

# 一种基于深度神经网络的基音检测算法

曹 猛, 李鸿燕, 赵蓉蓉

(太原理工大学 信息工程学院, 山西 太原 030024)

**摘 要:** 针对强噪声干扰环境下的基音检测这一挑战性问题, 提出了一种基于深度神经网络的基音检测算法。该算法运用监督学习方法估计基音, 在混合语音数据中得到概率性的基音状态。通过静态帧级声音特征训练, 采用深度神经网络模拟每个基音状态产生的后验概率, 最后执行维特比算法将基音状态连接成基音曲线。实验结果表明: 提出的基音检测算法在不同强噪声甚至混响环境中具有很强的鲁棒性, 而且明显胜过其他的基音检测算法。

**关键词:** 基音检测; 深度神经网络; 监督学习; 维特比算法

**中图分类号:** TP31

**文献标识码:** A

**文章编号:** 1000-7180(2016)06-0143-04

## A Pitch Detection Method Based on Deep Neural Network

CAO Meng, LI Hong-yan, ZHAO Rong-rong

(College of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China)

**Abstract:** Based on deep neural network (DNN), a pitch detection algorithm is proposed to solve the challenging problem of pitch detecting in strong noisy background. In this paper, we estimate pitch using supervised learning, where the probabilistic pitch states are directly learned from noisy speech data. The feedforward deep neural network is trained on static frame-level acoustic features to produce accurate probabilistic outputs of pitch states, which are then connected into pitch contours by Viterbi decoding. The experimental results show that the proposed algorithm are robust to different noise conditions and can even be applied to reverberant speech. The proposed approach outperforms other state-of-art pitch tracking algorithms.

**Key words:** pitch detection; deep neural network; supervised learning; Viterbi decoding

## 1 引言

基音<sup>[1]</sup>或基频, 是语音信号最重要的特征之一。在语音识别和语音分离等很多应用中, 一个基音检测算法鲁棒性高于背景干扰至关重要。近年来对稳健性基音检测的研究已经探索到频域的谐波结构, 时域的周期性, 甚至时频域中独立频带的周期性。PEFAC<sup>[2]</sup>结合非线性振幅压缩减弱窄频带噪声, 并从滤波后的频谱中选择基音候选值。JIN 和 WANG 用互通道相关性选择可靠的通道, 并用隐马尔科夫模型<sup>[3]</sup>提取连续的基音曲线。在强噪声环境或麦克风距离说话者的距离较远的情况下, 语音信号的谐波结构遭受到了严重的破坏, 而使基音估计算法的

性能降低, 导致很多方法在低于 0 dB 条件下变得不适用。因此, 负信噪比条件下的基音跟踪检测仍是一个很大挑战。

2006 年, 加拿大多伦多大学教授 Geoffrey Hinton 提出了深度学习<sup>[4]</sup>理论, 指出多隐层的人工神经网络具有优异的特征学习能力, 学习得到的特征对数据有更本质的刻画, 从而有利于可视化或分类。

基于以上研究, 利用频谱域特征为噪声环境下的基音提取提供稳健性表现, 运用先进的分类器——深度神经网络<sup>[5]</sup>, 产生具有准确概率的基音状态, 增强基音检测的性能。该算法首先提取每帧的频谱域<sup>[6]</sup>特征, 然后采用深度神经网络计算产生基音状态的后验概率, 最后利用维特比解码算法组织

**收稿日期:** 2015-08-27; **修回日期:** 2015-10-19

**基金项目:** 山西省自然科学基金项目(2013011016-1); 教育部博士点基金项目(2011081047)

连续的基音曲线。

## 2 特征提取

对于一个理想周期性声源信号,功率谱密度域中,频率为  $f_0$  时间为  $t$  的信号模型<sup>[7]</sup> 如下式:

$$X_t(f) = \sum_{k=1}^K a_{k,t} \delta(f - kf_0) + N_t(f) \quad (1)$$

式中,  $N_t(f)$  表示噪声的功率谱密度;  $a_{k,t}$  表示第  $k$  个谐波的能量. 对数频域中,信号模型表示成:

$$X_t(q) = \sum_{k=1}^K a_{k,t} \delta(q - \log k - \log f_0) + N_t(q) \quad (2)$$

式中,  $q = \log f$ . 标准化功率谱密度可计算如下:

$$X'_t(q) = X_t(q) \frac{L(q)}{\bar{X}_t(q)} \quad (3)$$

式中,  $L(q)$  代表长时平均语谱;  $\bar{X}_t(q)$  表示平滑平均语谱. 有了标准化的功率谱密度后,利用扩展峰值的滤波器进一步加强谐波性,定义冲击响应如下:

$$h(q) = \begin{cases} \frac{1}{\gamma - \cos(2\pi e^q)} - \beta, & \text{if } \log(0.5) < q < \log(K+0.5) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

式中,  $\beta$  中选, 则  $\int h(q) dq = 0$ ,  $\gamma$  控制峰值宽度, 设定为 1.8.

在此域中,谐波之间的距离不依赖于  $f_0$ , 谐波能量通过卷积  $X_t(q)$  与滤波器冲击响应结合起来:

$$h(q) = \sum_{k=1}^K \delta(q - \log k) \quad (5)$$

卷积  $\hat{X}_t(q) = X'_t(q) * h(q)$  包含的峰值<sup>[8]</sup> 对应于谐波和谐波的倍数和因数,从中选择可信度高的基音频率(60 ~ 400 Hz)的频谱分量作为特征. 由此,得到时间帧  $t$  的频谱特征向量:

$$\hat{X}_t = (\hat{X}_t(q_1), \dots, \hat{X}_t(q_m))^T \quad (6)$$

基频  $F_0$  是存在于  $\hat{X}_t$  的最高峰值. 虽然特征向量的目的是对付噪声语音,但是由于忽略非峰值频谱信息,基于规则的基音候选值的选择可能会丢失有用信息. 本文研究中,把  $\hat{X}_t$  作为提取的特征并运用监督学习方法评估基音的概率,即学习从特征到基音频率的映射,期望监督学习收到更好的效果.

由于相邻时间帧包含基音检测的有用信息,合并相邻时间帧融入特征向量将提高检测准确性. 因此,最终的帧级特征向量是:

$$\mathbf{X}_t = (\hat{X}_{t-d}, \dots, \hat{X}_{t+d})^T, d = 2 \quad (7)$$

## 3 学习基音状态的分配

与选择基音候选值的方法不同,在给定每帧特征向量情况下,运用监督训练的方法学习后验概率的分配. 深度神经网络拥有不止一个隐藏层,具有强大的学习能力,在信号处理应用中卓有成效. 因此,提出通过深度神经网络模拟每个基音状态的后验概率.

首先,将可信的基音频率范围量化,分为  $M$  个频点,对应  $M$  个基音状态,  $s^1, \dots, s^M$ . 在对数尺中,利用每 8 度 24 个频点将可信的基音频率范围(60 ~ 400 Hz)量化,分为 67 点,即第  $m$  频点的量化频率是  $60 \times 2^{(m-1)/24}$  Hz. 此外,合并对应清音的  $s^0$  状态,所以共有 68 个状态.

DNN 训练阶段中,使用带 mini-batch 的随机梯度下降的反向传播训练 DNN 模型. 每个训练样本是时间帧  $t$  (和它的相邻帧) 的特征向量  $\mathbf{X}_t$ , 目标是基音状态  $s_t$  的  $M+1$  维向量,如果地面真值基音落在对应的频率点,那么它的元素  $s_t^i$  是 1,反之为 0.

使用交叉熵  $L(Y, \mathbf{X}; \Theta)$  作为目标函数,学习概率输出如式下:

$$L(Y, \mathbf{X}; \Theta) = - \sum_{m=0}^M y_m \ln f_m(X) \quad (8)$$

式中,  $Y = (y_0, \dots, y_m)^T$  是期望输出;  $f_m(\cdot)$  是输出层第  $m$  个神经元的实际输出.  $\Theta$  表示需要学习的参数. 隐藏层的激活函数是 Sigmoid 函数,输出层使用 Softmax 函数. 文中 DNN 包含三个隐藏层,每层有 1 600 个 Sigmoid 单元. Softmax 输出层的输出单元数目与基音状态的数目相同,即 68 个,由式(8)计算每个 mini-batch 的实际成本. 训练后的 DNN 产生每个基音状态的后验概率为:  $P(s_t^i | \mathbf{X}_t)$ .

## 4 维特比解码算法

经过 DNN 得到每个时间帧的后验概率分布后,使用维特比算法连接这些基音状态. 为产生最佳基音序列<sup>[9]</sup>,维特比算法使用似然性和转移概率计算成本,每帧的似然性  $P(\mathbf{X}_t | s_t^i)$  与后验概率  $P(s_t^i | \mathbf{X}_t)$  和先验概率  $P(s^i)$  的比值成正比:

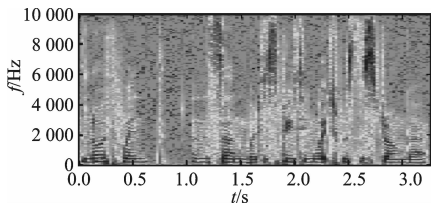
$$P(\mathbf{X}_t | s_t^i) \propto \frac{P(s_t^i | \mathbf{X}_t)}{P(s^i)} \quad (9)$$

式中,  $P(s_t^i | \mathbf{X}_t)$  是神经网络的输出. 计算训练数据可直接得到先验概率  $P(s^i)$  和转移矩阵. 维特比算法输出一系列基音状态,然后将基音状态序列转变为频率序列,最后使用三分平滑移动平均线产生最

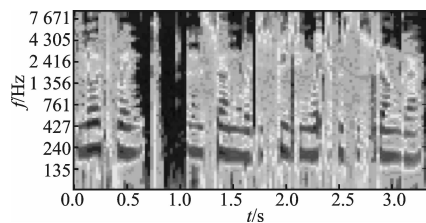
终的基音曲线。

## 5 实验结果

本文采用的原始纯净语音取自 TIMIT 语音库,该语音采样频率是 16 kHz,16 bit 量化.其发音内容为“Note Witt and Sam were disput- ing which was stronger”.图 1 为该语音信号的频谱图和对数频谱图。



(a) 频谱图



(b) 对数频谱图

图 1 原始纯净语音信号

训练集的噪声取自 NOISE-92 语音库,包括白噪声,工厂噪声和低频噪声.训练阶段中,原始纯净语音和三种类型噪声在不同信噪比(-5 dB,0 dB,5 dB)水平下混合.测试集中噪声类型包含三个训练噪声类型和三种新噪声类型:粉红噪声,汽车噪声和背景人噪声.尽管三种训练噪声包含在测试集内,但噪声录音都是不同的剪切片段.测试阶段原始语音在信噪比为-10 dB,-5 dB,0 dB,5 dB,10 dB,20 dB 水平下与噪声混合.该算法检测到混合语音中基音曲线如图 2 所示。

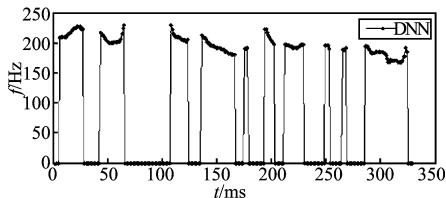


图 2 本文算法基音检测结果

为了验证本方法的有效性和抗噪能力,采用以下两项性能指标进行评估<sup>[10]</sup>。

(1)基音周期检测的正确率 DR:如果估计基音周期和实际真实基音周期的偏差小于 5%,则认为

基音估计正确.该项指标越高,则算法的精确度越好。

$$DR = \frac{N_{0.05}}{N_p} \quad (10)$$

式中, $N_{0.05}$  表示偏差小于 5% 的帧数, $N_p$  表示实际基音帧数。

(2)浊音判断的错误率 VDE:错误判断浊音帧占语音总帧数的百分比.该项指标越低,则清浊音判断的准确性越好。

$$VDE = \frac{N_{p \rightarrow n} + N_{n \rightarrow p}}{N} \quad (11)$$

式中, $N_{p \rightarrow n}$  和  $N_{n \rightarrow p}$  分别表示被错误分类为非浊音和浊音的帧数; $N$  表示语音中总帧数。

本文提出的方法与两种基音检测算法比较. PEFAC 提取标准化光谱特征处理强噪声,然后产生竞争性的基音检测结果. JIN 和 WANG 的多基音检测算法通过计算互通道相关函数选择可靠的通道,然后利用隐马尔科夫模型生成基音曲线.三种基音检测算法基音周期检测正确率的比较,如图 3 所示.可以看出基音周期检测的正确率随信噪比的增加而增加.基于 DNN 的检测方法取得比较高的正确率,尤其是在低信噪比条件下.虽然未在信噪比为-10 dB 情况下训练模型,但提出的算法仍能胜出其他算法.对于未训练的高信噪比条件,提出的算法也能达到良好的性能.由图 3 可直观地看出 DNN 算法基音周期检测的正确率比其他算法平均高 5%,并且当信噪比低于 5 dB 时检测率更是高出 10%。

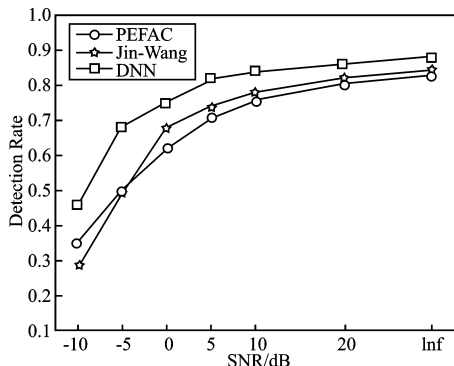


图 3 三种基音检测算法基音周期检测的正确率对比

理想的基音检测算法不仅能获得高基音周期检测的正确率,而且能降低浊音判断的错误率.图 4 表明不同信噪比条件下浊音判断的错误率的结果对比.从图中可以看出,基于 DNN 的基音检测方法相对于其他算法浊音判断的错误率低,在所有噪声和信噪比条件下,DNN 的语音判决错误约为 15%,而

PEFAC 的为 18%，JIN-WANG 算法约为 20%，结果表明提出的方法在基音和浊音检测上有很大的优势。

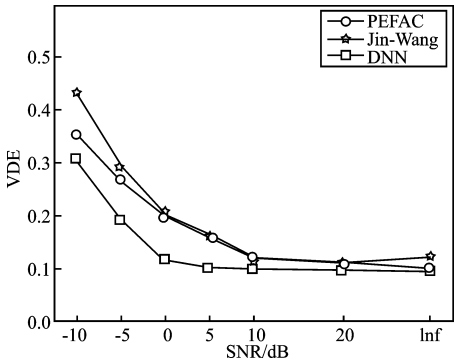


图 4 三种基音检测算法浊音判断的错误率对比

同时,也对三种算法的计算复杂性进行比较. 本实验运行是在 Win7 操作平台上, PC 主频为 2.7 GHz,内存 2 GB. 表 1 表明了不同算法的平均处理时间. 其中,DNN 算法所使用的处理时间有很大的改善.

表 1 不同算法的运行时间对比

|        | PEFAC | JIN-WANG | DNN |
|--------|-------|----------|-----|
| Time/s | 18    | 47       | 11  |

6 结束语

本文提出了一种基于深度神经网络的基音检测算法方法. 该方法通过计算过滤的标准化对数频率功率频谱图提取特征向量,这些特征采用信号处理技术衰减干扰噪声并促进随后神经网络的训练. 在强噪声干扰条件下,训练后的 DNN 评估基音状态的后验概率. 最后利用维特比解码算法组织连续的基音曲线. 用 Matlab 软件进行仿真实验,实验结果表明,该方法的监督式学习无论是在可预见还是不可预见情况下,都可以产生准确的基音检测结果. 特别是在低信噪比条件下,提出的方法在基音和浊音

检测上有很大的优势,提高了系统的鲁棒性.

参考文献:

[1] 胡立波,吴月萍,孙光中. 基于同态滤波的基音检测算法[J]. 微电子学与计算机,2009,26(4): 95-97.

[2] S Gonzalez, M Brookes. A pitch estimation algorithm robust to high levels of noise[J]. Audio Speech Lang. Process., 2014, 22(2): 518-530.

[3] Z Jin, D L Wang. A multipitch tracking algorithm for noisy and reverberant speech[C]// proceedings of the 2010 IEEE international conference on Acoustics, Speech and Signal Processing. Dallas, TX: IEEE, 2010: 4218 - 4221.

[4] Geoffrey Hinton, Ruslan Salakhutdinov. Discovering binary codes for documents by learning deep generative models[J]. Topic in Cognitive Science, 2011, 3(1): 74-91.

[5] 潘伟深,金连文,冯子勇. 基于多尺度梯度及深度神经网络的汉字识别[J]. 北京航空航天大学学报,2015,41(4):751-756.

[6] 由红,陈健. 改进的频域基音检测算法[J]. 上海交通大学学报,2001,35(6):855-858.

[7] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2009:37-43.

[8] 王民,曹绘,要趁红. 一种改进的小波变换基音检测算法[J]. 重庆邮电大学学报(自然科学版),2012,24(3): 283-287.

[9] 王民,孙洁,负卫国,等. 基于两层神经网络的基音检测算法[J]. 计算机工程与应用,2014,50(5):199-202.

[10] 王耀琦,王小鹏. 一种低信噪比环境下的基音检测方法[J]. 铁道学报,2012,34(2):58-62.

作者简介:

曹 猛 男,(1988-),硕士研究生. 研究方向为盲语音分离,语音识别.

李鸿燕 女,(1973-),博士后,副教授. 研究方向为盲信号处理,模式识别. E-mail:tylihy@163. com.