基于深度神经网络的语音端点检测

殷实[1,2]， 张之勇[2]， 王东[2]， 郑方[2]， 李银国[1]
（1.重庆邮电大学 计算机科学与技术学院，重庆 400065；
2.清华大学 语音和语言技术中心，北京 100084）

摘 要：语音端点检测 (voice activity detection, VAD) 是在连续信号中检测出语音片段的技术，在语音编码、说话人识别、语音识别等领域具有广泛应用。随着移动设备的普及，差异化噪声下的端点检测成为研究的热点与难点。本文提出一种基于深度神经网络 (deep neural network, DNN) 的端点检测方法。这一方法利用DNN在表征复杂模式上的高度灵活性来学习各种语音和噪声模式，实现对语音片段更准确的检测。实验结果表明，基于 DNN 的端点检测方法与基于能量、谱熵、基频等传统检测方法相比具有明显优势，特别是引入带噪训练技术后，该方法在高噪声环境下表现出优异的性能。

关键词：语音识别；端点检测；深度神经网络

Deep neural network based voice activity detection

Yin Shi[1,2], Zhang Zhiyong[2], Wang Dong[2],
Zheng Fang[2], Li Yinguo[1]
**(1. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
2. Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, China)**

Abstract：Voice activity detection (VAD), with the aim of detecting voice segments from continuous speech signals, has been applied in a wide range of applications such as speech coding, speaker recognition and speech recognition. With the popularity of mobile devices, VAD in heterogeneous noise situations has gained much interest. This paper proposes a novel VAD approach based on deep neural networks (DNN). The basic idea is to utilize the flexibility of DNNs in representing complex signal patterns to learn various speech and noise patterns, leading to more precise voice detection. Our experiments show that the DNN-based VAD approach is clearly superior to conventional approaches based on energy, entropy and pitch. Particularly, DNN models trained with noisy data show significant performance improvement in situations with strong noises.

Key words: speech recognition, voice activity detection, deep neural network

从音频流中检测出语音片端，即端点检测技术，对语音编码、说话人分离和识别、语音识别等具有重要意义。一般而言，端点检测定义为从连续音频信号中检测出实际语音片段的起始点和终止点，从而提取出有效的语音片段，排除噪声等其他非语音信号的干扰，为后续语音处理系统提供可靠的语音数据；同时，语音端点检测去除了不必要的非语音片段，减少了后续语音处理系统的计算压力，有利于提高系统的响应速度。

一般来说，在低噪音条件下，端点检测相对容易，传统基于能量或谱熵的检测方法即可得到较高的检测精度。然而，当语音信号受到噪声污染时，端点检测的困难显著提高。特别是随着移动设备的普及，噪声变得更加差异化，检测起来也更为困难。如音乐声、敲门声、背景说话声、咳嗽声等都和待检测的语音信号具有很高的混淆度。在这种差异化复杂噪声环境下，传统的端点检测方法很难取得让人满意的效果[1]。

近年来，DNN 在信号处理领域，特别是语音识别任务上取得了巨大成功，一些研究者也将目光转向了基于 DNN 的语音端点检测。在文[2]中，作者利用 DNN 的学习能力，将多种 VAD 特征进行融合训练 DNN 模型，以此作为语音端点检测的判决模型，取得了很好的效果。该研究的一个不足是各种 VAD 特征需要人为设计，实现起来较为复杂，同时该模型没有提供一个较好的抗噪音方法。

事实上，DNN 具有从原始数据中学习层次特征的能力，可以利用这一能力，在初级特征 (FBank) 上学习逐层学习高层特征，从而避免了人为设计特征的困难。同时，DNN 具有学习各种复杂信号模式的能力，因而可以基于同一模型学习多种差异性噪声特性，从而解决传统 VAD 方法对不同噪声需要分别设计区分性模型的困难。

本文依上述思路，探讨利用 DNN 模型进行端点检测的方法。与文[2]不同的是，本文方法不依赖于人为设计的判决特征（如能量、过零率等），而是从 FBank 特征直接训练 DNN 模型。同时，本文提出利

用带噪训练方法增强 DNN 抗噪性能，进一步增强基于 DNN 的端点检测方法在噪声环境下的鲁棒性。实验结果表明，基于 DNN 的端点检测方法与基于能量、谱熵、基频等传统检测方法相比具有明显优势，特别是引入带噪训练技术后，基于 DNN 的端点检测方法在高噪声环境下表现出优异的性能。

1 传统语音端点检测

传统的端点检测算法主要包括两大类，一类是基于特征提取的端点检测算法，一类是基于模型匹配的端点检测算法。

基于特征提取的端点检测算法从语音信号中提取时域或频域上的特征参数，根据语音/非语音在这些特征参数上的不同分布规律，设定某一阈值或建立区分性模型来区分语音/非语音段。比较有效的时域特征参数包括：短时能量、过零率[3]、自相关函数、基频等。主要的频域特征参数有包括：LPC 倒谱距离、频率方差、谱熵等。本文选择三种常用的特征提取检测法作为对比系统，分别为基于能量的方法，基于谱熵的方法和基于基频的方法。

基于模型的端点检测算法是将语音信号端点检测问题转化成语音帧分类问题，通过建立语音/非语音帧的二分类模型实现语音段起止点检测[4]。 这一方法考虑了相邻语音帧之间的相关性以及误差的先验概率，因此能够比较正确的找到语音/非语音段的分界面。然而，当前绝大数模型方法所采用的模型很难同时学习多种噪声特性，不同噪声往往互相干扰，且很难扩展到集外噪声。本文提出的基于 DNN 的端点检测法即属于模型检测法，同时解决了传统模型方法无法同时学习多种噪声的困难。

2 基于 DNN 的语音端点检测

DNN 是一个包含多个隐藏层的神经网络。神经网络在语音信号处理领域有广泛应用，例如在语音识别中，神经网络常被用来代替传统的高斯混合模型(Gaussian mixture model, GMM)来计算语音帧的状态输出概率。然而，长期以来，神经网络只是作为替代方法存在，并没有表现出对传统方法的绝对优势。直到最近几年，伴随着深度学习技术的兴起和 DNN 的出现，神经网络的优势才得以充分显现，并开始全面取代传统建模方法。

2.1 基于 DNN 的端点检测方法

DNN 模型的一个显著优势是其层次性学习能力。基于其多层网络特性，DNN 在较低层次上学习通用模式，在较高层次上学习复杂模式。这一分层学习方法有利于更充分利用模型参数，同时也更符合人类的学习方式。基于这一特性，可以利用 DNN 从初级特征中学习语音/非语音的高级区分性特征（如能量、谱熵、基频等），而无需人为设计。

同时，DNN 具有学习复杂分类任务的能力。这一方面得益于 DNN 的多层非线性，另一方面得益于其区分性模型的本质。这一特性，使得 DNN 能从大量数据中学习多种噪音模式而互不干扰。

本文提出基于 DNN 的端点检测方法，其基本思路是，利用 DNN 的分层学习能力和区分性建模能力，基于大规模标注的语音库，以音素区分性为学习目标，利用 DNN 从初级 FBank 特征中学习多种语音和非语音模式，实现帧层次上的语音/非语音判决，进而实现适用于差异化复杂噪声环境的端点检测。

具体而言，首先训练一个对音素（实际上是上下文相关音素的特定状态，见第 3 节实验设置）进行分类的 DNN 网络，其输入为某一语音帧的初级 FBank 特征，输出为该语音帧对应的音素。本文使用一个训练好的语音识别系统得到语音帧和音素的对应。 该 DNN 网络可表示为一个由输入到输出的映射函数 $f_\theta : R^M \to R^K$，其中 M 是输入的 FBank 特征向量维度，K 是音素集的大小，θ 表示网络中所有可变参数。设输入 FBank 特征向量为 $\mathbf{x} \in R^M$，对应的目标输出为 $\mathbf{y} \in \{0,1\}^K$，其中 $\mathbf{y}$ 仅在 $\mathbf{x}$ 所对应的音素所在维度取 1，其余维度上取 0。 DNN 的优化目标函数定义为 DNN 输出结果与目标分类的交叉熵：

$$E(\theta) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\{\mathbf{y}^{(n)} ln f_k(\mathbf{x}^{(n)})\} \qquad (1)$$

其中，N 表示训练样本数。依(1)式对该 DNN 模型参数进行优化，即可得到音素区分模型。

对某一帧 Fbank 特征输入，依上述方法训练的 DNN 模型将输出该帧语音在音素集中每一个音素上的后验概率。将所有非噪声/静音音素对应的输出加和，即可得到该帧为语音的概率，通过与某一设定阈值比较，即可判断该帧是否为语音。

2.2 DNN 的加噪训练方法

通过上节所述方法得到的 DNN 模型，在训练条件与测试条件相匹配时，通常可以取得较好的分类效果。然而，当训练条件与测试条件不匹配时，例如训练数据是原始音频信号，而测试数据是含有噪

声的音频信号，则会导致过拟合问题。 这是因为 DNN 模型具有庞大的参数空间，可以学习语音信号中的很多细节，而这些细节在不匹配的测试集中并不存在，因此导致所学模型在测试集上产生偏差。为提高 DNN 模型对噪声的鲁棒性，本文提出带噪训练方法：在训练过程中，人为对训练数据加入不同信道、不同量级的噪声，使得这些噪声能够被 DNN 所学习。如前所述，基于 DNN 模型的区分性模型本质，这些差异性的噪声可以同时被 DNN 学习而互不干扰[5]。

为说明带噪训练的基本原理，假设一种独立同分布的噪声 $\mathbf{v}$，它的一阶矩和二阶矩分别满足：

$$\mathrm{E}\{\mathbf{v}\}=0 \qquad \mathrm{E}\{\mathbf{v}^2\}=\varepsilon I \tag{2}$$

其中，I 是 M 维的单位矩阵，ε 代表一个小的正数。使用泰勒级数展开式(1)中的 $lnf(x)$，则加入噪声之后的误差函数变为：

$$\begin{aligned} E_v(\theta) &= -\sum_{n=1}^{N}\sum_{k=1}^{K}\{\mathbf{y}_k^{(n)} ln f_k(\mathbf{x}^{(n)}+\mathbf{v}^{(n)})\} \\ &\approx -\sum_{n=1}^{N}\sum_{k=1}^{K}\{\mathbf{y}_k^{(n)} ln f_k(\mathbf{x}^{(n)})\} \\ &\quad -\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{y}_k^{(n)}\{\mathbf{v}^{(n)^T}\frac{\nabla f_k(\mathbf{x}^{(n)})}{f_k(\mathbf{x}^{(n)})}+\frac{1}{2}\mathbf{v}^{(n)^T}H_k(\mathbf{x}^{(n)})\mathbf{v}^n\}. \end{aligned} \tag{3}$$

其中，$H_k(x)$ 为：

$$H_k(x)=\frac{-1}{f_k(\mathbf{x})}\nabla f_k(\mathbf{x})\nabla f_k(\mathbf{x})^T+\frac{1}{f_k^2(\mathbf{x})}\nabla^2 f_k(\mathbf{x}).$$

由于 $\mathbf{v}^{(n)}$ 是独立于 $\mathbf{x}^{(n)}$ 的，并且 $\mathrm{E}\{\mathbf{v}\}=0$，则一阶项不存在，误差函数可进一步表示为：

$$E_v(\theta)\approx E(\theta)-\frac{\varepsilon}{2}\sum_{k=1}^{K}tr(\tilde{H}_k). \tag{4}$$

其中，tr 表示矩阵迹操作，$\tilde{H}_k=\sum_{n\in C_k}H_k(\mathbf{x}^{(n)})$，$C_k$ 表示属于第 k 类训练样本集合。

为了进一步了解式(3)的含义，引入一个辅助函数：$E(\theta,\mathbf{v})=-\sum_{n=1}^{N}\sum_{k=1}^{K}\{\mathbf{y}_k^{(n)} ln f_k(\mathbf{x}^{(n)}+\mathbf{v})\}$，其中 $\mathbf{v}$ 是输入向量 $\{\mathbf{x}^{(n)}\}$ 上的一个小小的变化。不同于 $E_v(\theta)$ 中的 $\mathbf{v}$，$E(\theta,\mathbf{v})$ 中的 $\mathbf{v}$ 对每个输入向量 $\mathbf{x}^{(n)}$ 都是一个固定值，而 $E_v(\theta)$ 中的 $\mathbf{v}^{(n)}$ 是一个随机变量，对不同的训练样本值也不同。对 $E(\theta,\mathbf{v})$ 运用拉普拉斯算子，可得：

$$\begin{aligned} \nabla^2 E(\theta,\mathbf{v}) &= tr\{\frac{\partial^2 E(\theta,\mathbf{v})}{\partial v^2}\} \\ &= -tr\{\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{y}_k^{(n)}H_k(\mathbf{x}^{(n)}+\mathbf{v})\} \\ &= -tr\{\sum_{k=1}^{K}\sum_{n\in C_k}H_k(\mathbf{x}^{(n)}+\mathbf{v})\}. \end{aligned} \tag{5}$$

结合式(5)和式(4)可得：

$$E_v(\theta)\approx E(\theta)+\frac{\varepsilon}{2}\nabla^2 E(\theta,0). \tag{6}$$

式(6)表明在输入单元上加入随机噪声等价于在目标函数上增加了一个与目标函数的二阶导有关的正则项。当目标函数趋于优化时，目标函数取最小值，则 $\nabla^2 E(\theta,0)$ 为正数，这意味着正则化的目标函数将更倾向于较平滑位置的最优解。换句话说，依式(6)训练的 DNN 模型对输入的改变较不敏感，过拟合问题得到相应的改善。

3 实验设计

本节首先介绍实验所用的数据库及实验设置，然后报告基于原始语音训练的 DNN 与加噪训练的 DNN 在端点检测上的性能。

3.1 实验数据

本实验采用来自 Tencent 公司提供的 100 小时语音数据作为训练语料，来自 Pachira 公司提供的 600 句语音数据作为测试语料。语音数据的采样率为 16kHz，采样精度为单声道 16bits。

3.2 实验设置

本文使用 Kaldi 工具包[6]进行声学模型建模，绝大部分过程与该工具包中的 WSJ s5.0 流程一致。具体而言，首先依 WSJ s5.0 流程建立一个基于 GMM 模型的语音识别系统，该系统对上下文相关音素进行 HMM-GMM 建模。每个语音音素对应汉语中的一个声母或带调韵母，另有一个静音音素，用来表征所有非语音信号。包含静音音素，音素集中共包括 112 个单音素，经过决策树聚类，最终模型中含有

3611 个概率密度函数(probability density function, PDF)，7400 个共享状态。该识别系统用来对语音数据进行音素对齐，同时为 DNN 模型提供原型，即音素与共享状态的对应。

DNN 模型的输入为基于 Mel 滤波器组的 Fbank 特征，其中每帧语音长度为 25ms， 帧移为 10ms，特征维数为 40。DNN 模型训练和识别时，首先以当前语音帧为中心，前后各取 5 帧组成上下文相关特征向量。这一特征向量经过线性判别式分析(LDA)降维为 200 维向量，再经过全局倒谱归一(global cepstral mean and variance normalization, CMVN)去除信道影响后作为 DNN 的输入。

本文采用的 DNN 的结构如下：输入层含有 200 个输入单元，对应输入的 200 维特征向量(Fbank+LDA+CMVN)；每个隐层含有 1200 个单元，共包含 4 个隐层； 输出层包括 3611 个单元，对应 HMM 系统中的 3611 个概率密度函数(PDF)。训练时采用随机梯度下降(stochastic gradient descendent, SGD)算法, 训练准则为(1)式所示的交叉熵。

3.3 实验结果

实验分为 2 组。在第 1 组实验中，对比在不同 SNR 级别上 DNN 方法与传统方法的性能；在第 2 组试验中，比较使用原始语音和带噪语音数据训练的 DNN 模型在端点检测上的性能差异。

在评价检测结果时，只考虑对语音起始点和结束点的检测，而不考虑中间静音段。评价方法如下：如果对一句话的起始点./结束点的预测值在相应的人工标注点前后 500ms 的误差范围内，则认为该句检测正确，否则认为检测错误。定义检测正确的句子占所有被检测句子的百分比为检测正确率，并作为实验评价的标准。

3.3.1 原始语音训练的 DNN 端点检测

本实验用原始语音训练的 DNN 模型进行端点检测，并与其它三种常用的端点检测方法进行对比，分别为：基于能量的检测方法，基于谱熵的检测方法和基于基频的检测方法。

在检测中，首先对每帧语音进行语音/非语音判决。对于 DNN 模型，将某一语音帧的 DNN 输出进行整理，得到该帧为语音的概率值，如果该值大于某一个预定阈值（本文取 0.2），则判断该帧为语音，反之则认为是非语音。该阈值法同样适用于其它三种端点检测方法，每一方法的阈值依各自的取值范围调节为最优。

得到每帧语音的独立判决之后，通过一个平滑算法得到每一句话的端点判决。实验中选择了一个阈值 T（本文取 17）， 只有当连续语音或非语音帧数大于 T 时，才会触发一个语音开始或结束的事件。这一平滑方法可以避免短时杂音的干扰。为保证比较公平，该平滑方法被应用于本实验中所有端点检测方法中。

表 1 给出了各种检测方法在测试语音上的性能。首先依信噪比将测试语音分为 6 组，统计每组测试的正确率，最后统计在全集上的检测正确率。

表 1　不同端点检测方法对比实验结果

	正确率%						
SNR	$[40,\infty)$	[30,40)	[20,30)	[10,20)	[0,10)	$(-\infty,0)$	$(-\infty,\infty)$
能量	94	83	57	41	16	0	48.5
谱熵	95	82	46	40	15	0	46.33
基频	96	88	81	74	33	0	62
DNN	98	93	87	80	52	11	70.16

从表 1 中的结果可以观察到，DNN 方法比三种传统方法具有明显优势，在各组测试中都取得了最高的正确率。比较不同测试组，可以发现在高信噪比条件下，各种方法性能都较好，DNN 方法的优势并不明显。随着信噪比的降低，传统基于能量以及基于谱熵的方法性能下降严重，很难进行正确的检测；基于基频的方法虽然比前两种方法较为稳定，但是在信噪比低于 0 的条件下，也无法进行有效检测。基于 DNN 的方法在高信噪比的条件下性能比其它方法更好，在低信噪比条件下，性能有所下降，但没有完全失效。这一发现证明 DNN 方法相较于其它传统方法有更好的区分能力，但是对噪声的鲁棒性还有明显不足。

3.3.2 带噪训练的 DNN 端点检测

在第 2 组实验中，使用带噪训练的 DNN 模型进行端点检测。由前文分析可知，带噪训练可以有效提高 DNN 模型的抗噪能力，因而可以提高 DNN 端点检测方法在低信噪比条件下的性能。

表 2 给出了基于原始语料训练的 DNN 和带噪训练的 DNN 在端点检测上的性能对比结果。

表 2　带噪训练的 DNN 与原始语料训练的 DNN 端点检测结果

正确率%

SNR	[40,∞)	[30,40)	[20,30)	[10,20)	[0,10)	(−∞,0)	(−∞,∞)
原始 DNN	98	93	87	80	52	11	70.16
带噪 DNN	98	96	88	83	76	39	80

从表 2 中可以看出，带噪训练的 DNN 在高噪声环境下表现出明显优势，同时并没有降低低噪声环境下的检测性能。这一结果证明，在训练 DNN 模型时有意加入一些噪声语料，会使这些噪声模式被 DNN 所学习，使之对噪音条件下的判决更为准确。同时，加入噪声使 DNN 模型对输入变化的敏感度下降，部分解决了 DNN 训练中的过训练问题。

在实验设置说明中提到，在计算端点检测正确率时，允许在手工标注的端点前后存在 500ms 的误差，这意味着同样正确率的两个模型可能具有不同的检测能力。为了进一步分析不同 DNN 模型在噪声条件下的端点检测性能，随机从测试集中挑选了两个模型都检测正确的 136 条语音，对比两个模型的检测误差，即模型检测结果与手工标注的偏差绝对值与 500ms 的比值。统计结果如图 1 所示。

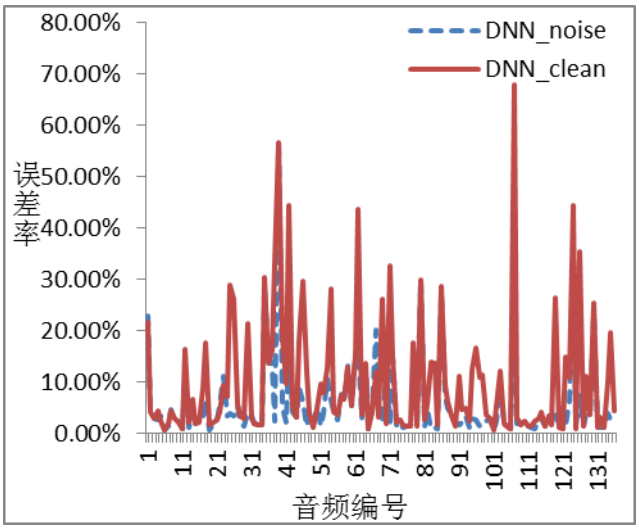

图 1 不同 DNN 模型端点检测误差对比结果

可以发现，带噪训练的 DNN 带来的检测误差要明显低于原始语音训练的 DNN。结合表 1 和表 2 的数据，作者认为基于 DNN 的语音端点检测方法可以在中高信噪比条件下取得让人满意的检测结果，但是在高噪声条件下，由于 DNN 自身的灵活性导致的过拟合问题凸显了出来，导致检测能力下降。通过带噪训练可以明显改善 DNN 对噪声的鲁棒性，使得在低信噪比条件下依然能保持较高的语音端点检测性能。

4 结束语

本文提出一种基于 DNN 的端点检测方法。通过与三种传统检测方法（短时能量、谱熵、基频）在不同信噪比条件下的对比研究，发现基于 DNN 的语音端点检测方法在高噪声环境下具有更好的性能。同时，通过人为加入噪声语料训练的 DNN 模型具有更强的抗噪能力，进一步提高了 DNN 模型在低信噪比条件下的端点检测能力。DNN 的这些特点使得基于 DNN 模型的端点检测方法具有普遍的应用价值。

未来工作将研究各种抗噪特征（如加入基频的 FBank 特征）对 DNN 语音端点检测系统的影响，研究进一步提高端点检测算法对噪声的鲁棒性。

参考文献(References)

[1] Pham C K. Noise Robust Voice Activity Detection[D]. Singapore: Nanyang Technological University, 2012.

[2] Zhang X L, Wu J. Deep Belief Network Based Voice Activity Detection[J]. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2013, 21(4):691-710.

[3] Rabiner L R, Sambur M R. An algorithm for determining the endpoints of isolated utterances[J]. B*ell System Technical Journal*, 1975, 54(2): 297-315.

[4] 朱杰, 韦晓东. 噪声环境中基于 HMM 模型的语音信号端点检测方法[J]. 上海交通大学学报. 1998(10):14-16.

[5] Meng X T, Liu C, Zhang Z Y, and Wang D, Noisy training for deep neural networks[J], *in Proc. of ChinaSIP 2014*, 2014, pp. 16-20.

[6] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]//Proc ASRU. 2011.