发展

语音增强算法引入深度学习大抵是一种必然，讨论基于深度学习的语音增强算法我们必须首先把目光放在Prof. DeLiang Wang的OSU-PNL身上。Prof. Wang 是用于语音分离的计算听觉场景分析(CASA)领域的代表任务，他提出用理想二值掩蔽(IBM)解决语音增强（分离）问题。理想二值掩蔽，其中掩蔽广泛用于图像处理之中，具体到语音增强领域是指对选定的时频区域进行遮挡以控制要处理的区域，二值则是对每个时频点量化的精度非0即1，理想二值掩蔽的含义则是根据纯净语音和噪声之间的能量关系，将语音能量占主导的时频点标记为1，噪声能量占主导的时频点标记为0，由此得到一个滤波器，对带噪特征进行滤波，在语音时频稀疏性假设下留下的即使语音成分。于是，熟悉机器学习的读者应该很容易发现，语音增强问题在IBM的定义下可以被看做一个分类问题，当时OSU-PNL的Y. Wang已经开展了支持向量机(SVM)估计IBM的研究。机器学习中特征提取和目标选择十分重要，作为机器学习模型的SVM当然也不例外。Y. Wang的研究当然也包括评估众多利用听觉特性构造的手工特征对SVM进行语音增强的性能影响，在深度学习大行其道的今天，利用深度学习提取更有效的特征的思维并不难理解，而Y. Wang在2013年也的确选择这样去做。如果你对机器学习有一定了解的话，应该了解机器学习任务可以分为分类任务和回归任务两种，基于回归模型的语音增强算法也于2014年被中科大USTC SPRAT的Y. Xu 等人提出。基于回归的语音增强模型直接利用神经网络建立从带噪语音谱特征到纯净语音谱特征的映射，利用网络直接生成增强的对数功率谱。如果说Y. Wang的工作是基于监督学习的CASA算法的延申的话，Y. Xu的这项工作则更具有基于统计模型的传统语音增强算法的风格。工作中沿用了MMSE-LSA中的对数谱特征以及传统语音增强方法中人耳对相位信息不敏感的假设。然而利用分类和回归去分类这两个极具代表性的工作并不利于对后续方法的研究，因为后续的理想比值掩蔽(IRM)、相位敏感掩蔽(PSM)等训练目标也明显受到了传统语音增强统计模型的影响，这类掩蔽估计由于其定义也不再能被视为分类问题（而同样是回归问题）。这类掩蔽本质上是在设计某种意义上最优的滤波器，毕竟，回顾IBM的介绍“IBM的含义正是根据每个时频点上语音和噪声之间的能量关系，将语音能量占主导的时频点标记为1，噪声能量占主导的时频点标记为0”中“主导”二字充满了不安全感。一个极端的例子是，在极低信噪比情况下IBM会出现非常多的0值导致理想掩蔽的滤波结果的语音质量和可懂度都会十分糟糕。IRM将IBM的硬分类变成了一种软分类，每个时频点的滤波器系数定义为语音能量和语音噪声二者能量之和的比值。熟悉传统语音增强算法的读者不难发现，这个定义和频域维纳滤波算法尽管并不相同但还是十分相像。因此，Prof. Wang在综述中从训练目标角度将上述两大类工作分别定义为基于掩蔽(masking-based)和基于谱映射(mapping-based)的算法用于分类，这种分类方式得到了广泛地接受。

尽管上述内容足以概括基于深度学习的语音增强在早期阶段的历史，然而除了训练目标，模型架构的发展同样值得我们关注。不过在这一段中我们对文献中的训练目标使用掩蔽还是谱特征并不感兴趣，因此，如无必要在介绍中不会涉及映射目标。另一方面，虽然本节介绍的是深度学习语音增强的发展，然而为了保证叙述的连贯性（毕竟对于初学者通史似乎比通鉴更友好），本段刻意忽略了算法的提出时间，只保证了相关联工作的顺序。由于语音具有时序性，因此时间建模是网络结构设计上不可忽视的问题。前面提及的方法都是利用全连接网络，考虑时序关系的方式也十分朴素——将当前帧前后几帧组合在一起形成上下文窗(context window)后送入网络。很明显，这种方式并不足以建模长时信息。而循环神经网络(RNN)天然适合时序建模，因此2017年OSU-PNL的J. Chen和USTC-SPRAT的Sun分别开展了基于LSTM的语音增强模型的工作。这类网络和全连接网络唯一结构上的区别在于全连接网络的其中几层（通常是前几层）被替换成了LSTM层，但其在增强性能、未见说话人和噪声类型的泛化能力上都明显优于全连接网络。RNNs的提出使得目前基于深度学习的语音增强模型几乎全部放弃了全连接网络。另一个需要提及的内容在于卷积神经网络的应用，熟悉图像分割的读者应该已经感受到了语音增强与图像分割的相似之处，因此卷积神经网络的引入也十分自然。不过在介绍卷积神经网络在语音增强的应用之前，有必要至少提一点语音增强与图像分割的区别，尽管我们可以将语音特征看成一张图片，但是由于语音增强有很多实时应用场景，因此在这张“图片”时间维度的处理需要十分谨慎“实时陷阱”。卷积层由于“局部连接、权重共享”的特点相比全连接层和循环层拥有更少的参数，2015年清华 SATLab的L. Hui提出的maxout卷积网络首次在语音增强领域性能超过全连接网络。而后中央研究院Bio-ASP Lab的S. Fu和卡内基梅隆大学的S. R. Park分别将图像领域常用的普通卷积神经网络和卷积编解码器结构成功应用到了语音增强中。由于卷积编解码器结构对于时频点级预测任务的天然适配，基于卷积编解码器的语音增强模型逐渐成为主流。自然地，将卷积层的特征提取能力和循环层的时序建模能力结合的工作应运而

生。2018年H. Zhao在微软亚研完成了卷积层特征提取、LSTM时序建模、最后由全连接层进行幅度谱估计的EHNet，同年OSU-PNL的K. Tan将LSTM插入卷积编解码器之间提出一种卷积循环网络(CRN)。除了RNN适合序列建模外，时序卷积网络 (TCN)利用一维空洞卷积实现了并行时序建模过程。2018年K. Tan引入门控(gating)TCN取代RNN构造了一种全卷积增强网络。

随着基于深度学习的幅度谱估计算法性能的快速发展，两个很有意思的现象出现了。一是相对简短可以介绍完的，是深度学习反求诸传统，深度学习作为一种参数估计器为传统语音增强中的MMSE、MMSE-LSA和卡尔曼滤波等算法提供依赖先验假设更少且更为准确参数估计。一般将这类深度学习与传统算法结合的方法称为混合式模型，然而混合式模型类工作相比性能提升更重要的贡献在于让基于深度学习的语音增强渐渐褪去纯机器学习或深度学习意味的工作，而逐渐带有语音信号处理的特色。当然，相比于上述这种"主动"结合语音特色，另一种相对"被动"的结合方式产生于另一个有意思的现象——基于深度学习语音增强的性能第一次遇到了瓶颈。当然人们很快意识到了问题——这一性能限制是只进行幅度谱估计带来的，相位恢复似乎对人耳感知是重要的。以至于解决该问题的方案出现得如此之快在今天看来似乎并没有瓶颈的意味在（也可能本身也并不是瓶颈，只是这里为了阐释后续勃勃生机万物竞发的发展强行解释为瓶颈）。在相位恢复重要性思潮下，用于同时进行幅度估计和相位恢复的方案大方向上可以分为时域的方案和谱域的方案。不论是哪种方案他们都是为了解决由于相位谱没有清晰的模式结构导致其不易优化的问题。首先介绍时域模型，原因依旧是因为介绍起来可以比较简略。时域模型的兴起主要由于其在语音分离领域大获成功，以一种波形到波形的映射方式提供了一种解决相位谱不易优化问题的方案——放弃建模相位。不过也许是由于噪声模式的复杂性，抑或是房间冲激响应的敏感多变，时域模型在噪声混响场景下的性能和鲁棒性使其在语音增强中并不能成为主流。网络结构上和之前的幅度谱架构基本还是类似的，有卷积编解码器结构也有卷积循环（或TCN）网络，只不过对应的卷积层大多用一维卷积代替了二维卷积（当然也同样有使用二维卷积作为编解码器的架构）。相比之下，谱域的方案解决相位估计的方式则略显复杂。众所周知，复数可以由幅度相位表示也可以由实部虚部表示。其中实部虚部表示和之前幅度谱模型一样，主要分为基于掩蔽和基于谱映射的方式。2016年OHU-PNL的Williamson提出了一种复数比值掩蔽(cRM)，分别估计实部和虚部的掩蔽；而2017年S. Fu提出了一种谱映射的CNN，2019年K. Tan将CRN应用到复数谱映射直接利用网络估计增强语音的实部和虚部。2019年首尔国立大学的H. S. Choi提出了一种基于极坐标系的复值掩蔽策略(pcRM)，通过网络估计的cRM得到有界的幅度掩蔽和无界的相位掩蔽。尽管这篇文章成功应用了复值网络并且西工大ASLP的Y. Hu继复值U-Net发展为复值CRN取得了令人瞩目的性能，然而本篇文章对于在不同损失函数情况下的几种典型复值掩蔽的分析可能更加有趣：尽管这类复数目标算法都宣称致力于相位的恢复，然而其带来的性能提升究竟是源于相位的恢复还是噪声成分幅值的抑制（相位可能调整的极少）值得探究，D. Yin等人同样通过实验发现了使用cRM虚部几乎为0的问题。而为了体现相位的修复，基于幅度相位表示的方案得到了考虑，其中包括采用双分支结构同时估计幅度和相位，通过多任务学习的方式达到相位恢复的方法；也有一些算法采用了更具结构性的相位目标与幅度谱同时和先后估计，比如OHU-PNL的Z. Wang提出的群时延(GD)以及西工大CIAIC的N. Zheng提出的瞬时频率倒数(IFD)。

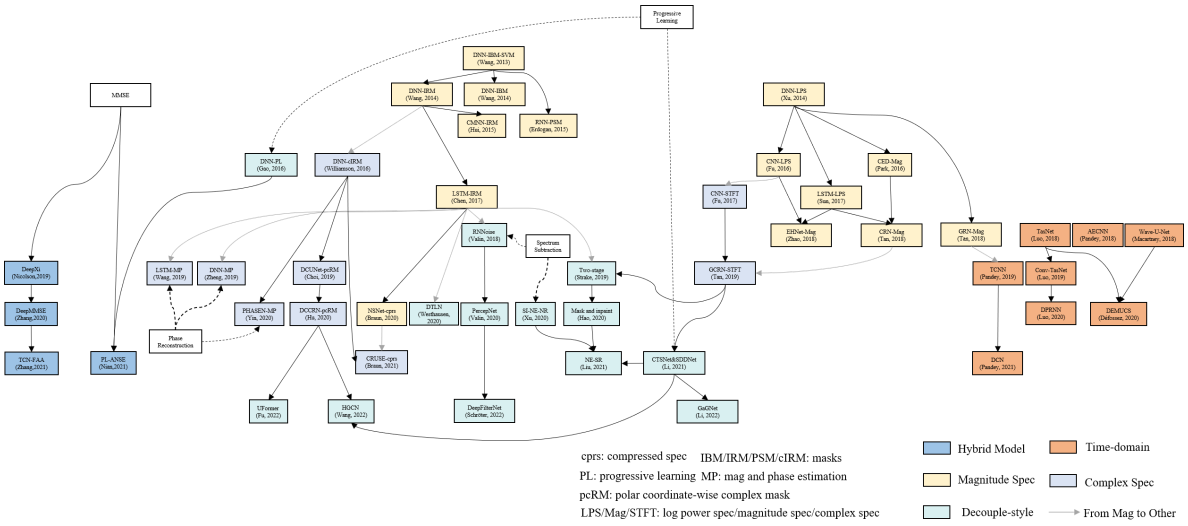
尽管Z. Wang的分析要晚于接下来要介绍的一些工作，然而在此处提及也许是必要。尽管这篇分析的重点在于解释RI+Mag损失函数为何是有效的，但其提供的幅度估计和相位估计的补偿作用无疑为解耦风格工作有效性提供了一种合理解释。所谓幅度和相位估计的补偿作用，是指在迫使网络输出逼近理想复数谱时，幅度谱估计的准确度会大打折扣。而如Griffin-Lim相位重构的很多相位估计算法都在幅度谱估计的基础上进行。一个解决幅度相位估计补偿作用的想法是，将复数谱的优化问题解耦为幅度初估计和包括相位信息在内的"全信息"估计两部分。尽管这仍属于复数谱估计的一类方法，然而这类算法在近年来备受关注且性能优异，因此有必要单独作为成段介绍。2020年，奥尔登堡大学通信声学的N. L. Westhausen先后利用两次信号变换——首先利用STFT幅度谱上进行幅度估计，而后再利用一维卷积生成的可学习的分析/合成基函数直接将语音时域波形映射到高维空间进行语音和噪声成分的分离，利用时域模型的方式完成相位恢复。哈工大SPLab的Z. Du在Mel谱(一种符合人耳听觉特性的幅度谱变换)域进行降噪，而后利用语音合成声码器根据增强的Mel谱合成语音波形从而实现相位信息的引入。中科院声学所IACAS-Lab9的A. Li则在2021年提出一种完全在STFT域上先后完成幅度谱估计和复数谱残差修正的框架，利用复数谱残差修复相位信息。完全在STFT域上处理两部分任务的一个好处是，原本级联的两个子任务出现了并行优化的可能，2022年A. Li和西工大ASLP的Y. Fu分别提出了幅度估计和复数谱修正的两种并行架构。其实，从复数谱的角度考虑，这种解耦是由于幅度相位估计的补偿作用导致幅度谱估计不准确造成的，因此，除了首先约束幅度谱后再根据相位差微调增强结果，也可以先得到复数谱结果后补偿不准确的幅度信息。于是在2022年，北交大IIS的T. Wang则先进行复数谱估计再利用谐波特征进

行幅度谱修正的方法。而解耦风格的工作也并非只由从训练目标解耦一种方式。从语音被噪声污染的过程角度，早在2016年USTC-SPRAT的T. Gao就提出以不同信噪比的被污染语音作为目标将一次降噪过程分解为渐进地、多阶段地、逐信噪比地提升。从语谱高低频特性差异角度，搜狗的J. Li对语谱进行子带分解，多阶段地处理各个子带。而对语音特性进行分析，语音可以分解为包络和周期两部分，2018年，J.-M. Valin首先利用GRU估计Bark域(另一种符合人耳听觉特性的幅度谱变换)的增益因子，而后利用信号处理的方式进行基频滤波抑制谐波间噪声残留，而后在2020年，其又对模型中各部分进行进一步提升，以可观的复杂度达到超过众多复数谱模型的性能。而对该工作神经网络化最杰出的代表当属PAU模式识别实验室的H. Schröter提出的在估计包络增益因子外利用网络估计深度滤波器(deep filtering)系数的框架。 除此之外，语音增强任务也可以理解为噪声去除和语音恢复两部分，因为单通道语音增强具有噪声抑制和语音失真之间的折衷问题，彻底去除噪声势必会对语音造成严重损伤。因此，一种先进行激进地噪声抑制而后再恢复在降噪过程中的失真语音的框架首先由TU Braunschweig通信技术所信号与机器学习组的M. Strake于2019年提出，2020年内蒙古大学IMUAI的X. Hao在该思路的基础上引入计算机视觉的概念形象地将该过程命名为“masking and inpainting”即将所有噪声的时频点去除后根据语谱相关性修补被破坏的语音时频点。而传统的谱减法则从带噪语音的合成过程反推，将语音增强看做噪声估计和移除噪声成分的过程。2020年，哥伦比亚大学C2G2的R. Xu受传统谱减法的启发将降噪过程分解为噪声的“masking and inpainting”过程(利用VAD得到静音段而后通过静音段推测出完整的噪声谱)并将噪声信息和带噪语音信息一同作为输入完成降噪。2021年，IACAS-Lab9的W. Liu将基于训练目标解耦的思路与基于带噪语音合成的解耦思路结合首先利用多任务学习幅度域对语音降噪并直接估计噪声成分，而后修正复数谱恢复过度抑制的语音。需要注意的是，尽管本段从解耦角度介绍了以上算法，但是这些算法在提出时受到的启发性工作和动机可能并非如此。这种分类目前也并不作为一种共识，而更多的是一家之言。

需要注意的是，网络架构和学习目标虽然的研究虽然更多然而只是性能提升的一方面，更长时间范围的输入特征以及精心设计的损失函数对模型性能的提升可能更加有益，其中损失函数由于不会在推理过程中引入额外的时延和复杂度而备受关注。一个有力的证据是在ICASSP2021 DNS-Challenge中微软提供的结合一种幂律压缩的复数谱均方误差损失的复数掩蔽网络相比包括解耦风格框架在内的一众高参数量和计算量方案仍表现出最优的性能。损失函数包括基于距离的损失函数、基于能量比的损失函数和基于感知指标的损失函数等。常见的距离表征最常用的莫过于基于范数的损失函数L_p损失，即平均绝对误差(MAE)和均方误差(MSE)，这两类误差都被用于度量时域波形、掩蔽、频谱以及嵌入式特征。时域波形的L_p度量形式相对简单，表示为$\mathcal{L}_{time-L_p}=\|s-\hat{s}\|_p$，有时会加入对噪声项$L_p$损失作为正则，以提升语音估计性能，有文章表示，对于时域样本而言MAE相比MSE更关注小信号并具有更强的噪声抑制能力。除了IBM有使用交叉熵损失进行度量外，其余掩蔽大多都会采用L_p损失直接计算掩蔽间的误差$\mathcal{L}_{mask}=\|M-\widehat{M}\|_p$，J. Valin提出了幂律压缩的掩蔽MSE损失$\mathcal{L}_{RNNoise}=\|M^c-\widehat{M^c}\|_2$，利用指数项$c$控制噪声抑制的激进程度。然而，目前普遍认为基于信号近似(SA)技术的损失相比直接度量掩蔽间差异对语音增强的性能更有益：$\mathcal{L}_{SA}=\|S-\widehat{M}\cdot Y\|_p$。可以看出，此时对掩蔽的范数损失已经和基于频谱的范数损失基本一致，只需将滤波项$\widehat{M}\cdot Y$改为网络映射的谱$\widehat{S}$：$\mathcal{L}_{spec-L_p}=\|S-\widehat{S}\|_p$。当然$S$和$\widehat{S}$可以是幅度谱或其变换(如对数谱、Mel谱等)，也可以是复数谱。由于幅度谱和复数谱是如此常用，这里我们分别给出其具体形式：$\mathcal{L}_{Mag}=\mathbb{E}[|A-\widehat{A}|^p]$和$\mathcal{L}_{com}=\mathbb{E}[|\mathcal{R}(X)-\mathcal{X}(\widehat{X})|^p+|\mathcal{I}(X)-\mathcal{I}(\widehat{X})|^p]$(其中$A$和$X$分别代表幅度谱和复谱)。$\mathcal{L}_{spec-L_p}$还可能根据如AMR编码等心理声学机制进入加权，以期望其更符合人耳听觉，而复数谱范数损失又常与幅度损失结合以缓解幅度相位估计的补偿作用，该损失被称为RI+Mag损失：$\mathcal{L}_{RI+Mag}=\alpha\mathcal{L}_{Mag}+\beta\mathcal{L}_{com}$。在此基础上，与时域波形$L_p$损失结合的$\mathcal{L}_{time-spec-L_p}=\mathcal{L}_{RI+Mag}+\mathcal{L}_{time-L_p}$、基于幂律压缩的RI+Mag损失$\mathcal{L}_{cprs}=\alpha\mathbb{E}[|A^c-\widehat{A^c}|^2]+\beta\mathbb{E}[|\frac{X\cdot A^c}{A}-\frac{\widehat{X}\cdot\widehat{A^c}}{\widehat{A}}|^2]$以及多分辨率(假设有$I$种STFT点数对应的$I$个分辨率)的$\mathcal{L}_{cprs}$函数$\mathcal{L}_{MR-STFT}=\sum_i\mathcal{L}^i_{cprs},i=1,\cdots,I$都成为目前广泛使用的$L_p$损失。另外由于$\mathcal{L}_{cprs}$对噪声的抑制能力较强导致语音失真较多，一种非对称损失有时也被作为正则项$\mathcal{L}_{asym}=\mathbb{E}[|max(A^c-\widehat{A^c},0)|^2]$。可以看出，尽管有文献提及由于STFT频点分布更类似拉普拉斯分布因此建议使用$\mathcal{L}_1$范数损失，然而大多数工作更喜欢选择$\mathcal{L}_2$范数损失。L_p范数损失最后也见于对嵌入式特征的距离度量，利用wav2vec、PANN等预训练模型生成嵌入式特征之间的距离作为谱距离损失的正则项。此外，一些除了$\mathcal{L}_p$范数作为距离度量，KL散度也有一些工作将其与$\mathcal{L}_p$范数联合度量谱距离：$\mathcal{L}_{KL}=\mathbb{E}[\hat{X}\cdot log(\frac{\hat{X}}{X})]$。基于信号能量比的损失函数主要用于时域波形，常见的有SDR(及其Log形式)和SI-SDR损失，分别定义为$\mathcal{L}_{SDR}=-10log(\frac{|s|^2}{|\hat{s}-s|^2})$和$\mathcal{L}_{SI-SDR}=-10log(\frac{|\xi\cdot s|^2}{|s-\xi\cdot\hat{s}|^2})(\xi=\frac{s^T\hat{s}}{\hat{s}^T\hat{s}})$。尽管SI-

SDR被广泛应用于语音分离当中，但是由于其过多的语音失真导致最近的工作常将其与L_p范数谱距离损失联合使用。上述特征常被诟病并不能反映人耳的听感，因此，一些感知指标，如PESQ、ESTOI和CD，一度被引入作为损失函数训练模型。然而这类损失函数仅能提升被优化指标的性能，并没有带来其余指标的提升。要知道，目前的客观指标与人类的主观听感并不完全一致，单纯提升某些客观指标大多也并未带来主观听感的明显提升。于是近年来，这类感知指标类损失更多是以一种联合形式作为验证集损失被用于选择最优的模型。

尽管生成对抗网络(GAN)常作为一种不同的网络架构，然而这里我们更希望将其作为一种训练手段，即对抗性训练。众所周知，GAN由生成器和鉴别器两部分，在GAN在语音增强的使用中，其生成器可以用前文提到的众多网络代替，而之前提到的各种损失同样可以作为生成器损失函数的其中一项或等价于生成器损失函数。因此，GAN在语音增强中目前表现的作用，更类似于一种类损失函数”用于帮助网络专注于那些原本损失函数难以强调的噪声成分和语音伪影(artifacts)部分。于是，一言以蔽之GAN在语音增强的整体发展：之前的语音增强模型(如LSTM、UNet等)代入生成器，各式GAN的发展技术(如LSGAN、 Wasserstein GAN 、Relativistic GAN、Conditional GAN和 Geometric GAN等)代入鉴别器。如此导致GAN始终并未在语音增强领域至少在实时语音增强领域成为主流技术 ，这很可能与语音增强任务与图像生成、语音合成等生成类任务的差异过大有关，然而，在无监督或自监督情况下，包括GAN在内的生成式模型(还包括变分自动编码VAE等)也许是值得研究的方法。



自2013年至今基于深度学习语音增强已近十载，取得的成果颇丰，面临的问题却也不少。以上对基于深度学习的单通道语音增强的发展做了大致梳理，然对于目前处于发展之初的无监督(包括自监督)语音增强、超宽带语音增强、多模态语音增强、个性化语音增强等并未动笔，此外去混响等问题也并未讨论。成文匆匆，难免错漏，亦请诸君斧正。