# 基于深度神经网络的语音端点检测

# Voice Activity Detection Based on Deep Neural Networks （CSLT-TRP-20150007）

**殷实（Shi Yin）**

**2015/2/10**

**CSLT, RIIT, Tsinghua Univ.**

## 1. 背景简介

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信号；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等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。

近年来，DNN 在信号处理，特别是语音识别领域取得了巨大成功，一些研究者也将目光转向了基于 DNN 的语音端点检测。在某些文章中，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，同时该模型并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 VAD 分类模型，避免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，这可以被利用到在同一模型中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特征的困难。

本文依上述思路，探讨利用 DNN 模型进行端点检测的方法。首先，设计了一种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 模型的方法；同时，本文提出利用带噪训练增强 DNN 抗噪性的方法，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；最后，本文也进行了分别融合帧能量（frame）、teager 能量（teager）、harmonics 能量（harmonic）三种具有较高区分语音/噪音性特征的端点检测实验。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，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；而 frame 能量、teager 能量和 harmonics 能量在纯净语音信号上也能表现出优越的端点检测能力，但在带噪语音信号的端点检测上却仍显不足。

## 2. 语音端点检测技术

语音端点检测本质上是通过语音和噪声对于相同参数所表现出的不同特征来区分两者的，其基本流程如图 1 所示。其中预处理通常包括分帧和预滤波等。分帧是指将语音信号分段（称为语音帧，各帧

通常是有交叠的），预滤波一般是指采用高通滤波器滤除低频噪声；参数提取是指选取可以反映语音和噪声差别的特征参数；端点判决是指采用一种判决准则（如门限判决或模式分类等）来区分语音帧与非语音帧；后处理是指对上述判决结果进行平滑滤波等处理，得到最终的语音端点判决结果。在语音端点检测的流程中，参数提取和端点判决是两个关键步骤。

参数提取是指选取能够反映语音和噪声差别的特征参数，是以语音和噪声的特性为基础。语音信号是一种典型的非平稳信号。但是，语音的形成过程是与发音器官的运动密切相关的，这种物理运动比起声音振动速度要缓慢得多，因此语音信号常常可假定为短时平稳的。语音可粗略分为清音和浊音两大类。浊音在时域上呈现出明显的周期性，在频域上出现共振峰，而且能量大部分集中在较低频段内。但清音段相对于很大一类噪声没有明显的时域和频域特征，类似于白噪声。在语音端点检测算法研究中，可利用浊音的周期性特征，而清音则难以与宽带噪声区分。

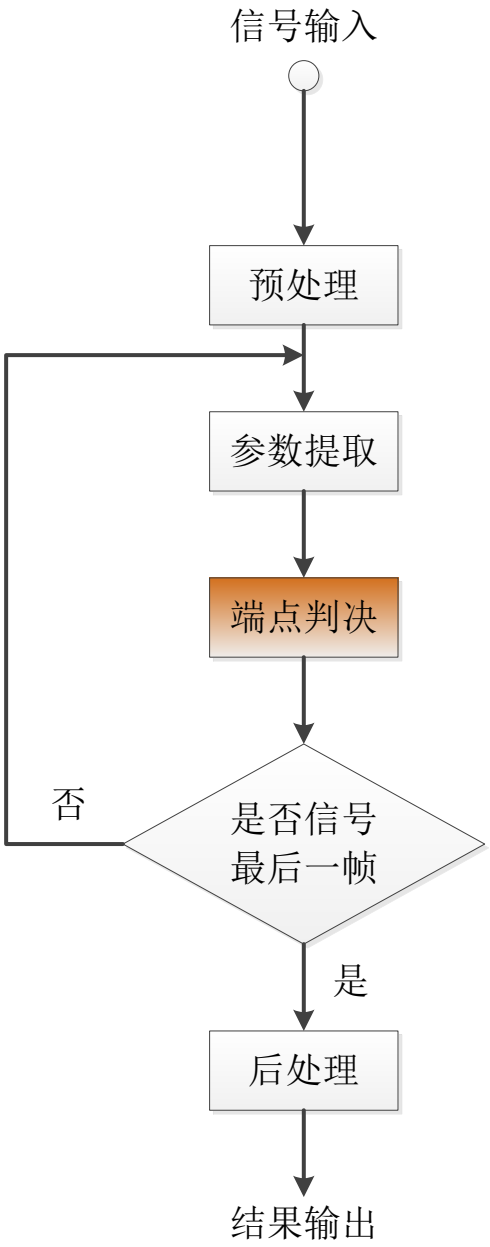


图 1　语音端点检测流程图

由于噪声来源于实际的应用环境，其特性会变化无穷，所以语音端点检测问题是非常复杂的，我们不能通过手工指定规则或使用专家知识轻易的得到判别结果。因此，语音端点检测的重点即找出一种有效的、鲁邦的端点判决方法来区分实际环境中的噪声与语音段。

混入语音中的噪声可以是加性的，也可以是非加性的。考虑到加

性噪声更普遍而且易于分析，并且对于非加性噪声，有些可以通过一定变换转换为加性噪声，因此几乎所有研究语音端点检测技术的学者都是针对加性噪声展开研究的。

## 2.1 传统语音端点检测技术

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、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。本文中，选择了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看成对每帧语音信号进行分类，将端点检测问题转化成分类问题来解决。 基于模型的端点检测算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/非语音的分界面。然而，当前绝大数模型方法所用的模型很难同时学习多种噪声特征，不同噪声往往互相干扰，且无法扩展到其它噪声环境中去。本文提出的基于 DNN 的端点检测法即属于模型法，同时解决了传统模型法无法同时学

习多种噪声的困难；而融合 DNN 方法则是结合了传统模型法和传统特征提取方法的优点。

## 2.2 基于 DNN 的语音端点检测技术

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，也可以用来产生上下文相关的长时语音特征以补充或者代替传统的短时倒谱特征。然而，长时间以来，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。

### 2.2.1 基于 DNN 的语音端点检测

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料库，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）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实现语音帧和音素的对应。 该网络可表示为一个由输入到输出的映射函数为$f_\theta: R^M \rightarrow R^K$，其中$M$是输入的FBank特征向量维度，$K$是音素集的大小，$\theta$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 分类结果与目标输出的交叉熵：

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

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。

对某一帧特征输入，依上述方法训练的 DNN 模型将输出该帧在音素集中每一个音素的后验概率。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。

### 2.2.2 DNN 的加噪训练技术

通过 2.2.1 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，这些各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：

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

其中，$I$是$M$维的单位矩阵，$\varepsilon$代表一个小的正系数。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$\mathbb{E}\{\mathbf{v}\} = 0$，则一阶项不存在，误差函数可进一步表示为：

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

其中，$tr$表示矩阵迹操作，$\widetilde{H}_k = \sum_{n\in C_k} H_k\left(\mathbf{x}^{(n)}\right)$，$C_k$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：

$$
E(\theta, \mathbf{v}) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{\mathbf{y}^{(n)} ln f_k\left(\mathbf{x}^{(n)} + \mathbf{v}\right)\right\}
$$

其中$\mathbf{v}$是输入向量$\left\{\mathbf{x}^{(n)}\right\}$上的一个小小的变化。不同于$E_{\mathrm{v}}(\theta)$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$E_{\mathrm{v}}(\theta)$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：

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

结合式(5)和式(4)可得：

$$E_{\mathbf{v}}(\theta) \approx E(\theta) + \frac{\varepsilon}{2}\nabla^2 E(\theta,0) \quad (6)$$

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。

### 2.3 特征融合 DNN 模型训练技术

为了充分利用 DNN 的分层学习能力和区分性建模能力，可以人为从原始音频文件中提取一些更具噪音/语音区分性的特征，与基本的语音特征 Fbank 结合形成新的特征输入到模型中供 DNN 学习。其基本过程如下：首先，使用一层神经元从同一音频中分别提取不同的特征；然后，其他的层次在一个正则化的框架下被用于特征融合，正则化是为了网络权值的分布更能表征特征之间的关系；最后，使用网络的最后一层来表征融合后的特征，从而构建相应的分类模型。特征融合框图如图 2 所示。

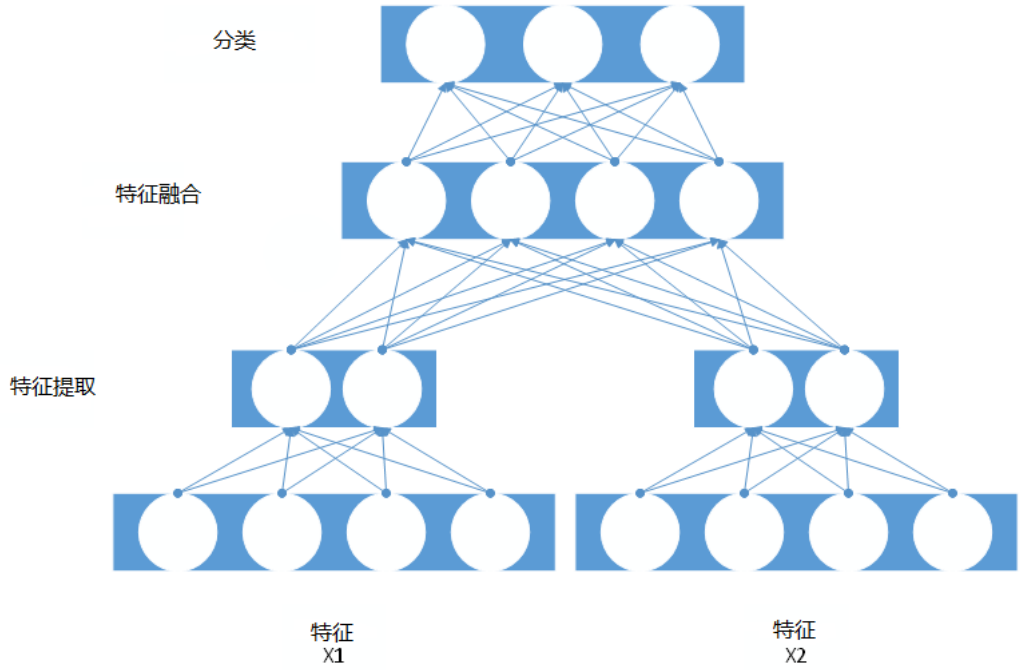


图 2　特征融合框图

### 2.3.1 帧能量融合

语音信号的能量随着时间变化比较明显，一般清音部分的能量比浊音的能量小的多，所以在区分清音和浊音，有声段和无声段的应用中效果比较明显。对于信号$\{x(n)\}$，短时能量的定义如下：

$$E_n = \sum_{m=-\infty}^{+\infty} [x(m)w(n-m)]^2 = \sum_{m=-\infty}^{+\infty} x^2(m)h(n-m) = x^2(n) * h(n)$$

式中$h(n) = w^2(n)$，$E_n$表示在信号的第$n$个点开始加窗函数时的短时能量。通过上式可以看出，短时能量是语音信号的平方经过一个线性低通滤波器的输出，该线性低通滤波器的单位冲激响应应为$h(n)$。

本文中根据相关短时能量的定义提取 1 维短时能量，然后与 40 维基础 Fbank 特征融合，构成新的 41 维特征用于 DNN 网络训练，以期获得对噪声更具区分的模型。

### 2.3.2 Teager 能量融合

Teager 能量算子（Teager Energy Operator，TEO）是一个非线性算子，能够跟踪信号的瞬时能量。美国科学家 H. M. Teager 在研究非线性语音建模时，提出了一种简单的信号分析算法，记作Ψ，设有信号$\mathrm{x}(t)$，则：$\Psi[\mathrm{x}(t)] = \left(\frac{d\mathrm{x}(t)}{dt}\right)^2 - \mathrm{x}(t)\frac{d\mathrm{x}^2(t)}{dt^2}$。对一个作无衰减自由振动的线性振子的振动位移$\mathrm{x}(t) = \mathrm{A}\cos(\omega_c t + \theta)$，有$\Psi[\mathrm{x}(t)] = \Psi(\mathrm{A}\cos(\omega_c t + \theta)) = (\mathrm{A}\omega_c)^2$。又知该振子的瞬时总能量是一个常数，$\mathrm{E} = \mathrm{m}\frac{(\mathrm{A}\omega_c)^2}{2}$，m为振子的质量，这个能与上式的Ψ运算结果只差一个常数因子$\frac{m}{2}$，故将这种Ψ算子称为 Teager 算子。

Teager 能量算子提取包络线是对被测波形相邻的 3 个采样点进行计算，具有优良的时间分辨率，实现起来简单而快速，能实时跟踪被测信号波形变化。

本文中根据相关 Teager 能量的定义，根据当前帧、前帧、后帧，提取 1 维 Teager 能量，然后与 40 维基础 Fbank 特征融合，构成新的 41 维特征用于 DNN 网络训练，以期利用 Teager 能量的实时跟踪信号波形变化的特性得到更具噪声区分性的模型。

### 2.3.3 Harmonics 能量融合

语音信号随时间变化的谱特性可以利用语图仪用图形显示，也称为语谱图。语谱图可以对语音信号进行短时 Fourier 变换得到，分为窄带语谱图和宽带语谱图。前者具有较高的频率分辨率，后者具有较

高的时间分辨率。浊声的窄带语谱图具有一条条有规律的清晰的谐波结构，干净语音的能量大部分集中在这些谐波上。在宽带语谱图上，窄带语谱图的谐波成分的平滑成共振峰，共振峰是谐波成分的包络。大部分噪声不具有共振峰和谐波这些特征。例如咳嗽声、喘气声这些高能量突发噪声不具备这些特征。在噪声背景下，语音的这个特征也非常明显。基于这些现象，本文拟提取共振峰谐波能量参数，共振峰谐波能量参数是指语音帧在窄带语谱图上的各个谐波成分的能量之和。

语音的有规律的谐波成分的时频分布和共振峰是人类发音的一个显著特点。本文通过提取语音信号中的谐波成分，得到的共振峰谐波能量参数作为语音检测的特征。把这个参数作为特征，构建语音和端点检测系统。共振峰谐波能量提取框图如图 3 所示。

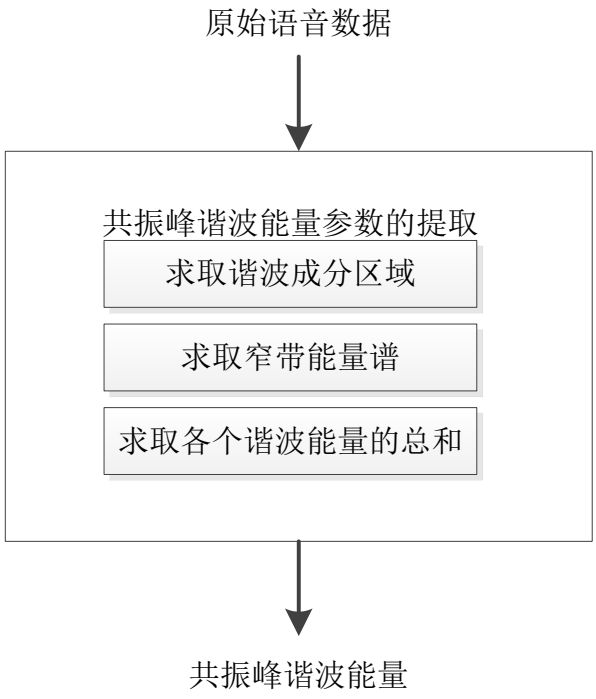


图 3　共振峰谐波能量提取框图

## 3. 基于深度神经网络的语音端点检测技术实现

本文所述基于深度神经网络的语音端点检测均属于基于模型的语音端点检测方法，所有方法的实现都依赖于 DNN 模型，而所有 DNN 模型的训练都在 kaldi 平台上完成，相关测试实验也是基于 kaldi 平台。具体思路即首先在 kaldi 上根据需求训练相应的声学模型，然后修改 kaldi 解码部分的输出，使其专用于语音端点检测，最后，用模型检测出的端点与手工标注的端点进行对比，求取正确率。

### 3.1 基于纯净语料的 DNN 端点检测

在这部分技术实现中，我们使用纯净的语料来训练 DNN 声学模型，然后依赖所得声学模型进行端点检测。该部分所使用的特征就是传统的 40 维 Fbank 特征，所有配置都与传统 DNN 声学模型训练保持一致。具体实现如下：

(1) 传统声学模型详见：

/work0/yinshi/dnn_vad/100_fbank_tec_oldtk

(2) 解码器详见：

/nfs/disk/work/users/yinshi/dnn_vad

(3) 对比程序代码详见：

/nfs/disk/work/users/yinshi/py/tmp_ys.sh。

### 3.2 基于带噪语料的 DNN 端点检测

在这部分技术实现中，为了提高对带噪语音信号的端点检测能力，我们在纯净语料（未人为加噪）的基础上，人为加入了一些带噪语料（包括人声干扰、汽车噪声等）。具体实现如下：

(1) 安装 kaldi；详见：

/nfs/disk/work/users/zhangzy/work/kaldi-201311

(2) 提取纯净语料的特征，详见：

/work0/yinshi/dnn_vad/100_fbank_tec_oldtk/data

(3) 提取噪声语料的特征，并将相应的 text 置为 sil，详见：

/work0/yinshi/dnn_vad/100_fbank_tec_oldtk/data/train.100/train.100_fbank/text

(4) 将上述配置用于训练 DNN 声学模型。

### 3.3 融合帧能量的 DNN 端点检测

在这部分技术实现中，提取了语音信号的短时能量，以此增强 DNN 模型对噪声/语音的区分能力。具体实现如下：

(1) 安装 kaldi 和提取纯净语料特征见 3.2 中(1)(2)

(2) 通过 openSMILE 工具包提取相应的帧能量，详见：

配置文件：

/work0/yinshi/dnn_vad/tools/opensmile/config/demo/demo1_energy.conf

代码：/work0/yinshi/dnn_vad/tools/opensmile/run_frame_energy.py

特征：/work0/yinshi/dnn_vad/tools/opensmile/train_tec_energy_csv

(3) 通过 paste-feats 将纯净语料的 40 维 Fbank 特征与对应的帧能量特征拼接，形成 41 维特征用于 DNN 训练，详见：/work0/yinshi/dnn_vad/100_fbank_tec_oldtk/data/train.100/train.100_fbank/frame_energy

(4) 将上述配置用于训练 DNN 声学模型。

### 3.4 融合 teager 能量的 DNN 端点检测

在这部分技术实现中，提取了语音信号的 teager 能量，具体实现如下：

(1)安装 kaldi 和提取纯净语料特征见 3.3 中(1)(2)

(2)通过 MATLAB 工具包提取相应的 teager 能量，详见：/work0/yinshi/dnn_vad/tools/matlab/Teager.m

(3)同 3.3 节(3)

(4) 同 3.3 节(4)

### 3.5 融合 harmonics 能量的 DNN 端点检测

在这部分技术实现中，提取了语音信号的 formant 能量，以共振峰的特性最大限度的减小噪声对端点检测的干扰，具体实现如下：

(1)安装 kaldi 和提取纯净语料特征见 3.4 中(1)(2)

(2) 通过 openSMILE 工具包提取相应的 formant 能量，详见：

配置文件：

/work0/yinshi/dnn_vad/tools/opensmile/myconfig/formant.conf

代码：

/work0/yinshi/dnn_vad/tools/opensmile/ run_harmonic_energy.py

特征：/work0/yinshi/dnn_vad/tools/opensmile/ train_tec_harmonic_csv

(3)通过 paste-feats 将纯净语料的 40 维 Fbank 特征与对应的 harmonics 能量特征拼接，形成 48 维特征用于 DNN 训练，详见：

/work0/yinshi/dnn_vad/100_fbank_tec_oldtk/data/train.100/train.100_fbank/harmonic_energy

(5) 将上述配置用于训练 DNN 声学模型。

（注：3.3-3.5 节中所使用的解码器与 3.1-3.2 节的不同，由于三种能量特征是通过 openSMILE 工具包提取的，未免麻烦，将 3.1-3.2 节中的解码器由读取 wav 文件改成了读取特征文件，详见：

/nfs/disk/work/users/yinshi/dnn_vad/dnn_vad_frame_energy）

## 4. 后续工作

基于 DNN 模型的语音端点检测还能结合 DAE(Deep Auto Encoder)的方式，进一步加强 DNN 模型对噪声/语音的区分性能。大致思路如下：首先，将带噪语音通过 DAE 映射成降噪之后的语音；然后，通

过纯净语料训练一个纯净的 DNN 模型，通过带噪语料训练一个带噪的 DNN 模型；最后，将映射之后的语音分别通过这两个模型得到端点检测的结果，对比取结果强的即可。

## 5. 总结

本文从理论及实现两方面详述了基于深度神经网络的语音端点检测技术，并且对比分析了 5 种不同的基于 DNN 模型的语音端点检测方法的优劣。最后还简单介绍了一下可以进一步围绕 DNN 语音端点检测开展的工作。现有实验结果也表明，基于 DNN 的语音端点检测方法都要优于传统基于特征提取的端点检测方法，不过基于 DNN 的方法在计算时间、存储空间上较传统方法而言略显不足。