适的压缩函数来减小动态范围，平衡不同光谱区域之间的损失差距，网络有望在弱区域捕获更详细的信息，这将有助于感知质量的提高。

对数功率谱特征被认为是一种通用的降噪方法，因为它更好地描述了人耳对声强级的特征最近，Zhao 等人[20] 采用了一种不同的特征预处理方法，采用立方根压缩的方法来完成反混响任务，并获得了不错的性能。在此基础上，我们提出了一种广义压缩方法，$|X|^\beta$，其中 $\beta \in (0,1]$，β 越小，代表压缩效果越强。

在极坐标下，混合语音的复谱可以写成 $X = |X|e^{i\theta_x}$。由于相位谱几乎没有时间谱上的规律，很难准确估计，所以我们在这里只压缩幅度而不改变相位信息。因此，压缩后的复谱可以表示为 $X^\beta = |X|^\beta e^{i\theta_x}$ 也可以用笛卡尔坐标来表示：

$$X^\beta = X_r^\beta + X_i^\beta \tag{3-21}$$

$$X_r^\beta = |X|^\beta cos\theta_x \tag{3-22}$$

$$X_i^\beta = |X|^\beta sin\theta_x \tag{3-23}$$

参考公式3-20，本章采用以下的公式恢复实部和虚部中的信息：

$$\mathcal{L}_{RI} = \|X_r^\beta - \hat{X}_r^\beta\|^2 + \|X_i^\beta - \hat{X}_i^\beta\|^2 \tag{3-24}$$

Wang 等人[21] 证明添加了幅度谱损失之后，语音质量也会得到提升，因此，本章在公式3-24基础上添加了幅度谱损失：

$$\mathcal{L}_{Mag} = \|\sqrt{|X_r^\beta|^2 + |X_i^\beta|^2} - \sqrt{|\hat{X}_r^\beta|^2 + |\hat{X}_i^\beta|^2}\|^2 \tag{3-25}$$

$$\mathcal{L}_{RI+Mag} = \mathcal{L}_{RI} + \mathcal{L}_{Mag} \tag{3-26}$$

3.4 添加人声惩罚项的损失函数

传统的 MSE 损失函数仅仅估计预测语音和干净语音之间的距离，以 $\mathcal{L}_{Mag}$ 为例，该损失函数的目标为减少估计人声和目标人声幅度之间的距离，但不会考虑方向性。如图3-10所示，估计语音 1(over-suppression) 和估计语音 2(under-suppression) 距离干净语音的损失相同，但对于语音识别任务来讲，估计语音 2 可能为更优的选择。因为该信号经受了较少了语音失真，而后端 ASR 系统对于语音失真异常敏感。

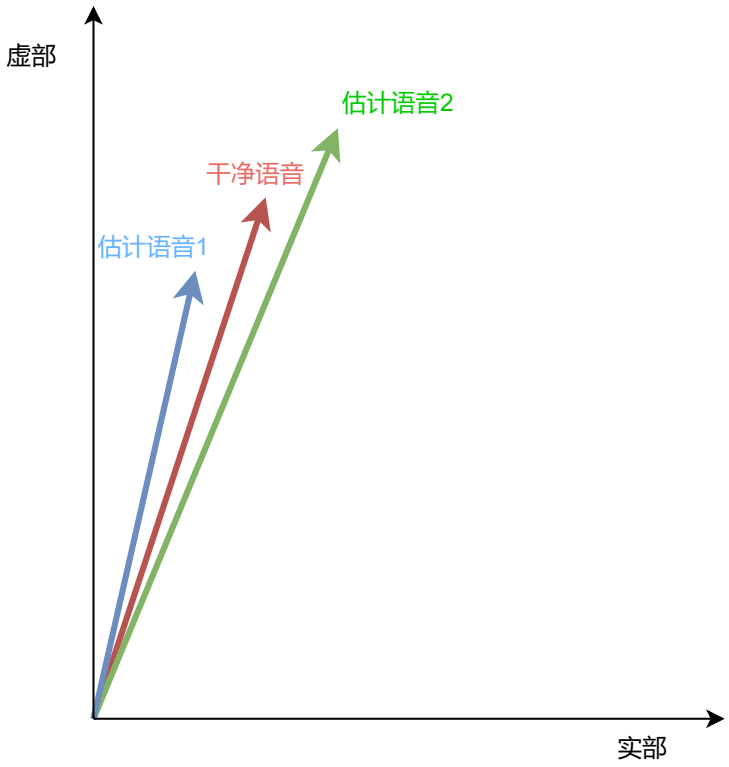


图 3-10　无人声惩罚项的语音估计示意图。

受此启发，本节提出的损失对于 under-suppresion 的情况容忍度更高，而对于 over-suppression 降低容忍度。因此本节对这两种情况添加不同的惩罚项，以减少语音失真：

$$g_{penalty}(x,a)=\begin{cases} x, if\, x \le 0 \\ ax, if\, x > 0 \end{cases} \tag{3-27}$$

$$\mathcal{L}_{penalty} = \|g_{penalty}(\sqrt{|X_r|^2+|X_i|^2}-\sqrt{|\hat{X}_r|^2+|\hat{X}_i|^2}, a)\|^2 \tag{3-28}$$

3.5 系数压缩和人声惩罚项结合

由于人声惩罚损失只考虑了幅度谱之间的距离，而忽略了相位之间的关系，因而本节提出将系数压缩和人声惩罚项结合的损失函数：

$$\mathcal{L}_{combine} = \mathcal{L}_{RI} + \|g_{penalty}(\sqrt{|X_r^\beta|^2+|X_i^\beta|^2}-\sqrt{|\hat{X}_r^\beta|^2+|\hat{X}_i^\beta|^2}, a)\|^2 \tag{3-29}$$

该函数结合了惩罚项损失和系数压缩损失的优点，对于人声的优化更加敏感。

3.6 语音增强评价指标

为了验证训练的模型在语音质量和可懂度方面的能力，需要通过不同的指标去衡量。常用的评价指标可以分为客观和主观评价两部分。主观评价是通过人类测试者主观听觉来感受语音质量的好坏，比较符合人类的听觉认知，但需要有大量的测试者参与，需要花费大量的人力财力。而客观评价指标通过既定的规则来计算相应的分数，根据分数来评价语音的质量。

3.6.1 信号失真比

信号失真比（signal to distortion ratio, SDR）通常和信号干扰比（signal to interference, SIR）以及信号人造成分比（signal to artifacts ratio, SAR）统称为基于信号分解的语音分离评价指标[22]。其中，SDR 代表信号整体的失真情况，SIR 表示抑制干扰的能力，SAR 表示产生人造干扰成分的情况，他们的单位均为分贝（dB）。

基于 Vincent 等人[22] 的研究，假设估计的目标语音为 $\hat{s}$，首先我们需要将其拆解为目标信号成分 s，干扰人声 e_{interf}，噪声误差成分 e_{noise} 以及人造误差成分 e_{artif}：

$$\hat{s} = s + e_{interf} + e_{noise} + e_{artif} \tag{3-30}$$

因此，SDR 的公式可以表示为：

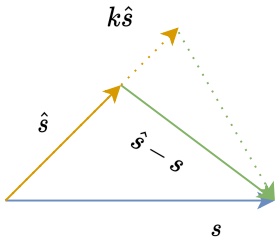
$$SDR(\hat{s}, s) = 10log_{10}\frac{||s||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} = 10log_{10}\frac{||s||^2}{||\hat{s} - s||^2} \tag{3-31}$$

结合上述定义和公式可以看出，SDR 主要是表示的是目标语音信号相对于其他成分所表现出来的整体分离效果。然而，我们需要注意到，不同数据集本身混合语音信号的平均 SDR 水平就不同，我们直接通过两者预测语音信号的平均 SDR 水平来对比某语音分离模型在不同数据集上的效果，这可能会存在些许的片面。对比不同数据集上的效果，常常会使用 SDRi 评价指标来替代，其表示的是应用语音分离模型后 SDR 的增益，定义如下：

$$SDRi(\hat{s}, s) = 10log_{10}\frac{||s||^2}{||\hat{s} - s||^2} - 10log_{10}\frac{||s||^2}{||y - s||^2} \tag{3-32}$$

3.6.2 尺度不变信号失真比

尺度不变的信号失真比（scale invariant signal to distortion ratio, SI-SDR）是信号失真比 SDR 评价指标的改进版本。如图3-11所示，当预测的信号 $\hat{s}$ 变换尺度变为 $k\hat{s}$ 时，SDR 的评价指标分数值也会发生变化，这使得该指标对于语音的响度非常敏感。



$$SDR = 10log_{10}\frac{||s||}{||\hat{s} - s||}$$

$$\neq 10log_{10}\frac{||s||}{||k\hat{s} - s||}$$

图 3-11 SDR 对信号响度大小比较敏感。

基于上述问题，Jonathan 等人提出了尺度不练的信号失真比 SI-SDR，该方法将预测的语音信号 $\hat{s}$ 投影到目标语音信号 s 的方向以及垂直于 $\hat{s}$ 的方向，分别用 s_{target} 和 e_{noise} 符号表示，该指标可以表示为：

$$\begin{cases} s_{target} = \frac{\hat{s}^{\mathrm{T}}s}{||s||^2}s \\ e_{noise} = \hat{s} - s_{target} \\ SI\text{-}SNR(s, \hat{s}) = 10log_{10}\frac{||s_{target}||^2}{||e_{noise}||^2} \end{cases} \tag{3-33}$$

3.6.3 短时客观可懂度

短时客观可懂度（short-time objective intelligibility, STOI）是衡量语音可懂度的重要指标之一[23]。它能将可懂度只分为能被听懂和不能被听懂两种情况，即

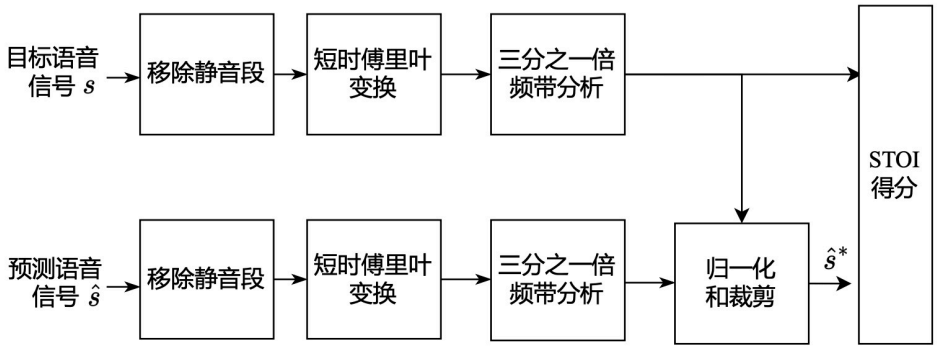


图 3-12　　短时客观可懂度 STOI 的原理流程。

二值判断，因而 STOI 的取值范围会被量化到 0-1 之间，代表这语音信号中能被正确理解的占比，STOI 的流程如图3-12所示。

3.6.4　字符错误率

中文的基本组成单位为字符，因此常常使用字符错误率（Character error rate）来衡量 ASR 效果的好坏。计算公式可以表示为：

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{3-34}$$

假设有一个参考例句 Ref 和一段 ASR 系统转写语音后生成的预测文本 Hyp。带入上面公式，S 表示将 Hyp 转化为 Ref 时发生的替换数量，D 表示将 Hyp 转化为 Ref 时发生的替换数量，I 代表将 Hypo 转化为 Ref 时发生的插入数量，N 代表 Ref 句子中总的字数或者英文单词数。C 代表 Hyp 句子中识别正确的字数。即原参考句子总字数 N = S+ D + C。

3.7　数据和实验介绍

3.7.1　数据介绍

本章实验采用 AISHELL-1[24] 和部分 DNSChallenge[25] 的数据。

AISHELL-1 是由北京希尔公司发布的一个中文语音数据集，其中包含约 178 小时的开源版数据。该数据集包含 400 个来自中国不同地区、具有不同的口音的人的声音。录音是在安静的室内环境中使用高保真麦克风进行录音，并采样降至 16kHz。通过专业的语音注释和严格的质量检查，手动转录准确率达到 95％以上。该数据集包括训练集、验证集和测试集。其中训练集包含 340 位演讲者的 120098 句话; 验证集包含了来自 40 个说话者的 14,326 个话语; 测试集包含来自 20 个说话者的 7176 个话语。对于每个说话人来讲，大约 360 句话 (约 26 分钟)。

表 3-2 AISHELL-1

Subset	Duration (h)	#Male	#Femal
训练集	150	161	179
验证集	10	12	28
测试集	5	13	7

DNS 挑战赛旨在激励研究者在语音增强方向的创新，以实现语音质量的提升。本章使用 DNS 语音混合脚本[1]对数据进行混合。干净语音采用 AISHELL-1 中的数据，噪声选取 DNS 数据集中的噪声，其中噪声数据来源于 AudioSet[2] 和 Freesound[3]两个数据集，按照 snr -5 ~ 20 进行混合。训练集中混合语音长度设为 3s，总时长设为 30h，包括 27,000 条音频，验证集包括 14,334 条音频，测试集包括 7,176 条音频。图3-13为其中一条音频加噪前后的语谱图对照。

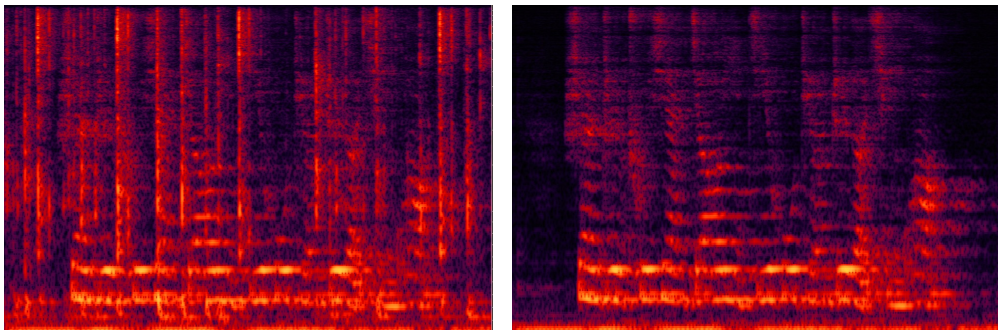


图 3-13 干净语音和带噪语音的语谱图对比。

3.7.2 实验设置

本章选择 Adam 作为优化器，初始学习率设为 10^{-3}，当损失在验证集上连续 12 次不下降，或者训练轮次达到 100 时，则停止训练，使用 ESPNet [4]的框架进行训练。STFT 和 iSTFT 的窗长和窗移分别为 320 ms 和 160 ms。

[1]https://github.com/microsoft/DNS-Challenge
[2]https://research.google.com/audioset/
[3]https://freesound.org/
[4]https://espnet.github.io/espnet/tutorial.html

3.8 结果分析

表格3-3显示的为测试集中干净的语音和混合语音在不同 SNR 条件下的 CER 指标差别。从结果对比中可以看出，当 SNR 越小，即噪声所占的比重越大时，对于后端语音识别的性能的影响越大。用来测试 CER 的 ASR 模型为 ESPNET 框架下使用干净的 AISHELL-1 语音训练的模型。

表 3-3 混合语音和干净语音在不同 SNR 情况下的 CER 指标对比。

Data	data set	SNR					
		-5 ~ 0	0 ~ 5	5 ~ 10	10 ~ 15	15 ~ 20	avg
clean	dev	＼	＼	＼	＼	＼	4.5
	test	＼	＼	＼	＼	＼	4.9
mixture	dev	46.7	35.9	27.0	20.1	16.5	29.7
	test	48.9	36.5	31.0	24.1	22.2	32.9

表 3-4 不同的损失函数在增强性能上的影响。所有的指标都是越大越好，代表增强模型的去噪效果越好。

Loss	data set	SAR	SDR	SI-SNR	STOI(%)
$\mathcal{L}_{MSE}$	dev	16.0	16.0	15.32	82.66
	test	15.57	15.57	14.87	80.09
$\mathcal{L}_{RI}$	dev	15.0	15.0	13.69	82.70
	test	14.65	14.65	13.38	80.37
$\mathcal{L}_{RI+Mag}$	dev	15.16	15.16	13.90	83.41
	test	14.84	14.84	13.61	81.16
$\mathcal{L}_{penalty}$	dev	**16.25**	**16.25**	**15.91**	**86.30**
	test	**15.80**	**15.80**	15.30	82.30
$\mathcal{L}_{combine}$	dev	16.19	16.19	15.86	86.27
	test	15.74	15.74	**15.42**	**84.36**

3.8.1 增强指标

表3-4显示的为不同损失函数在 SAR，SDR，SI-SNR 以及 STOI 上指标的影响。$\mathcal{L}_{MSE}$ 仅仅驱使模型让输出语音和目标语音之间的距离更接近，因此该损失在去除噪声的同时也会噪声部分语音失真。$\mathcal{L}_{RI}$ 和 $\mathcal{L}_{RI+Mag}$ 损失在原始的 MSE 基础上对频谱信号进行了压缩，压缩系数 $\beta = 0.5$，缩减了频谱中的最大值和最小

值之间的距离，避免模型过度优化频谱能量较大的部分。$\mathcal{L}_{penalty}$ 在原始 MSE 的基础上添加了惩罚项 $\alpha = 3$，对于估计语音和目标语音失真的部分给予更大的权重，从而驱使模型向着减少失真的部分优化。表中显示，几种损失函数对增强指标的影响，$\mathcal{L}_{penalty}$ 取得了最好的效果，在信号指标 SAR，SDR 以及 SI-SNR 上以及语音可懂度 STOI 上取得了最好的效果。而添加了压缩系数的损失函数 $\mathcal{L}_{RI}$ 和 $\mathcal{L}_{RI+Mag}$ 则表现较差，猜测可能是添加了压缩系数之后，频谱能量较低的部分得到了优化，频谱能量较高的部分得到了较少的关注，因而在信号指标上表现不如从前。

表 3-5　不同的损失函数对 CER 指标的影响，CER 越小越好。

Loss	data set	SNR					
		-5 ~ 0	0 ~ 5	5 ~ 10	10 ~ 15	15 ~ 20	avg
$\mathcal{L}_{MSE}$	dev	59.1	49.0	39.9	30.4	24.4	41.0
	test	62.7	52.2	45.9	36.5	32.7	46.3
$\mathcal{L}_{RI}$	dev	56.9	46.2	37.2	28.8	22.9	38.9
	test	60.0	48.3	42.6	34.2	30.9	43.5
$\mathcal{L}_{RI+Mag}$	dev	54.9	44.7	35.7	27.9	22.4	37.6
	test	58.5	46.8	40.8	33.4	30.1	42.3
$\mathcal{L}_{penalty}$	dev	41.6	32.2	25.7	20.9	17.5	28.9
	test	44.7	34.8	29.9	24.9	23.6	32.7
$\mathcal{L}_{combine}$	dev	41.5	32.1	25.5	20.7	17.2	**27.7**
	test	44.5	34.2	29.9	24.5	23.1	**31.6**

图3-14展示了测试集上 STOI 和 SI-SNR 的数量分布，由图中可以看出，使用 $\mathcal{L}_{combine}$ 损失函数时，在 STOI 和 SI-SNR 上表现都优于其他函数。

3.8.2 CER 指标

接下来对比不同的损失函数在 CER 指标上的差别，结果如表3-5所示。可以看到传统的 $\mathcal{L}_{MSE}$ 损失函数对人声造成了巨大的损失，使用该损失增强后的语音，其效果表现比混合语音还要差，该结果也证明了本章的论点，也就是增强模型会损害语音，造成语音失真。$\mathcal{L}_{RI}$ 的结果好于 $\mathcal{L}_{MSE}$ ，证明对语音幅度谱进行压缩的方法能有效平衡不同幅度之间的差距，提升感知质量，减少语音失真。$\mathcal{L}_{RI+Mag}$ 的损失函数证明了 Wang 等人[21] 提出的幅度谱损失能提升语音质量。$\mathcal{L}_{penalty}$ 选择 $\alpha = 3$ 作为惩罚项系数的值，其结果表明该函数对于语音拟合方向的控制是有效的，减少了增强过程中语音的损失，同时也有效消除了部分噪声。该结果甚至

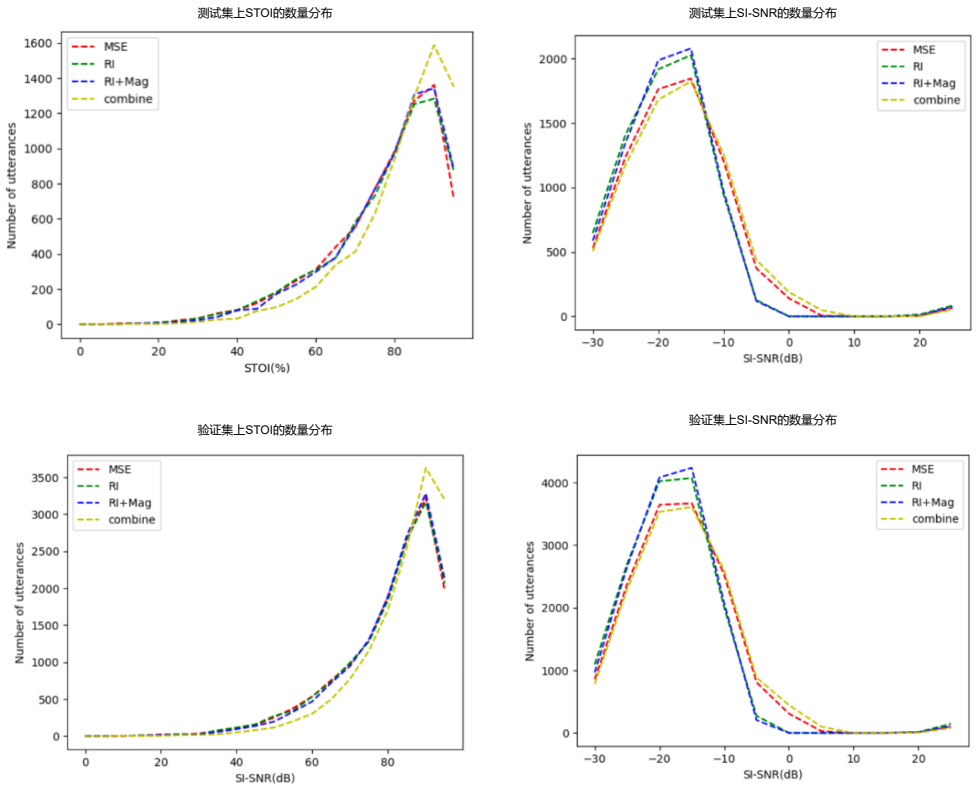


图 3-14 不同损失函数下 STOI 和 SI-SNR 的数量分布。

优于表3-3中混合语音的表现。鉴于系数压缩和人声惩罚项两种损失函数的有效性，本章继续验证了 $\mathcal{L}_{combine}$ 损失函数的效果，该函数如公式3-29所示，结合了系数压缩和人声惩罚两种函数的优点，以期待达到对人声保持同时对噪声消除的目的。表格中的结果也证明了该损失函数的有效性，在所有损失函数中取得了最优的表现。

本章提出的损失函数在不同的 SNR 条件下都能取得优于 MSE 的效果，证明本文提出的损失函数的有效性，能解决增强模型对于目标人声损失的问题。此外，当 SNR 较低时，CER 指标的表现比较差，证明了噪声对于 ASR 系统的影响非常大，因此提出一种有效的去除噪声的方法显得尤为重要。

为了进一步分析语音识别错误时，哪种类型的错误占比最多，本章对插入错误，替换错误以及删除错误进行了分析，如表3-6所示。可以看出，替换错误是造成语音识别准确率下降的主要原因，即文本中的句子被识别成了其他的词语。例如将“广州二手住宅市场表现一直相对稳健”识别为“广州二手住宅市场表现一起消费明显”。其次为删除错误，即语音的某些单词没有被识别出，而造成空缺。例如将“但是改善型置业入市积极性下降”识别为“但是改善 *** 置业入市积极性下降”，缺少了字符“型”。影响最小的为插入错误，即语音中没有出现该读音，但是却识别出了该字符。例如“ 故宫博物院迎来九零周年 *** 院庆”识别为“故

表 3-6 不同损失函数上的插入错误（Ins），替换错误（Sub）和删除错误（Del）。

Loss	data set	Sub	Del	Ins	CER
$\mathcal{L}_{MSE}$	dev	35.8	3.6	1.7	41.0
	test	39.6	4.4	2.3	46.3
$\mathcal{L}_{RI}$	dev	33.8	3.7	1.4	38.9
	test	37.2	4.4	2.0	43.5
$\mathcal{L}_{RI+Mag}$	dev	33.3	2.6	1.7	37.6
	test	36.8	3.1	2.3	42.3
$\mathcal{L}_{combine}$	dev	**24.6**	**2.5**	**0.6**	**27.7**
	test	**27.5**	**3.1**	**0.9**	**31.6**

宫博物院迎来九零周年毕院庆”，多识别出了字符 “毕”。综合以上多种类型的识别错误，本章提出的 $\mathcal{L}_{combine}$ 能够有效降低各种类型的错误的概率，证明了该损失对于后端识别的有效性。

3.8.3 频谱对比

为了能够直观的观察到不同损失函数对于语音去噪以及人声失真的影响，本节对于不同损失函数增强后的语音进行频谱可视化，如图3-15所示。本节选择了 $\mathcal{L}_{MSE}$，$\mathcal{L}_{RI}$，$\mathcal{L}_{RI+Mag}$ 以及 $\mathcal{L}_{combine}$ 四种损失，进行对比。从图中看出使用 $\mathcal{L}_{MSE}$ 优化后的结果静音部分（黑色部分）和语音部分之间的割裂感非常明显，会造成一定程度上的语音失真。产生该问题的原因是，$\mathcal{L}_{MSE}$ 仅仅考虑减少目标语音和估计语音之间的距离，在优化过程中，静音段和人声段所占的比重会比较大，易于优化。

$\mathcal{L}_{RI}$ 和 $\mathcal{L}_{RI+Mag}$ 两种损失函数的效果和 $\mathcal{L}_{MSE}$ 差别并不是很明显。而 $\mathcal{L}_{combine}$ 的结果可以明显看出语音和静音之间的过度部分，说明该方法有效减少了语音失真，对于人声中如轻音等部分进行了很好的保留。

3.9 本章小结

语音增强是语音识别等后端任务的前端处理模块，其作用为消除混合语音中的噪声，但语音增强模块预测的语音，不仅仅会消除噪声，而且会对目标人声造成损失，这会严重影响 ASR 系统的性能，增强模块不仅仅不会提升后端的性能，反而会使性能下降，因此设计一种对 ASR 系统友好的前端模型是很重要的。为了提升语音识别系统的鲁棒性，本章从损失函数的角度出发，设计一系列的损

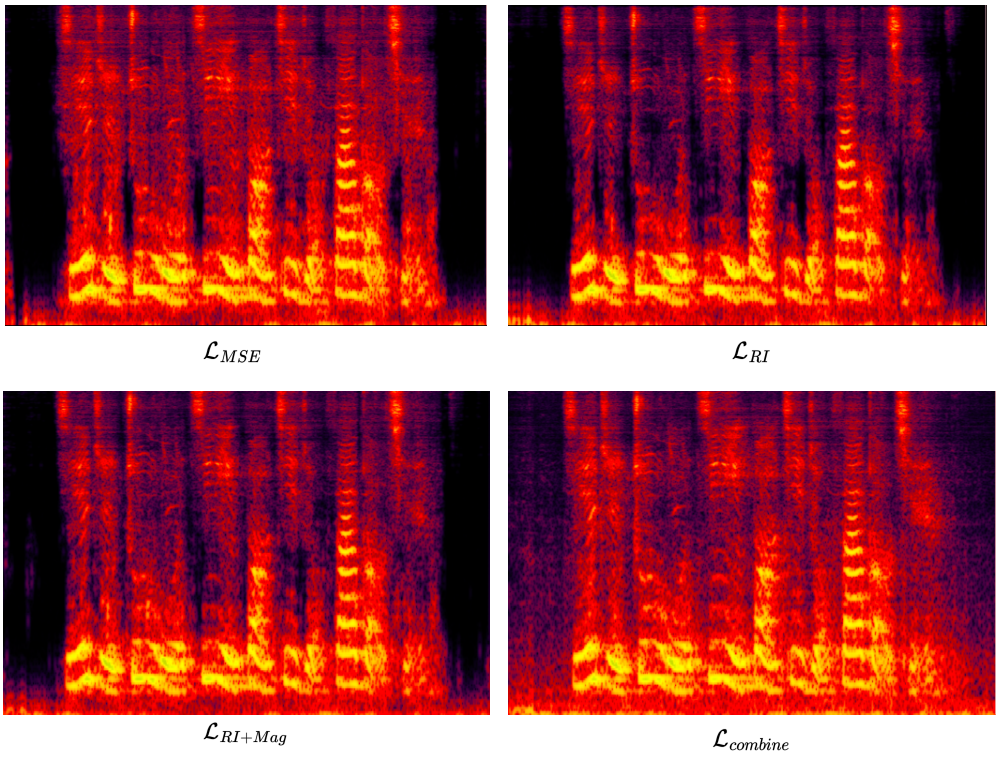


图 3-15 使用不同损失函数增强后的语谱图对比。

失函数来减少语音增强过程中的人声失真问题。本章提出几种不同的损失函数：系数压缩的损失函数 $\mathcal{L}_{RI}$，$\mathcal{L}_{RI+Mag}$，添加惩罚项的损失函数 $\mathcal{L}_{penalty}$ 以及结合系数压缩和惩罚项的损失函数 $\mathcal{L}_{combine}$。

通过在 AISHELL-1 数据上进行实验，结果验证了相比于传统的 $\mathcal{L}_{MSE}$ 损失，本章提出的几种损失函数能有效减少语音失真，在后端 ASR 系统上表现出了较好的效果，其中 $\mathcal{L}_{combine}$ 取得最好的效果，证明了系数压缩和添加人声惩罚项对于人声保持的作用。同时分析了不同损失函数在插入错误，替换错误以及删除错误的表现，结果表明替换错误是影响 ASR 系统的主要因素。

参考文献

[1] Xiong W, Wu L, Alleva F, Droppo J, Huang X, et al. The Microsoft 2017 conversational speech recognition system [C]. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018: 5934–5938.

[2] Saon G, Kurata G, Sercu T, Audhkhasi K, Thomas S, et al. English conversational telephone speech recognition by humans and machines [J]. arXiv preprint arXiv:1703.02136, 2017.

[3] Vincent E, Watanabe S, Barker J, Marxer R. The 4th CHiME speech separation and recognition challenge [J]. URL: http://spandh. dcs. shef. ac. uk/chime challenge {Last Accessed on 1 August, 2018}, 2016.

[4] Wang Z-Q, Wang D. A joint training framework for robust automatic speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24 (4): 796–806.

[5] Boll S. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Transactions on acoustics, speech, and signal processing, 1979, 27 (2): 113–120.

[6] Scalart P, et al. Speech enhancement based on a priori signal to noise estimation [C]. In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1996: 629–632.

[7] Fujimoto M, Kawai H. One-Pass Single-Channel Noisy Speech Recognition Using a Combination of Noisy and Enhanced Features. [C]. In INTERSPEECH, 2019: 486–490.

[8] Seltzer M L. Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays [C]. In 2008 Hands-Free Speech Communication and Microphone Arrays, 2008: 104–107.

[9] Li F, Nidadavolu P S, Hermansky H. A long, deep and wide artificial neural net for robust speech recognition in unknown noise [C]. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[10] Seltzer M L, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition [C]. In 2013 IEEE international conference on acoustics, speech and signal processing, 2013: 7398–7402.

[11] Park D S, Chan W, Zhang Y, Chiu C-C, Zoph B, et al. Specaugment: A simple data augmentation method for automatic speech recognition [J]. arXiv preprint arXiv:1904.08779, 2019.

[12] Liu B, Nie S, Liang S, Liu W, Yu M, et al. Jointly Adversarial Enhancement Training for Robust End-to-End Speech Recognition. [C]. In Interspeech, 2019: 491–495.

[13] Ma D, Hou N, Xu H, Chng E S, et al. Multitask-based joint learning approach to robust asr for radio communication speech [C]. In 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021: 497–502.

[14] Sato H, Ochiai T, Delcroix M, Kinoshita K, Kamo N, et al. Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition [C]. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 6287–6291.

[15] Fan C, Yi J, Tao J, Tian Z, Liu B, et al. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 198–209.

[16] Iwamoto K, Ochiai T, Delcroix M, Ikeshita R, Sato H, et al. How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr [J]. arXiv preprint arXiv:2201.06685, 2022.

[17] Luo Y, Chen Z, Yoshioka T. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation [C]. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 46–50.

[18] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation [J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27 (8): 1256–1266.

[19] Oord A v d, Dieleman S, Zen H, Simonyan K, Vinyals O, et al. Wavenet: A generative model for raw audio [J]. arXiv preprint arXiv:1609.03499, 2016.

[20] Zhao Y, Wang D, Xu B, Zhang T. Monaural speech dereverberation using temporal convolutional networks with self attention [J]. IEEE/ACM transactions on audio, speech, and language processing, 2020, 28: 1598–1607.

[21] Wang Z-Q, Wang P, Wang D. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR [J]. IEEE/ACM transactions on audio, speech, and language processing, 2020, 28: 1778–1787.

[22] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation [J]. IEEE transactions on audio, speech, and language processing, 2006, 14 (4): 1462–1469.

[23] Taal C H, Hendriks R C, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19 (7): 2125–2136.

[24] Bu H, Du J, Na X, Wu B, Zheng H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline [C]. In 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA), 2017: 1–5.

[25] Reddy C K, Dubey H, Gopal V, Cutler R, Braun S, et al. ICASSP 2021 deep noise suppression challenge [C]. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 6623–6627.