基于卷积神经网络的鲁棒性基音检测方法

张 晖 [1] 苏 红 [1] 张学良 [1] 高光来 [1]

摘 要 在语音信号中, 基音是一个重要参数, 且有重要用途. 然而, 检测噪声环境中语音的基音却是一项难度较大的工作. 由于卷积神经网络 (Convolutional neural network, CNN) 具有平移不变性, 能够很好地刻画语谱图中的谐波结构, 因此我们提出使用 CNN 来完成这项工作. 具体地, 我们使用 CNN 来选取候选基音, 再用动态规划方法 (Dynamic programming, DP) 进行基音追踪, 生成连续的基音轮廓. 实验表明, 与其他方法相比, 本文的方法具有明显的性能优势, 并且对新的说话人和噪声有很好的泛化性能, 具有更好的鲁棒性.

关键词 信号处理, 基音检测, 卷积神经网络, 动态规划

引用格式 张晖, 苏红, 张学良, 高光来. 基于卷积神经网络的鲁棒性基音检测方法. 自动化学报, 2016, **42**(6): 959−964

DOI 10.16383/j.aas.2016.c150672

Convolutional Neural Network for Robust Pitch Determination

ZHANG Hui[1] SU Hong[1] ZHANG Xue-Liang[1]
GAO Guang-Lai[1]

Abstract Pitch is an important characteristic of speech and is useful for many applications. However, pitch determination in noisy conditions is difficult. Because shift-invariant property of convolutional neural network (CNN) is suitable to model spectral feature for pitch detection, we propose a supervised learning algorithm to estimate pitch using CNN. Specifically, we use CNN for pitch candidate selection, and dynamic programming (DP) for pitch tracking. Our experimental results show that the proposed method can obtain accurate pitch estimation and that it has a good generalization ability in terms of new speakers and noisy conditions.

Key words Signal processing, pitch determination, convolutional neural network (CNN), dynamic programming (DP)

Citation Zhang Hui, Su Hong, Zhang Xue-Liang, Gao Guang-Lai. Convolutional neural network for robust pitch determination. *Acta Automatica Sinica*, 2016, **42**(6): 959−964

基音频率, 简称基频, 它决定了语音的音高. 在语音信号处理中, 基频信息可应用于语音识别、语音压缩编码以及语音分离等领域[1−2].

基频估计可以看作一个序列标注问题, 即需要标注出每一帧语音的基频. 我们常用隐马尔科夫模型 (Hidden Markov models, HMMs) 来解决这类问题. HMM 的隐状态 (Hidden state) 对应着基频, 其观察值 (Observation) 对应着输入的语音声学特征, 那么基频估计对应着隐马尔科夫模型的解码问题. 在基频估计中, 有两个关键步骤: 候选基音选取和基音追踪 (Pitch tracking), 分别对应着 HMM 的后验概率计算和解码. 选取候选基音时仅考虑当前帧, 而不考虑语音的连续性. 并把候选基音的可能性得分作为 HMM 的后验概率. 基音追踪根据语音的连续性约束, 将候选基音串联成连续的基音轮廓. 这一解码过程一般采用动态规划 (Dynamic programming, DP) 来完成.

基频估计一般采用信号处理、统计等方法[3−4]. 它们大多利用了语音信号的谐波结构. 但是, 谐波结构较容易受到噪声信号破坏, 使得基频估计错误. 尤其是当信噪比 (Signal-noise ratio, SNR) 较低时, 谐波结构被严重破坏, 基频估计变得尤为困难. 鲁棒性基音检测方法主要关注有噪声干扰的基频估计问题. 如 Wu 等[5] 利用统计方法在破坏较小的通道上为谐波结构建模. PEFAC (Pitch estimation filter with amplitude compression)[6] 方法通过非线性幅度压缩来减小窄带中的噪音成分. Zhang 等[7] 利用语音分离方法去除噪声后再进行基频估计. 这些方法都能缓解噪声干扰带来的损害, 从而提高了基频估计的鲁棒性.

深度学习在其他领域取得了极大的成功[8−9], 受此鼓舞, 一些学者提出用深度模型来选取候选基音. Han 等[10] 首次提出使用深度神经网络 (Deep neural network, DNN) 和递归神经网络 (Recurrent neural network, RNN) 来选取候选基音. 而我们在本研究中首次提出使用卷积神经网络 (Convolutional neural network, CNN) 来完成这一工作.

在本研究中, 我们将卷积神经网络应用于噪声环境下语音的基频估计任务. 实验表明, 与其他方法相比, 我们的方法性能更好, 可以适应说话人和噪声条件的改变, 具有很好的鲁棒性和泛化性能.

本文的结构安排如下: 第 1 节说明使用卷积神经网络的原因, 第 2 节详细介绍提出的基音检测方法, 第 3 节介绍实验过程, 第 4 节给出结论.

1 背景

短时语音信号可以表示为一系列谐波的加权和, 其中第 1 个谐波即为基频, 记做 $F0$, 其他谐波均为 $F0$ 的整数倍. 在语谱图中 (图 1), 这些谐波表现为一条条相互平行的深色曲线. 图中最下面一条曲线即代表基频, 并且每两条相邻曲线之间的距离也是一个基频.

利用谐波结构, 我们有两种获取基频的方法, 一是确定最下面一条曲线的位置, 二是确定相邻两条曲线之间的距离. 显然后一种方法更抗噪、更鲁棒. 为了确定相邻两条曲线之间的距离, 我们需要识别类似图 1 方框中这样的包含两条相邻曲线的局部模式. 我们发现这样的模式在整个语谱图中大量重复出现. 这种重复会同时表现在频率维度和时间维度上. 在本研究中, 我们使用卷积神经网络来挖掘包含在这些重复出现的局部模式中的有用信息.

我们使用卷积神经网络是因为卷积神经网络具有平移不变性 (Shift-invariant). 平移不变性是指一种模式无论它出现在输入的任何位置, 都可以被 CNN 识别出来. 这一特性恰好符合我们要识别出语谱图中大量重复出现的局部模式的需

收稿日期 2015-10-29 录用日期 2016-04-01
Manuscript received October 29, 2015; accepted April 1, 2016
国家自然科学基金 (61365006, 61263037) 资助
Supported by National Natural Science Foundation of China (61365006, 61263037)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 内蒙古大学计算机学院 呼和浩特 010020
1. Computer Science Department, Inner Mongolia University, Hohhot 010020

求, 这是我们选用 CNN 的主要原因. 此外, 相比于 Han 等[10] 所采用的深度神经网络和递归神经网络, 我们使用的 CNN 还有一些额外的优势. 首先, CNN 的识别粒度更小. 在本研究中, 我们识别的是局部模式, 而 Han 等[10] 识别的是整个语音帧, 相比之下, 我们的输入空间更小, 更容易建模刻画. 此外, 谐波结构可能会被噪声破坏, 对应着局部模式的破坏, 但是由于最终的识别结果是从大量局部模式的识别结果中产生的, 少量的局部模式识别错误不会影响到最终的识别结果, 从而能够表现出更好的鲁棒性能. 最后, CNN 采用权值共享, 相对而言训练参数更少, 进而所需的训练数据也会较少. 总之, 使用 CNN 恰好符合基频估计的应用需求, 并且可以带来更好的鲁棒性和泛化性能.

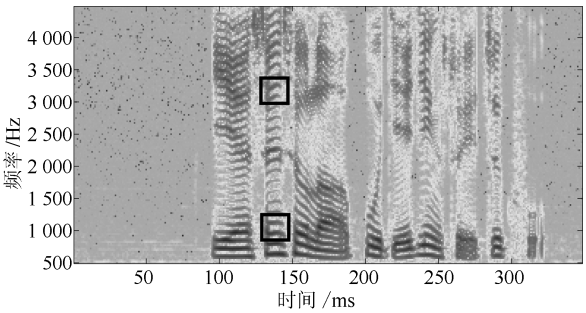


图 1 语谱图中的谐波结构 (小方框中的局部模式重复出现)

Fig. 1 Harmonic structure in spectrogram (The patterns in small windows are repeated. See the ones in the two black boxes.)

2 系统描述

本文提出的基频估计方法分作两个步骤: 候选基音选取和基音追踪. 具体算法如图 2 所示, 主要包括四大模块. 首先, 从输入的语音波形信号中抽取声学特征, 我们选用的声学特征为线性的 PEFAC 特征 (见第 2.1.1 节). 之后, 将语音的声学特征作为一个卷积神经网络的输入, 确定候选基音范围 (见第 2.1.2 节). 然后, 利用混合高斯模型 (Gaussian mixture model, GMM) 计算候选基音的后验概率, 完成候选基音选取 (见第 2.1.3 节). 最后, 使用动态规划进行基音追踪, 输出估计的基频结果 (见第 2.2 节).

2.1 候选基音选取

2.1.1 抽取特征

我们使用 PEFAC[6] 中提出的语谱特征, 该特征已被证明具有较好的鲁棒性. Han 等[10] 在其工作中也使用 PEFAC 特征作为其神经网络的输入. PEFAC 特征是语音的短时傅里叶谱经过一个梳状滤波器滤波后的结果.

需要指出的是, 这里的 PEFAC 特征在频率维度上是对数尺度 (Logarithmic scale) 的. 为了利用卷积神经网络的平移不变性, 我们将 PEFAC 特征从对数尺度变换到线性尺度上. 这是因为在线性尺度上, 语谱图中两条相邻曲线之间的距离相同, 这样的模式重复出现, 适合用卷积神经网络刻画. PEFAC 加强了有基音分布的频段, 同时削弱了没有基音分布的频段. 我们截取线性 PEFAC 特征中基音信息最明显的前 200 维, 将其作为本研究使用的特征, 该特征是卷积神经网络的输入.

2.1.2 确定候选基音范围

我们将确定候选基音视为一个分类任务, 即输入语音当前帧的声学特征, 输出其基频. 本研究中, 将输出基频限定在 80 Hz ~ 415 Hz 之间, 这一范围基本满足了日常会话中基频估计的需求. 然而即使以 1 Hz 作为识别的粒度, 也要求神经网络有超过 300 类的输出, 任务难度较大. 为了简化这个任务, 我们将此范围内连续的基频值用式 (1) 离散化[10] 为若干个基音状态, 每个状态对应真实基频的一个范围. 我们将基音状态作为卷积神经网络的分类目标, 那么卷积神经网络的输出就是候选基音的范围.

$$s = \left\lceil \left(\log_2\left(\frac{p}{60}\right) \cdot 24\right) \right\rceil \tag{1}$$

在式 (1) 中, p 是实际基频, s 是与之相对应的基音状态. 此外还需要考虑无声帧即没有基音的语音帧, 当加入一个无声状态后, 最终得到 59 个分类目标.

我们使用卷积神经网络来完成这个分类任务. CNN 包含多个卷积层和降采样层, 它们相互交替叠合形成一个多层的神经网络. 最后一个降采样层连接到一个多层感知机 (Multi-layer perception, MLP). 本研究中使用的 CNN 结构如图 3 所示. 网络最后一层的传递函数选用 Softmax, CNN 的输出是基音状态概率.

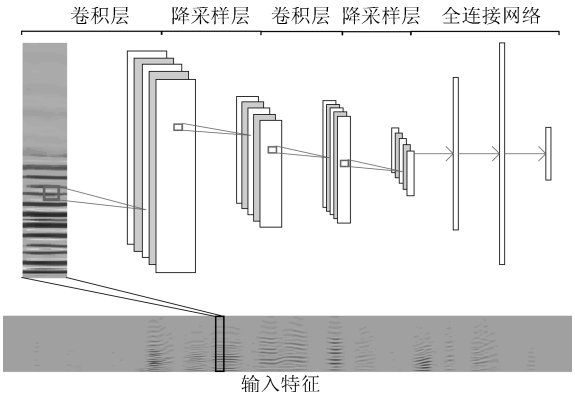


图 3 CNN 的网络结构

Fig. 3 Structure of the proposed CNN

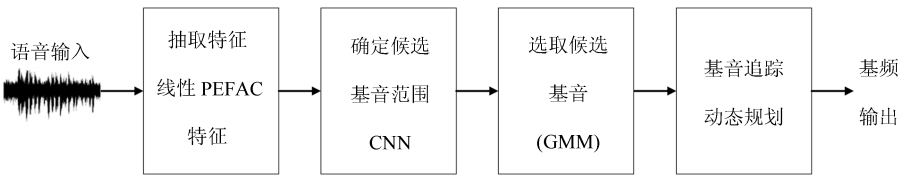


图 2 基频估计算法流程

Fig. 2 The proposed pitch determination algorithm

2.1.3 选取候选基音

前面确定了候选基音的范围, 得到了基音状态的概率, 这里用混合高斯模型将这个离散的基音状态概率转换为连续的基频概率.

对随机变量 z, 混合高斯模型的概率分布函数 $p(z)$ 定义如下:

$$p(z)=\sum_{k=1}^{K}\alpha_k \mathrm{N}(z;\mu_k,\sigma_k^2),\quad \sum_{k=1}^{K}\alpha_k=1,\alpha_k\geq 0 \tag{2}$$

其中, α_k 是系数, $\mathrm{N}(z;\mu_k,\sigma_k^2)$ 代表一个高斯分布, μ_k 和 σ_k^2 分别表示均值和方差. K 是高斯分量的个数.

为了得到候选基音, 我们用高斯分布对每一个基音状态建模, 其标准差 σ_k 是该基音状态带宽的一半, 均值 μ_k 是其中心频率. 我们根据卷积神经网络的输出选取前 K 个基音状态作为候选, 并将卷积神经网络的输出归一化后作为混合高斯模型的系数 α_k. 在本研究中, 根据开发集上的实验结果选择 $K=3$. 此时 $p(z)$ 即表示基频的概率.

这样就完成了候选基音的选取, 并得到了每个候选基音的后验概率.

2.2 基音追踪

基音追踪根据候选基音生成连续的基音轮廓. 由于语音信号的连续性约束, 相邻两帧之间的基频不会发生过大的变化. 根据 Kasi 等的研究[11], 使用 Laplacian 分布来对语音的连续性进行建模. 如式 (3) 所示:

$$p_t(\Delta)=\frac{1}{2\sigma}\exp\left(-\frac{|\Delta-\mu|}{\sigma}\right) \tag{3}$$

其中, Δ 表示相邻帧之间的基频变化量, 为了缩小搜索范围, 我们限制 $|\Delta|\leq 20$. μ 是位置参数, $\sigma>0$ 是尺度参数. 其中 $\mu=0.4$, $\sigma=2.4$[11]. 式 (3) 确定了有声帧之间基频的转移规律. 根据统计, 我们设定有声帧和无声帧之间相互转移的概率都为 0.005. 无声帧之间相互转移的概率为 0.2. 构成概率转移矩阵 (4):

$$\begin{matrix}\left[\begin{array}{c:c}0.2 & 0.005\\ \hdashline 0.005 & p_t(\Delta)\end{array}\right] & \begin{matrix}\}NP\\ \}P\end{matrix}\\ \underbrace{\qquad}_{NP}\ \underbrace{\qquad}_{P} & \end{matrix} \tag{4}$$

其中, NP 表示无声帧, P 表示有声帧.

这样我们就可以使用动态规划完成基音追踪了.

3 实验

3.1 实验数据

为了验证我们提出的方法, 我们用 RASC863 语音数据库中的普通话部分作为纯净语音来源. RASC863 语音数据库包含了来自不同的年龄、性别和教育背景下说话人朗读的大量语音. 在实验中, 噪声选自 Hu 收集的 100 种非语音噪声[12], 分别记做 n1-机器声、n2-鸡尾酒会噪声、n3-工厂噪声、n4-汽笛、n5-语谱噪声 (Speech shaped noise)、n6-白噪声、n7-鸟鸣、n8-鸡啼、n9-人潮噪声、n10-人声嘈杂 (Babble)、n11-马达声、n12-警报、n13-操场噪声、n14-车流噪声、n15-水流声和 n16-风声. 这些噪声覆盖了日常生活中的常见噪声. 为了进一步说明我们的方法对噪声具有更强的鲁棒性, 我们还选择了 IEEE AASP Audio Classification Challenge (ACC)[13] 中的噪声库, 该噪声库包含了 10 种噪声, 记作 n17～n26.

在实验中, 随机地选取一个女性说话人和一个男性说话人, 分别抽取其 50 句语音, 将这 100 句语音与 n1～n6 这 6 种噪声按 0 dB 混合得到的 600 句加噪的语音作为训练集. 在训练集中随机抽取 100 句作为开发集, 抽取到的语句从训练集中移除. 用于测试的数据分为三个集合: 第一个集合的说话人和训练集相同, 再选取新的 20 句语音; 第二个集合和第三个集合随机地选取了 20 个新的说话人, 再从每个说话人中随机地抽取一句语音. 前两个测试集的数据分别与 n1～n16 这 16 种噪声混合, 第三个测试集中的数据与 ACC 噪声库中的 n17～n26 这 10 种噪声混合. 所有测试集都按照 –10 dB、–5 dB、0 dB 和 5 dB 四种不同的信噪比产生加噪语音. n1～n6 是训练时见过的噪声, n7～n26 是新噪声. 由于第一个测试集中的说话人是训练时见过的, 我们称这个测试集为说话人相关测试集, 相对地, 第二个测试集称为说话人不相关测试集. 第三个测试集使用 ACC 噪声, 我们称之为 ACC 测试集.

作为参考目标的真实基频是从纯净语音中使用 PRAAT 软件[14] 提取得到的.

3.2 实验评估

为了验证我们所提出方法的性能, 我们用两个指标来评测实验结果: 基音检测率 (Detection rate, DR) 和错误决策率 (Voicing decision error, VDE). DR 和 VDE 的计算如式 (5) 所示:

$$\mathrm{DR}=\frac{N_{0.05}}{N_p},\quad \mathrm{VDE}=\frac{N_{p\to n}+N_{n\to p}}{N} \tag{5}$$

其中, $N_{0.05}$ 表示估计出的基频和实际基频偏差在 $\pm 5\,\%$ 范围的总帧数, $N_{p\to n}$ 表示将有声帧误判为无声帧的总帧数, $N_{n\to p}$ 表示将无声帧误判为有声帧的总帧数. N_p 和 N 分别表示有声帧的总帧数和所有数据的总帧数. 显然, DR 越大越好, VDE 越小越好.

3.3 实验配置

在实验中, 所有音频数据均降采样到 8 kHz, 并按照 25 ms 帧长, 10 ms 帧移, 分帧处理, 提取线性 PEFAC 特征作为卷积神经网络的输入.

在本研究中, 我们尝试过不同的 CNN 结构, 这些结构的区别在于不同的网络层数、节点数、卷积核大小等, 我们发现除第一个卷积层的卷积核大小对识别效果有较大影响外, 其他因素的影响均较小. 根据在开发集上的实验, 我们使用的 CNN 有 2 个卷积层和 2 个降采样层. 第一个卷积层包含 10 个卷积核, 每个卷积核的大小是 5×5, 实验表明这样的设置可以较好地捕捉到语谱图中的局部模式. 第二个卷积层包含有 20 个 5×5 的卷积核. 降采样层均采用均值降采样 (Mean-pooling), 其大小为 2×2. 之后连接到一个单隐层网络, 其隐层包含 500 个节点. 网络中的传递函数采用 Sigmoid 函数, 最终经过 Softmax 函数输出. 该结构是根据开发集选定的. CNN 训练使用 RMSprop 方法[15] 优化交叉熵目标函数.

3.4 系统分析

我们通过一个例子来展示提出的方法在基频估计上的效果. 图 4 中所用到的语料属于开发集, 是将一句男生语音和机器噪声按照信噪比 0 dB 混合而成的. 图 4 (a) 展示了卷积神经网络的分类效果, 图中点迹为网络输出, 颜色越深表示概率越大, 实线是目标基音状态. 从图中我们可以看出, 深色的点基本都落在了实线上, 这表明卷积神经网络可以较准确地预测基音状态. 在开发集上, CNN 的基音状态分类准确率可达 70 % 以上.

在图 4 (b) 中, 我们展示了基音追踪的效果. 首先图中点线是未经过基音追踪处理的结果, 这里直接将上一步得到的基音状态的中心频率作为输出. 实验表明多数情况下该方法已经可以得到较准确的基频估计, 但会出现离群点, 这是上一步 CNN 在少数帧上的基音状态预测错误所导致的. 图 4 (b) 中实线为基音追踪的结果, 经过处理后, 离群点被移除, 产生了连续的基音轮廓.

在开发集上, 对比基音追踪处理前后的结果, DR 略有提升, VDE 有 1 % 左右的下降. 基音追踪使整个算法的性能略有提升, 但幅度不大. 我们认为有两个方面导致这一结果: 首先, 在基音追踪前系统性能已经达到了较高水平, 提升空间不大. 其次, 在图 4 (a) 中, 几乎在每一帧都仅有一个深色点, 这表明卷积神经网络在预测基音状态时几乎只给出一个大概率的候选基音范围, 这使得基音追踪不能选中此范围以外的基频, 因而效果有限.

总之, 使用卷积神经网络能够较好的选取候选基音, 基于动态规划的基音跟踪能够输出连续的基音轮廓, 产生较好的基音检测结果.

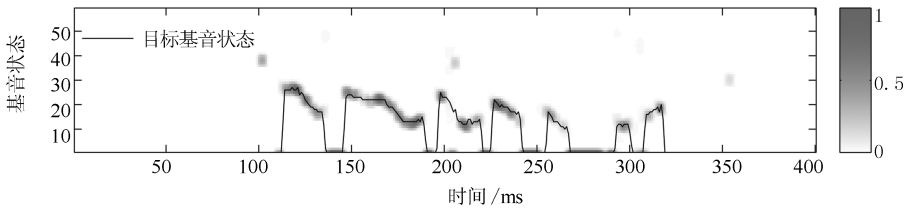
3.5 实验对比

为了评价提出方法的性能, 我们将其与 Jin 方法 (简称 "Jin")[16]、PEFAC 方法 (简称 "PEFAC")[6] 和 DNN 方法 (简称 "DNN")[10] 做对比. 其中 Jin 方法和 PEFAC 方法使用了其作者提供的开源代码, 我们根据文献 [10] 实现了 DNN 方法.

我们在表 1 中列出对比结果, 并在图 5 中给出可视化结果. 从图 5 中, 我们可以直观地发现, 提出的方法 (图中用圆圈表示的曲线) 能够得到较高的基音检测率 (DR) 和较低的错误决策率 (VDE). 表 1 显示与其他方法对比, 提出的方法在各条件下的基音检测率 (DR) 都是最高的, 错误决策率 (VDE) 虽未能保持一致优势但也与最优结果相当. 与 DNN 方法、PEFAC 方法和 Jin 方法相比, 我们提出的方法, DR 平均分别提升了: 5.58 %、5.75 % 和 16.41 %. VDE 则分别下降了 1.91 %、4.25 % 和 10.04 %. 实验表明, 提出的方法比其他方法性能更好. 图 5 中, 从左到右, 测试集与训练集的相似性越来越小. 但是我们所提出方法的优势也越来越明显. 与对比方法中综合性能最好的 DNN 方法相比, 在各测试集上 DR 和 VDE 分别提升 (下降) 了: 4.50 % (0.08 %)、3.68 % (0.06 %)、6.78 % (3.51 %)、4.68 % (0.09 %) 和 8.25 % (5.29 %). 实验表明, 我们提出的方法与其他方法相比有更强的泛化能力.

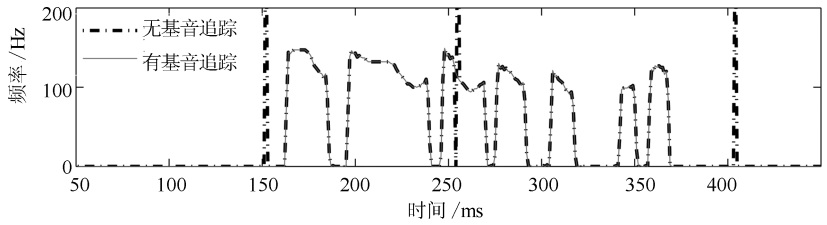
4 结论

在本文中, 我们提出将卷积神经网络应用于噪声环境下的语音基频估计任务中. 卷积神经网络具有平移不变性, 能够很好地刻画谐波结构, 有助于基频估计. 实验表明我们提出的方法明显优于其他方法并且具有更好的鲁棒性, 对新的说话人和新的噪声具有很好的泛化性能.



(a) CNN 输出

(a) CNN output



(b) 基音追踪结果

(b) Pitch tracking output

图 4 基音检测示例 (图中所用语料是一个男声语音和机器噪声按照 0 dB 混合而成的)

Fig. 4 Example output of the proposed pitch determination method (The example mixture is a male utterance which is mixed with machine noise at 0 dB.)

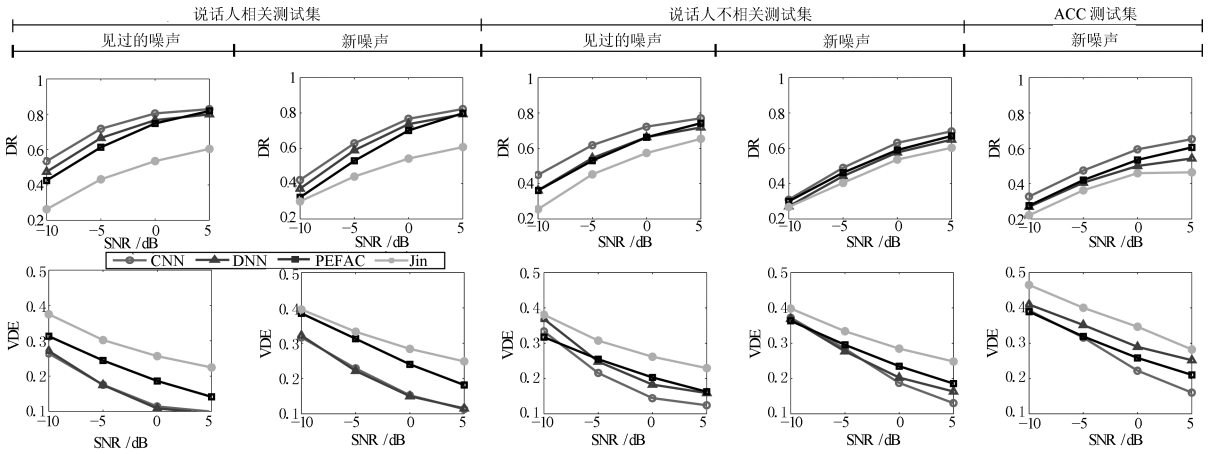


图 5 性能对比图

Fig. 5 Performance comparisons

表 1 本文方法参数设置表

Table 1 Parameters setting of our method

			DR				VDE			
		SNR	−5	0	5	10	−5	0	5	10
说话人相关测试集	见过的噪声	CNN	**0.5342**	**0.7179**	**0.8049**	**0.8292**	**0.2640**	0.1753	0.1140	0.0994
		DNN	0.4747	0.6659	0.7664	0.7994	0.2713	**0.1746**	**0.1083**	**0.0951**
		PEFAC	0.4248	0.6131	0.7478	0.8187	0.3127	0.2443	0.1862	0.1413
		Jin	0.2622	0.4316	0.5350	0.6042	0.3751	0.3021	0.2565	0.2244
	新噪声	CNN	**0.4211**	**0.6278**	**0.7671**	**0.8224**	**0.3166**	0.2287	0.1524	0.1133
		DNN	0.3720	0.5888	0.7369	0.7934	0.3216	**0.2216**	**0.1499**	**0.1154**
		PEFAC	0.3224	0.5291	0.7011	0.7988	0.3844	0.3125	0.2401	0.1815
		Jin	0.2998	0.4403	0.5420	0.6070	0.3954	0.3324	0.2838	0.2484
说话人不相关测试集	见过的噪声	CNN	**0.4495**	**0.6177**	**0.7228**	**0.7699**	0.3334	**0.2156**	**0.1445**	**0.1242**
		DNN	0.3624	0.5449	0.6635	0.7177	0.3685	0.2478	0.1827	0.1590
		PEFAC	0.3611	0.5302	0.6622	0.7421	**0.3172**	0.2546	0.2030	0.1624
		Jin	0.2552	0.4524	0.5731	0.6538	0.3807	0.3074	0.2616	0.2293
	新噪声	CNN	**0.3097**	**0.4899**	**0.6306**	**0.6961**	0.3724	0.284	**0.1875**	**0.1302**
		DNN	0.2714	0.4427	0.5762	0.6489	0.3689	**0.2769**	0.2026	0.1633
		PEFAC	0.2999	0.4619	0.5902	0.6701	**0.3631**	0.2953	0.2348	0.1857
		Jin	0.2680	0.4045	0.5362	0.6030	0.3981	0.3339	0.2845	0.2482
ACC 测试集	新噪声	CNN	**0.3268**	**0.4739**	**0.5938**	**0.6519**	0.3931	**0.316**	**0.2222**	**0.1600**
		DNN	0.2685	0.4053	0.5000	0.5425	0.4096	0.3516	0.2896	0.2519
		PEFAC	0.2751	0.4201	0.5342	0.6051	**0.3893**	0.3190	0.2583	0.2102
		Jin	0.2207	0.3624	0.4592	0.4642	0.4647	0.4002	0.3465	0.2822

References

1 Kun H, Wang D L. A classification based approach to speech segregation. *The Journal of the Acoustical Society of America*, 2012, **132**(5): 3475−3483

2 Zhao X J, Shao Y, Wang D L. CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, & Language Processing*, 2012, **20**(5): 1608−1616

3 Huang F, Lee T. Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique. *IEEE Transactions on Audio, Speech, & Language Processing*, 2013, **21**(1): 99−109

4 Rabiner L. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, 1977, **25**(1): 24−33

5 Wu M Y, Wang D L, Brown G J. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech & Audio Processing*, 2003, **11**(3): 229−241

6 Gonzalez S, Brookes M. PEFAC — a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Transactions on Audio, Speech, & Language Processing*, 2014, **22**(2): 518−530

7 Zhang H, Zhang X, Nie S, Gao G, Liu W. A pairwise algorithm for pitch estimation and speech separation using deep stacking network. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP). South Brisbane, QLD: IEEE, 2015. 246−250

8 Ciresan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2012. 3642−3649

9 Hinton G, Deng L, Yu D, Dahl G E, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, **29**(6): 82−97

10 Han K, Wang D L. Neural network based pitch tracking in very noisy speech. *IEEE/ACM Transactions on Audio, Speech, & Language Processing*, 2014, **22**(12): 2158−2168

11 Kasi K, Zahorian S A. Yet another algorithm for pitch tracking. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, FL, USA: IEEE, 2002. I-361−I-364

12 Hu G N. 100 nonspeech sounds [Online], available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, April 1, 2006.

13 Giannoulis D, Benetos E, Stowell D, Rossignol M, Lagrange M, Plumbley M D. Detection and classification of acoustic scenes and events: an IEEE AASP challenge. In: Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY: IEEE, 2013. 1−4

14 Boersma P, Weenink D J M. PRAAT, a system for doing phonetics by computer. *Glot International*, 2001, **5**(9−10): 341−345

15 Tieleman T, Hinton G. Lecture 6.5 — RMSprop. *COURSERA: Neural Networks for Machine Learning*, 2012.

16 Jin Z Z, Wang D L. Hmm-based multipitch tracking for noisy and reverberant speech. *IEEE Transactions on Audio, Speech, & Language Processing*, 2011, **19**(5): 1091−1102

张 晖 内蒙古大学博士研究生. 分别于 2011 年和 2014 年获得内蒙古大学学士和硕士学位. 主要研究方向为语音信号处理, 语音分离和机器学习. E-mail: alzhu.san@163.com
(**ZHANG Hui** Ph. D. candidate at Inner Mongolia University. He received his B. S. and M. S. degrees from Inner Mongolia University in 2011 and 2014, respectively. His research interest covers audio signal processing, speech separation, and machine learning algorithms.)

苏 红 内蒙古大学硕士研究生. 2013 年获得内蒙古师范大学学士学位. 主要研究方向为语音信号处理和机器学习.
E-mail: sh123imu@163.com
(**SU Hong** Master student at Inner Mongolia University. She received her B. S. degree from Inner Mongolia Normal University in 2013. Her research interest covers audio signal processing and machine learning.)

张学良 内蒙古大学计算机学院副教授. 2003 年获得内蒙古大学学士学位, 2005 年获得哈尔滨工业大学硕士学位, 2010 年获得中国科学院自动化研究所博士学位. 主要研究方向为语音分离, 听觉场景分析和语音信号处理. 本文通信作者. E-mail: cszxl@imu.edu.cn
(**ZHANG Xue-Liang** Associate professor in the Department of Computer Science, Inner Mongolia University. He received his B. S. degree from the Inner Mongolia University in 2003, the M. S. degree from Harbin Institute of Technology in 2005, and the Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2010. His research interest covers speech separation, computational auditory scene analysis, and speech signal processing. Corresponding author of this paper.)

高光来 内蒙古大学计算机学院教授. 1985 年获得内蒙古大学学士学位, 1988 年获得国防科技大学硕士学位. 主要研究方向为人工智能与模式识别. E-mail: csggl@imu.edu.cn
(**GAO Guang-Lai** Professor in the Department of Computer Science, Inner Mongolia University. He received his B. S. degree from Inner Mongolia University in 1985, and received his M. S. degree from the National University of Defense Technology in 1988. His research interest covers artificial intelligence and pattern recognition.)