环境鲁棒性

2019 年 1 月 2 日

1 环境鲁棒性简介

我们常有这样的体验，本来好好的语音输入法，在办公室里基本不会有什么错误，但在大街上使用时就会感到性能明显下降；挺好的语音助手，平常百试不爽，但在公交车上问个问题经常答非所问。这主要是因为实际应用场景中的声学环境非常复杂，这些复杂场景不可能在训练时被全部覆盖，因此形成识别场景与模型的不匹配，导致系统性能急剧下降。总体而言，声学场景的复杂性主要可归结为三类：

1. 背景噪音。实际应用场景中可能包含各种不同类型的噪音，如机器声、汽车引擎声、开门声、背景音乐声、其它人的谈话声等。语音信号在混杂这些噪声时会发生发生显著变化，引起识别性能的下降。

2. 混响和回声。在一个房间里，发音会在房间四壁反射，形成混响。房间越大，反射声音和原声音延迟越久，产生的混响效果越明显，在电话通信中，语音有可能通过对方的听筒和麦克风反射回来，形成回声。混响和回声会显著降低声音的清晰度，严重影响识别性能。

3. 信道差异。基于其不同的物理特性，不同麦克风录出的声音有显著区别；即使同一种麦克风录制的声音，在采样、预处理、传输、保存过程中也会产生各种差异，如增益设置、编码编码方式、压缩方法等不同会导致实际得到的语音信号差异较大，从而引起识别系统的性能下降。

一个语音识别系统如果可以对抗这些环境复杂性，在实际应用场景下依然可以得到较好的识别性能，即可称为一个环境鲁棒的识别系统。为了提高识别系统的鲁棒性，人们提出了各种方法，这些方法可以大体上分为前端信号处理方法和后端模型增强方法。前端信号处理方法通过各种信号处理算法减少噪音、回声和信道对语音信号的影响，使之接近正常安静的语音；后端模型增强

方法是通过对模型及相应解码过程的调整，使之适应实地场景的声学特性。一般来说，前端处理方法计算量小，灵活方便，但性能提升有限；后端模型增强方法计算量较大，需要的数据较多，但性能更好。下面我们将对这两种方法做简单介绍。

2 前端信号处理方法

前端信号处理方法通过对语音信号进行一系列变换，目的是去除信号中各种噪声和失真，恢复清晰语音。不同方法基于不同假设，对不同类型的环境影响产生的效果也不尽相同。总体来说，我们可以将各种环境影响分为加性噪声和卷积噪声，其中背景噪声可以认为是加性噪声，是在原有声音信号上叠加另一种信号，从而产生破坏；混响、回声和信道差异可以认为是一种卷积噪声，是在原有声音信号上的一种附加变换。我们下面介绍的方法对不同类噪声具有不同的效果。

2.1 语音增强方法

语音增强是一种频谱域上的信号处理方法。历史上，语音增强的目的是为了提高语音的可懂度，而不是语音识别性能的提高。仅管如此，在很多情况下，这种方法对语音识别依然有所帮助。

2.1.1 谱减法与加性噪声去除

谱减法（Spectral Substraction，SS）是一种常用的语音增强方法[Boll(1979)]。这一方法假设带噪语音的能量谱是原始语音和噪音能量谱的简单相加，如果可以估计出噪音的能量谱，将该能量谱从带噪语音的能量谱中减去，即可得到原始语音的能量谱。写成公式为：

$$|\hat{X}(f)|^2 = |Y(f)|^2 - |\hat{N}(f)|^2,$$

其中$Y(f)$和$\hat{N}(f)$为带噪语音和噪声的频谱，其中噪声的频谱是估计出来的；$\hat{X}(f)$为利用谱减法估计出的原始语音频谱。事实上，上述能量叠加假设省略了一个原始语音与噪音之间的相关作用项，因此只是一个近似估计。谱减法需要估计噪音信号的频谱，这可以通过确定一些非语音帧（如句子的开始和结束时的短时信号），对这些帧的能量谱进行平均得到。然而，基于这一平均能量得到的噪音估计未必能保证每一帧信号做谱减后都是正数，因此需要在应用时做适

当调整。这些调整有可能会导致相邻频谱间变化过于剧烈，从而引入音乐噪音，需要做进一步平滑处理[Berouti et al(1979)Berouti, Schwartz, and Makhoul]。

2.1.2 回声消除

谱减法面对的是加性噪声。对于回声和混响这种卷积噪声，直接应用谱减法并不合适。为了解决回声和混响问题，传统方法是估计原始语音到接收端的传递函数，一般用房间脉冲响应（RIR）来表示。基于这一传递函数，可以设计一个逆滤波器，从而抵消回声和混响的影响[Gillespie and Atlas(2002), Miyoshi and Kaneda(1988)]。然而，估计RIR本身就是很困难的问题，不精确的RIR估计一般不会带来性能提高。一些研究者发现带混响的语音在做LPC估计时，其残差往往更高斯。基于这一发现，可以直接设计逆滤波器，使生成的语音的LPC残差更加非高斯化[Yegnanarayana and Murthy(2000)]。当混响非常严重时（如在大礼堂中），可以估计出语音信号衰减60dB所需要的时间，称为T_{60}，这一参数可以用来设计一个RIR模型，基于此，延迟较大的混响可以被估计出来，并利用谱减法从带噪信号中减去。另一些研究者利用线性预测模型从历史信号观察中恢复当前原始信号（注意当前观察信号是由历史信号延迟衰减并与当前原始信号叠加而成），将逆滤波器的设计问题转化为线性预测模型的参数估计问题[Nakatani et al(2008)Nakatani, Yoshioka, Kinoshita, Miyoshi, and Juang]。

2.1.3 麦克风阵列

前面所述的各种去噪和去回声方法都是基于单一麦克风。由于现实场景的复杂性，基于单一麦克风很难达到很好的去噪效果。近年来，多麦克风设备开始普及，例如在几乎所有手机上，都装有两个以上的麦克风。多麦克的出现允许我们利用更多空间和时间信息，极大扩展了语音增强能力。最简单的方法是利用一个远端麦克风录制背景噪音，一个近端麦克风录制说话者语音，通过减单谱减法实现语音增强。更通用的场景是利用麦克风阵列技术[Nakatani et al(2008)Nakatani, Yoshioka, Kinoshita, Miyoshi, and Juang, Benesty et al(2008)Benesty, Chen, Kumatani et al(2012)Kumatani, McDonough, and Raj]。

麦克风阵列（Microphone Array）是按一定几何结构组合在一起的若干麦克风。最常用的阵列为线性阵列和环形阵列，如图1所示，其中每个麦克是一般是一个全指向麦克，即对各个方向的敏感度是一致的。和单个麦克风相比，麦克风阵具有极为强大功能，可以实现音源的定位、选择、去噪、去混响等。我们以一个线性麦克风阵列为例来说明。

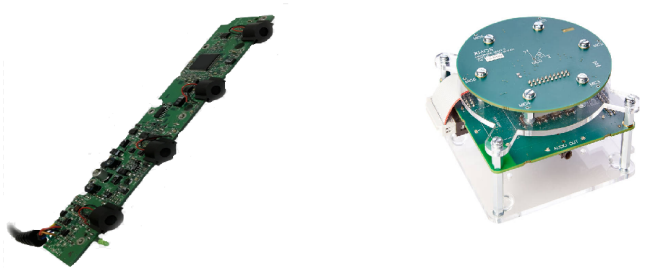

Figure 1: 线性和环形麦克风阵列。

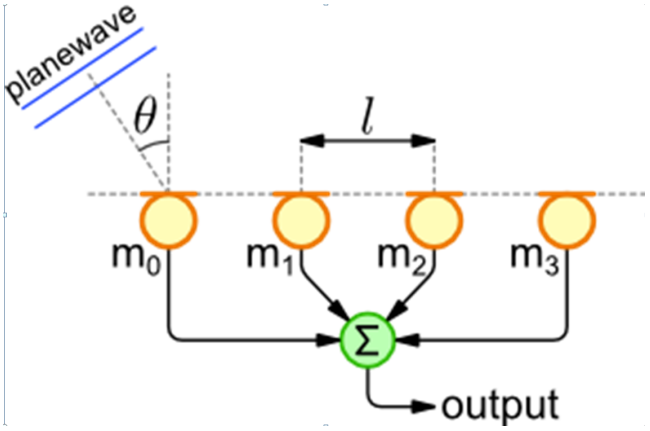

Figure 2: 线性麦克风阵列的声源组合。

如图4所示的四个麦克风组成的阵列，每个麦克风间隔为l，阵列的输出为四个麦克风输出的简单加合。对一个频率为f，由夹角θ入射的平面波，可以计算相临两个麦克风之间的接收到同一信号的延迟为:

$$\Delta t = \frac{lsin(\theta)}{c},$$

其中c为声速。由此可计算相临麦克风之间的相位差为$2\pi f\Delta t$。如果我们将最左边的麦克风接收到的信号计为$Ae^{j2\pi ft}$，则第i个麦克风的信号则为：$Ae^{j2\pi(t+i\Delta t)}$。由此，可计算这四路麦克风输出的结果为：

$$\frac{1}{4}\sum_{i=0}^{3} Ae^{j2\pi f(t+i\Delta t)} = \frac{1}{4}\sum_{i=0}^{3} Ae^{j2\pi f(t+\frac{ilsin(\theta)}{c})}.$$

与单独一个麦克风相比，可知其输出的增益（以dB为单位）为：

$$20\log_{10}\frac{1}{4}\sum_{i=0}^{3} e^{j2\pi f\frac{i\sin(\theta)}{c}}$$

由上式可知，该增益是一个与方向入射角θ的函数，如图3所示。为了更清晰地对比，图中给出了单一麦克风在不同方向上的增益函数。可见，阵列具有明显的方向选择性，即只有在正前方才有较好的增益，其它方向的输入都被抑制。这种方向指向性可以使阵列选择特定的方向，抑制其它方向的噪声输入，从而极大提高信噪比。同时，不同麦克风所接收的噪音是不相关的，这些不相关噪音在互相叠加时会互相抵消，因此可显著降低稳态随机加性噪声的影响。

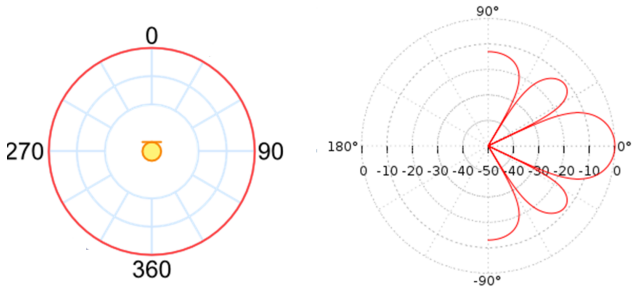

Figure 3: 单一麦克风（左）和线性麦克风阵列（右）的增益指向性。

简单加合的阵列其增益的指向性是固定的。如果我们对每路麦克风的输出做适当延迟，再对延迟后的信号做加合，即可选择阵列的指向性。事实上，如果我们想对入射角为θ的方向做最大增益，只需对由该入射角引起的延迟Δt进入补偿即可，这一方法称为延迟-加和算法（Delay-Sum），如图4所示。延迟-加和算法是最简单也是最常用的阵列算法，研究者提出了各种改进，包括为每个麦克风引入增益参数，调节阵列参数使之更适合语音识别任务等[Seltzer et al(2004)Seltzer, Raj, Stern et al]。

2.2 特征域补偿方法

如前所述，语音增强方法的目的是增加语音的清晰度和可懂度，这一目标与语音识别有一定差距。对语音识别系统来说，更重要的是提取出的特征具有更强的鲁棒性，或环境不变性。因此，在特征域上的补偿或正规化通常可起到更好的效果。我们将讨论三种方法，倒谱特征归一化（CMN和CVN），向量泰勒展开（VTS）和SPLICE。

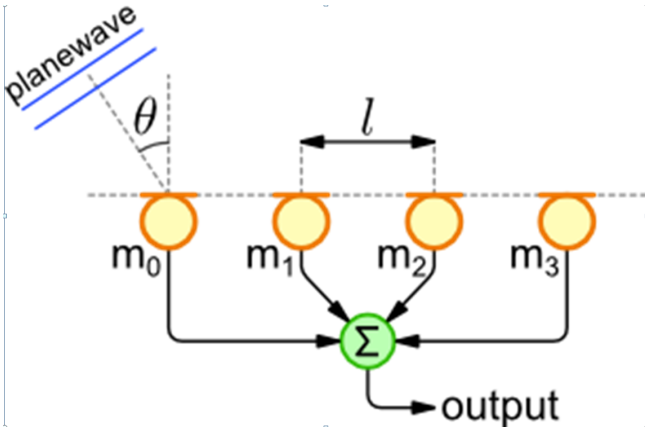

Figure 4: 线性麦克风阵列的延迟-加和算法。对目标语音方向，选择合理的延迟补偿，使得各路麦克风在做完补偿后的相位恰好一致，即可实现该方向的增益最大化。在这一设置下，其它方向的声音因相位失配，导致在输出信号中的增益减小。

2.2.1 CMN & CVN

CMN和CVN是最常用的特征补偿方法，主要用来对卷积噪声进行消除。我们首先讨论CMN。我们已经知道，Fbank和MFCC是最常用的两种特征。这两种特征基于共同的前端处理：加窗、预加重、FFT频谱变换、Mel频域归整、频域能量加窗、log压缩。对一个在时域上卷积的信道噪声，通过上述过程可以实现分解。为清楚看到这一点，设原始信号为$x(t)$，带噪语音信号为$y(t)$，信道的卷积噪声为$h(t)$，则有：

$$y(t) = x(t) * h(t) \quad (1)$$

$$Y(w) = X(w)H(w) \quad (2)$$

$$log|Y(w)|^2 = log|X(w)|^2 + log|H(w)|^2 \quad (3)$$

由此，可得:

$$\mathbf{y} = \mathbf{x} + \mathbf{h}$$

其中$\mathbf{x}$和$\mathbf{y}$分别为原始语音信号和带噪语音信号的Fbank特征，$\mathbf{h}$是和信道相关的卷积噪声。因此，如果我们可以估计出$\mathbf{h}$，即可估计出原始语音信号的Fbank特征$\mathbf{x}$。在实际操作中，可以选择非语音部分信号来估计$\mathbf{h}$；进一步，

如果我们假设$\mathbf{x}$是高斯的，也可以通过整句语音信号取平均来得到，即：

$$\mathbf{h} = \boldsymbol{\mu}_{\mathbf{y}}$$

$$\hat{\mathbf{x}} = \mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}$$

对于MFCC特征，是在Fbank特征基础上加入一个DCT变换，因为该变换是线性的，因此上述分解关系依然成立，即：

$$C\mathbf{y} = C\mathbf{x} + C\mathbf{h}$$

其中C是DCT的变换矩阵。由于MFCC是倒谱系数，上述方法称为倒谱均值正规化（Cepstra Mean Normalization, CMN）[Atal(1974)]。

形式上，CMN可以认为是对特征进行一阶归一化的方法。我们可以设计一种二阶归一化方法，即对方差进行归一化，称为倒谱方差正规化（CVN）。实际应用中，CVN一般和CMN联合使用，称为CMVN，计算公式为：

$$\hat{\mathbf{x}} = \frac{\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}}{\boldsymbol{\sigma}_{\mathbf{y}}}$$

其中除法为按位除。和CMN不同，CVN并没有特别明确的物理背景，但在实际应用中通常会取得一定的性能提高。

CMN和CVN只对一阶和二阶量进行正规化，类似的思路可以扩展到对特征向量的分布进行正规化。一种方法是将特征向量的每一维都正规化到标准高斯分布，称为特征的高斯化[De La Torre et al(2005)De La Torre, Peinado, Segura, Pérez-Córdoba, Benítez, and
高斯化通常采用统计方法，以直方图形式统计特征的实际分布，将其归整为累积概率分布，再将该分布映射到标准高斯分布的累积分布。高斯化对一些任务有一定效果，但在某些任务上的表现未必好于简单的CMN。

在实际系统中，为了保证实时性，需要设计一种在线CMN，在对一句话进行识别时，最初没有任何数据，这时某一缺省的CMN参数来对特征进行归一化，当数据逐渐积累后，对CMN参数进行更好估计，从而逐渐得到更好的归一化特征。这种在线估计可以理解为是一种在高通滤波器（滤掉了固定不变的成份）。将正规化过程表述为一种滤波过程具有很大启发性，一些著名的去噪方法，如ARMA滤波和RASTA滤波[Hermansky and Morgan(1994)]都遵循这一思路。

2.2.2 向量泰勒展开（VTS）

VTS是一种对加性噪声的建模方法。如在谱减法中所述，对于加性噪声，我们假设带噪语音的能量是原始语音和噪声语音的能量之和。假设我们使用的

是Fbank特征，这一关系可表示为：

$$e^{\mathbf{y}} = e^{\mathbf{x}} + e^{\mathbf{n}}$$

做简单变换，有：

$$e^{\mathbf{y}} = e^{\mathbf{x}}(1 + e^{\mathbf{n}-\mathbf{x}})$$

$$\mathbf{y} = \mathbf{x} + \ln(1 + \mathrm{e}^{\mathbf{n}-\mathbf{x}})$$

如果记$\mathbf{r} = \mathbf{n} - \mathbf{x}$，且：

$$g(\mathbf{r}) = \ln(1 + \mathrm{e}^{\mathbf{r}})$$

可得如下关系：

$$\mathbf{y} = \mathbf{x} + g(\mathbf{r})$$

注意，$g(\mathbf{r})$是一个非线性函数。如果对这一非线性函数做一阶展开，即可得到带噪语音、原始语音和噪声之间的简单对应关系，从而由带噪语音推导出原始语音，这一方法称为VTS方法。一般假设原始语音具有混合高斯形式，噪声具有高斯形式。在这一假设下，可通过迭代求出在每一个高斯成分s下，$\mathbf{r}$的期望$\mu_s^{\mathbf{r}}$，并由此得到原始语音的估计$\hat{\mathbf{x}}$如下[J.Droppo and Acero(2007)]：

$$\hat{\mathbf{x}} = \mathbf{y} - \ln(\mathrm{e}^{\mu_{\mathrm{s}}^{\mathbf{r}}} + 1) + \mu_{\mathrm{s}}^{\mathbf{r}}$$

由上述推导过程可知，VTS的基本假设是噪声是加性的，因此语音和噪声之间的能量是相加关系。基于这一基本假设，VTS推导出基于Fbank特征，带噪语音和原始语音之间的关系，并用泰勒展开对这一关系进行近似。值得说明的是，对倒谱特征，如MFCC，上述推导过程依然成立，只不过需要加入一个DCT变换。

2.2.3 SPLICE

SPLICE是另一种对特征进行建模的方法。和VTS不同，SPLICE并不假设噪音的加性，而是直接对原始语音和带噪语音的特征向量建立联合概率分布。为保证建模精确性，SPLICE采用GMM模型：

$$p(\mathbf{y}, \mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|\mathbf{y}, k)p(\mathbf{y}, k)$$

其中$p(\mathbf{y},k)$也是一个GMM:

$$p(\mathbf{y},k)=p(\mathbf{y}|k)p(k).$$

SPLICE定义条件概率$p(\mathbf{x}|\mathbf{y},k)$具有如下线性形式：

$$p(\mathbf{x}|\mathbf{y},k)=N(\mathbf{x};A_k\mathbf{y}+b_k,\Sigma_k),$$

则可由带噪语音估计出原始语音：

$$\hat{\mathbf{x}}=\sum_{k=1}^{K}(A_k\mathbf{y}+b_k)p(k|\mathbf{y}).$$

SPLICE的模型中的$p(\mathbf{y},k)$部分可通过对带噪语音的GMM建模实现，而条件概率$p(\mathbf{x}|\mathbf{y},k)$中的参数$\{A_k,b_k\}$一般需要基于原始语音和相应的带噪语音数据对（Stereo Data）进行训练。

2.3 基于DNN的特征映射

前面所述的大部分方法都假设了一个物理过程，基于该物理过程进行建模。这些方法可信度高，需要的数据和计算量通常较小。然而，这些建模方法都或多或少引入了一些人为假设，这些假设在实际应用中可能偏差较大，带来模型的不精确性。同时，对一些难以建模的场景（如传输过程中信道的即时改变），这些方法也很难凑效。近年来，深度神经网络（DNN）成为语音信号处理的强大工具。DNN的一个显著优势是可以近似任何映射函数，因此可以学习任何复杂的信号传递过程。我们可以利用这一能力，利用DNN将复杂环境中的语音信号或特征映射成安静环境下的信号或特征。研究表明，基于DNN的特征映射方法可取得非常好的效果[Feng et al(2014)Feng, Zhang, and Glass, Han et al(2015)Han, Wang, Wang, Woods, Merks, and Zhang, Wu et al(2017)Wu, Li, Yang, and Lee]。

去噪自编码器（Denoising Auto Encoder, DAE）是一种常见的特征映射模型。和SPLICE一样，我们需要准备一份干净数据和一份相应的带噪数据，将带噪数据输入DAE，输出来近似对应的干净数据。通过训练DAE的参数，即可学习到由带噪语音（或特征）还原出原始语音（或特征）的映射函数。图5给出一个利用DAE去除音乐噪音的例子[Zhao et al(2015)Zhao, Wang, Zhang, and Zhang]，可以看到，DAE可以极大恢复被音乐破坏的语音数据。

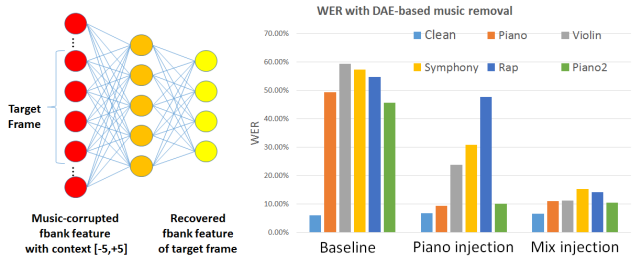

Figure 5: 基于DAE的音乐噪音滤除。左图是DAE的结构，右图是实验结果。其中每一组直方图表示一个系统，每一组里的一个直方图代表测试数据中包括某种音乐的结果。从第一组结果可以看到，训练数据中加入音乐后性能显著下降；第二组结果可以看到，即使只加入一种音乐做DAE训练，也可明显提高系统性能，即便对没见过的音乐也是如此；从第三组结果看到，当加入更多类型的音乐进行训练后，性能有了进一步提高。

3 后端模型增强方法

后端模型方法通过调整声学模型，使得系统适应更复杂的环境。依系统模型的不同，对模型增强的方式也不同。我们主要讨论三种模型增强方法：基于噪声模型的模型增强、模型自适应和带噪训练。

3.1 简单模型增强

对于一个HMM-GMM系统，如果我们假设噪音是加性的，则在2.2.2一节所讨论的对特征的补偿方法可以同样用于对模型参数的补偿。和特征补偿相比，这种在模型上的补偿更加灵活，因此一般性能也更好。

以Fbank特征为例，HMM-GMM系统假设分配到每个高斯成份上的语音帧（Fbank）是高斯分布的，同样，噪声也是高斯分布的。这一高斯分布映射到频域能量谱上，分别记为X和N，这两者也是随机变量，分布为对数高斯（即取对数后是高斯的）。引入加性噪声假设，可知得到的带噪语音的能量谱为$Y = X + N$。一般来说，视X和N的相关性和相对强弱，Y的分布是不规则的。如果我们假设Y依然是对数高斯的，即可求出对应到Fbank域的高斯分布的参数，由此实现对原模型的增强。这一方法称为平行模型加合（Parallel Model Combination）[Gales(1995)]。

另一种方法是将特征域上的VTS补偿应用到模型增强，即不对特征进行修正，而是对模型参数进行改进，以更好描述带噪后的语音。这一方法同样用到

加性噪声假设，但和PMC中的对数高斯近似不同，VTS基于泰勒展开对$\mathbf{y}$和$\mathbf{x}$之间的关系做近似[Moreno et al(1996)Moreno, Raj, and Stern]。

上述两种方法相对简单，但只能处理加性噪声，且只能应用于HMM-GMM系统，现在用的已经较少应用。

3.2 模型自适应

如果我们将实际应用场景看作是与训练不同的另一种场景，则可利用“说话人自适应”一节中所提到的领域自适应方法，利用应用场景的数据对模型进行更新。对HMM-GMM系统，一般采用MAP和MLLR两种方法；对DNN 系统，可在原模型基础上进行再训练，训练时选择较小的步长；或采用知识迁移方法[Tang et al(2016)Tang, Wang, and Zhang]，利用原系统的输出做约束，以减小过拟合的风险。当前对DNN模型最有效的自适应方法还是基于i-vector的条件学习方法。前面提到过，i-vector事实上是一种全信息向量，包括说话人、信道、语言、情绪等多种长时信息，因而可以充分覆盖噪声、混响、编码方式等声学变量。因此，将i-vector作为一种辅助信息引入到DNN模型训练和识别过程中，是一种非常有效的对抗环境影响的方案。

3.3 数据增强训练

DNN的一个显著优势是可以进行多场景进行学习。在传统GMM-HMM系统中，虽然我们可以通过收集更多实际应用场景的数据来提高具体应用场景的性能，但由于模型限制，当收集的数据具有较大差异性时，将导致音素的区分性下降。这意味着大量数据虽然可以提高对场景的覆盖能力，但对某一应用场景来说，并不能达到单一场景建模的效果。DNN极大改变了这种状况。实验表明，DNN模型可以有效学习多场景下的数据，这些各异场景的数据不仅不会降低音素的区分性，反而会互相促进，得到在各种场景下都能普适应用的模型[Yu et al(2013)Yu, Seltzer, Li, Huang, and Seide]。这一结果具有重要意义，说明如果我们可以收集到足够多、对场景覆盖足够全的数据，那么一个DNN系统即可在所有场景下都可以顺利工作。这事实上已经在原则上解决了环境鲁棒性的问题。从某个角度来看，DNN的这种多条件学习能力是今天大规模语音识别系统的基础。

仅管如此，我们依然要考虑如何有效利用DNN的这种多条件学习能力。这是因为数据天然具有长尾效应：绝大部分数据可能是正常的，但对很多特别场景（如特别强的噪音，特别强的混响，很少用的编码方式等），数据通常是不足的。数据在场景上的分布不均衡事实上带来了另一种更深刻的环境鲁棒性问题。

带噪训练（Noisy Training）是解决数据不均衡问题的有效方案，或称为数据增强（Data Augmentation)[Yin et al(2015)Yin, Liu, Zhang, Lin, Wang, Tejedor, Zheng, and Li]。具体来说，带噪训练方法对原始训练数据进行各种变换，以模拟不同场景下的语音信号。这些模拟包括在数据中随机加入不同类型的噪声，让数据通过随机生成的RIR以模拟混响和回声，通过各种编码-解码对以模拟不同信道和编码格式。实验发现，数据增强方法可极大提高系统的鲁棒性，特别是提高小概率场景下的识别性能[Kim et al(2017)Kim, Misra, Chin, Hughes, Narayanan, Sainath, and Bacchiani, Ko et al(2017)Ko, Peddinti, Povey, Seltzer, and Khudanpur]。

4 小结

我们简要介绍了提高识别系统鲁棒性的方法，这些方法可分为前端信号处理方法和后端模型增强方法。前端信号处理方法的目的是对不同环境下的语音或特征进行归一化，使之可以适应标准语音训练出的模型。最常用的前端处理方法是CMN，这种方法简单、高效，且有明确物理意义，被广泛应用于各种商用识别系统。另一种前端处理方法是基于DAE的前端去噪模型。归因于神经网络强大的函数学习能力，DAE可以学习各种环境下的特征映射，完成对复杂声学场景的归一化。后端模型增强方法的思路是对模型进行改进，使之对目标场景有更好效果。模型增强的主要方法有两种，一是对模型进行自适应，使其适应目标场景，二是多场景训练，使模型适应更广泛的场景。对DNN来说，多场景训练一般可取得较好的效果，利用数据增强方法可以进一步提高对非典型场景的覆盖。

References

[Atal(1974)] Atal BS (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America 55(6):1304–1312

[Benesty et al(2008)Benesty, Chen, and Huang] Benesty J, Chen J, Huang Y (2008) Microphone array signal processing, vol 1. Springer Science & Business Media

[Berouti et al(1979)Berouti, Schwartz, and Makhoul] Berouti M, Schwartz R, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise. In:

Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79., IEEE, vol 4, pp 208–211

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

[De La Torre et al(2005)De La Torre, Peinado, Segura, Pérez-Córdoba, Benítez, and Rubio] De La Torre A, Peinado AM, Segura JC, Pérez-Córdoba JL, Benítez MC, Rubio AJ (2005) Histogram equalization of speech representation for robust speech recognition. IEEE Transactions on Speech and Audio Processing 13(3):355–366

[Feng et al(2014)Feng, Zhang, and Glass] Feng X, Zhang Y, Glass J (2014) Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, pp 1759–1763

[Gales(1995)] Gales MJF (1995) Model-based techniques for noise robust speech recognition. PhD thesis, University of Cambridge Cambridge

[Gillespie and Atlas(2002)] Gillespie BW, Atlas LE (2002) Acoustic diversity for improved speech recognition in reverberant environments. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, IEEE, vol 1, pp I–557

[Han et al(2015)Han, Wang, Wang, Woods, Merks, and Zhang] Han K, Wang Y, Wang D, Woods WS, Merks I, Zhang T (2015) Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23(6):982–992

[Hermansky and Morgan(1994)] Hermansky H, Morgan N (1994) Rasta processing of speech. IEEE transactions on speech and audio processing 2(4):578–589

[J.Droppo and Acero(2007)] JDroppo, Acero A (2007) Environmental Robustness, springer

[Kim et al(2017)Kim, Misra, Chin, Hughes, Narayanan, Sainath, and Bacchiani] Kim C, Misra A, Chin K, Hughes T, Narayanan A, Sainath T, Bacchiani M (2017) Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. pp 379–383

[Ko et al(2017)Ko, Peddinti, Povey, Seltzer, and Khudanpur] Ko T, Peddinti V, Povey D, Seltzer ML, Khudanpur S (2017) A study on data augmentation of reverberant speech for robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, pp 5220–5224

[Kumatani et al(2012)Kumatani, McDonough, and Raj] Kumatani K, McDonough J, Raj B (2012) Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE Signal Processing Magazine 29(6):127–140

[Miyoshi and Kaneda(1988)] Miyoshi M, Kaneda Y (1988) Inverse filtering of room acoustics. IEEE Transactions on acoustics, speech, and signal processing 36(2):145–152

[Moreno et al(1996)Moreno, Raj, and Stern] Moreno PJ, Raj B, Stern RM (1996) A vector taylor series approach for environment-independent speech recognition. In: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, IEEE, vol 2, pp 733–736

[Nakatani et al(2008)Nakatani, Yoshioka, Kinoshita, Miyoshi, and Juang] Nakatani T, Yoshioka T, Kinoshita K, Miyoshi M, Juang BH (2008) Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, pp 85–88

[Seltzer et al(2004)Seltzer, Raj, Stern et al] Seltzer ML, Raj B, Stern RM, et al (2004) Likelihood-maximizing beamforming for robust hands-free speech recognition

[Tang et al(2016)Tang, Wang, and Zhang] Tang Z, Wang D, Zhang Z (2016) Recurrent neural network training with dark knowledge transfer. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, pp 5900–5904

[Wu et al(2017)Wu, Li, Yang, and Lee] Wu B, Li K, Yang M, Lee CH (2017) A reverberation-time-aware approach to speech dereverberation based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(1):102–111

[Yegnanarayana and Murthy(2000)] Yegnanarayana B, Murthy PS (2000) Enhancement of reverberant speech using lp residual signal. IEEE Transactions on Speech and Audio Processing 8(3):267–281

[Yin et al(2015)Yin, Liu, Zhang, Lin, Wang, Tejedor, Zheng, and Li] Yin S, Liu C, Zhang Z, Lin Y, Wang D, Tejedor J, Zheng TF, Li Y (2015) Noisy training for deep neural networks in speech recognition. EURASIP Journal on Audio, Speech, and Music Processing 2015(1):2

[Yu et al(2013)Yu, Seltzer, Li, Huang, and Seide] Yu D, Seltzer ML, Li J, Huang JT, Seide F (2013) Feature learning in deep neural networks-studies on speech recognition tasks. arXiv preprint arXiv:13013605

[Zhao et al(2015)Zhao, Wang, Zhang, and Zhang] Zhao M, Wang D, Zhang Z, Zhang X (2015) Music removal by convolutional denoising autoencoder in speech recognition. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific, IEEE, pp 338–341