文章编号:1671-4598(2025)02-0246-08 DOI:10.16526/j.cnki.11-4762/tp.2025.02.031 中图分类号:TP242 文献标识码:A

基于改进 MFCC 特征提取和 DNN 网络的机器人语音识别方法研究

秦垲忻[1]，王炜昕[2]，王砚生[3]
(1. 昆明市普通话培训测试中心，昆明 650000；
2. 昆明市网络安全应急指挥中心，昆明 650500；
3. 云南师范大学 能源与环境科学学院，昆明 650500)

摘要：为了实现机器人语音控制，并避免环境噪音的干扰，研究提出了基于改进 MFCC 特征提取和深度神经网络的机器人语音控制指令识别方法；该方法利用线性判别分析、最大似然线性变换和说话人自适应变换对 MFCC 特征进行处理，获得了新的语音特征；同时通过深度玻尔兹曼机对声学模型进行了改进，并利用深度神经网络和谐波增强技术构建了语音增强方法；实验结果显示，研究提出的基于改进 Mel 频率倒谱系数特征能显著降低语音识别的字错误率，通过辅以改进深度神经网络－隐马尔科夫模型能进一步降低字错误率；在 20 dB 条件下，该特征和改进深度神经网络－隐马尔科夫模型的平均字错误率分别为 24.9%和 22.1%，均低于其他方法；上述结果表明，研究提出的语音识别方法能实现带噪声语音的准确识别，提高机器人的语音控制指令识别能力。

关键词：语音识别；语音增强；声学模型；MFCC 特征；DNN

Research on Robot Speech Recognition Method Based on Improved MFCC Feature Extraction and DNN Network

QIN Kaixin[1], WANG Weixin[2], WANG Yansheng[3]
(1. Kunming Putonghua Training and Testing Center, Kunming 650000, China;
2. Kunming Network Security Emergency Command Center, Kunming 650500, China;
3. School of Energy and Environment Science, Yunnan Normal University, Kunming 650500, China)

Abstract: To achieve robot voice control and avoid environmental noise interference, a robot voice control instruction recognition method based on improved Mel frequency cepstrum coefficient (MFCC) feature extraction and deep neural network (DNN) is proposed. This method utilizes linear discriminant analysis, maximum likelihood linear transformation, and speaker adaptive transformation to process MFCC features and obtain new speech features. Meanwhile, the acoustic model is improved using the deep Boltzmann machine, and deep neural networks and harmonic enhancement techniques are used to construct the speech enhancement method. Experimental results show that the proposed feature based on improved Mel frequency cepstral coefficients can significantly reduce the error rate in speech recognition, and further reduce the error rate by using an improved deep neural network hidden Markov model. Under the condition of 20 dB, the average error rates of this feature and improved deep neural network hidden Markov model are 24.9% and 22.1%, respectively, which are both lower than those of other methods. The above results indicate that the proposed speech recognition method can achieve accurate recognition of noisy speech and improve the ability of robot speech control command recognition.

Keywords: speech recognition; speech enhancement; acoustic model; MFCC feature; DNN

收稿日期:2024-08-29； **修回日期**:2024-10-12。
基金项目:2022 年云南省哲学社会科学规划项目(YB2022085);2024 年全国教育规划青年课题(EHA210438)。
作者简介:秦垲忻(1984-),男,大学本科,工程师。
引用格式:秦垲忻,王炜昕,王砚生.基于改进 MFCC 特征提取和 DNN 网络的机器人语音识别方法研究[J].计算机测量与控制,2025,33(2):246-253.

0　引言

随着计算机技术的发展，语音交流已由人人交流扩展到了人机交流。通过人机之间的语音交流不仅能实现语音与文字之间的转换，还能通过语音直接控制各类设备。例如当前大多数的智能家居设备均能通过语音实现直接控制。而对于当前各类机器人而言，其大多也是通过语音进行控制的。但由于大多数机器人基本均处于公共环境之中，背景噪声较大，导致机器人难以准确理解语音指令的含义[1-2]。因此，在构建机器人的语音控制系统时，如何提高噪声环境中的语音识别性能是其面临的主要问题。当前国外针对语音识别的研究主要集中于深度学习、端到端语音识别及增强方面。例如文献［3］中为了实现阿马齐格语的语音识别，提出了基于 CNN 的语音识别方法。该方法利用 CNN 和 Mel 频谱图来评估音频样本并生成频谱图。实验结果表明，阿马齐格语中的口语数字能够以 93.62％的最高准确率、94％的准确率和 94％的召回率进行识别[3]。而国内针对语音识别的研究则主要涉及端到端语音识别、语音增强及多语种识别。例如文献［4］中提出了基于深度全序列卷积神经网络和联结时序分类的语音识别模型，该模型解决了语言模型的误差梯度无法传递给声学模型的问题。实验结果显示，该模型降低了 21％的语音识别字错误率[4]。但上述方法的实验环境大多为实验室，其噪声干扰较小。而机器人一般处于嘈杂环境之中，导致当前的语音识别方法难以准确识别指令含义。针对嘈杂环境中的语音识别问题，文献［5］中提出了基于混合特征提取技术的语音识别方法。该方法通过连接感知线性预测和 Mel 频率倒谱系数的核心块来提高语音识别器的性能。实验结果显示，该方法在嘈杂环境中的语音识别性能平均提高了 12.88％，但值得注意的是该方法存在帧丢失的问题。针对嘈杂环境中语音识别不准，进而导致翻译精度降低的问题，文献［6］中提出了一种基于语音合成的翻译方法，该方法通过数据预加重、分帧加窗和端点检测来提高嘈杂环境中语音特征提取的准确性。实验结果显示，该方法的识别准确率高达 96.78％。但由于文献［6］中的语音识别方法是以语音合成为基础的，导致其识别时间较长，难以满足机器人语音指令识别的需求。语音增强技术能从含噪语音中提取尽可能纯净的原始语音，进而避免环境噪声的干扰。此外，在进行语音识别时，特征提取和声学模型则是决定语音识别性能的关键因素。鉴于此，为了实现机器人的语音控制，降低环境噪音干扰，研究提出了基于 Mel 频率倒谱系数（MFCC，Mel frequency cepstrum coefficient）特征提取和深度神经网络（DNN，deep neural network）的语音增强及识别方法。研究的创新点在于利用线性判别分析（LDA，linear discriminant analysis）、最大似然线性变换（MLLT，maximum likelihood linear transformation）和说话人自适应变换（SAT，speaker adaptive transformation）对 MFCC 特征进行处理，获得了新的语音特征。对于声学模型而言，研究利用深度玻尔兹曼机（DBM，deep boltzmann machine）对深度神经网络－隐马尔科夫模型（DNN-HMM，deep neural network- hidden Markov model）进行了改进，同时利用 DNN 和谐波增强（HE，harmonic enhancement）技术构建了语音增强方法。

1　MFCC 特征提取及变换

1.1　MFCC 语音特征提取

在利用语音对机器人进行控制时，能否对语音进行准确识别是决定控制效果的关键因素。而对于语音识别而言，特征提取和声学模型匹配是其关键步骤。在语音特征提取中，MFCC 是基于人耳听觉特性提出来的，其与 Hz 频率成非线性对应关系，能更好地反映声音特征。人耳对实际声音频率的感知[7-8]见式（1）：

$$F_{\text{mel}} = 2595\log\left(1+\frac{f}{700}\right) \tag{1}$$

式中，F_{mel}为人耳的感知频率；f 为实际的声音频率。对于人耳而言，其耳蜗实质上相当于一个滤波器组。耳蜗的滤波作用是在对数频率尺度上进行的，在 1 kHz 以下为线性尺度，1 kHz 以上为对数尺度，使得人耳对低频信号敏感，而高频信号不敏感[9-11]。根据上述原则，可以设计出 Mel 滤波器组。Mel 滤波器组的频率响应曲线如图 1 所示。

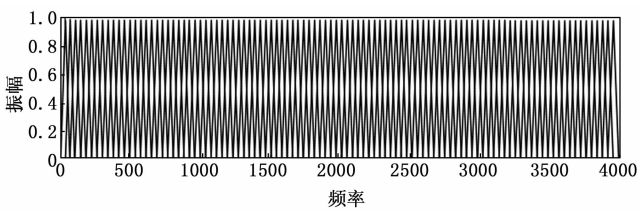

图 1　Mel 滤波器组的频率响应曲线

由图 1 可知，Mel 滤波器组就是一系列的三角形滤波器，在中心频率点响应值为 1，在两边的滤波器中心点衰减到 0。因此，Mel 滤波器组的传递函数见式（2）：

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{f(m-1)-k}{f(m+1)-f(m)}, & f(m)\leqslant k\leqslant f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{2}$$

式中，$H_m(k)$ 为 Mel 滤波器组的传递函数，$f(m)$ 为第 m 个滤波器的中心频率；k 为快速傅里叶变换后的编号。滤波器的中心频率计算公式见式（3）：

$$f(m)=\left(\frac{N}{F_s}\right)B^{-1}\left(F_{\text{mel}}(f_l)+m\frac{F_{\text{mel}}(f_h)-F_{\text{mel}}(f_l)}{M+1}\right) \tag{3}$$

式中，N 为快速傅里叶变换的长度；F_s 为采样频率；B^{-1} 为人耳频率感知公式的逆变换；f_l 为滤波器的带宽下限；F_h 为滤波器的带宽上限；M 为滤波器的数量。在对 MFCC 参数进行提取时，其首先需要对声音率进行预加重、分帧和加窗。接着即可对频率的时域信号进行变换，以获取相应的频域信号。时域信号变换公式见式（4）：

$$X(i,k)=FFT[x_i(m)],\quad 0\leqslant m\leqslant 320 \tag{4}$$

式中，$X(i,k)$ 为频域信号；FFT 为快速傅里叶变换；$x_i(m)$ 为语音信号。语音信号的谱线能量计算公式见式（5）：

$$E_i(k)=X^2(i,k) \tag{5}$$

式中，$E_i(k)$ 为语音信号的谱线能量。经 Mel 滤波器处理后，谱线能量的输出公式见式（6）：

$$S(i,m)=\sum_{k=0}^{N-1}E_i(k)H_m(k),\quad 0\leqslant m<M \tag{6}$$

式中，$S(i,m)$ 为 Mel 滤波器的输出。将 Mel 滤波器的输出进行离散余弦变换后，即可得到 MFCC 特征，其计算公式见式（7）：

$$MFCC(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M}\log_{10}[S(i,m)]\cos\left[\frac{\pi n(2m-1)}{2M}\right] \tag{7}$$

式中，$MFCC(i,n)$ 为 MFCC 特征；n 为特征维数。将得到的 MFCC 特征进行离散傅里叶变换后，即可得到频率的 Delta 特征，其计算公式见式（8）：

$$D_i=\frac{\sum_{\theta=1}^{\Theta}(MFCC_{i+\theta}-MFCC_i)}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{8}$$

式中，D_i 为 Delta 特征；Θ 为 Delta 特征的维数；$MFCC_{i+\theta}$ 为第 i 帧频率的 θ 维 MFCC 特征。

1.2 MFCC 特征变换

通过上述方法虽然能对声音频率的特征进行提取，但由于特征向量的维数与冗余信息成正比，导致所提取到的声音频率特征可能存在大量无关信息。为此，研究利用 LDA 变换对其进行降维处理。LDA 是一种监督学习的降维技术，其通过给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点中心尽可能远离。降维时，首先需要对样本均值进行计算，其计算公式见式（9）：

$$\begin{cases}\mu_j=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n_j}\boldsymbol{x}_i^j\\ \mu=\dfrac{1}{d}\displaystyle\sum_{i=1}^{d}\boldsymbol{x}_i\end{cases} \tag{9}$$

式中，μ_j 为第 j 类样本的均值；n_j 为第 j 类样本的数量；$\boldsymbol{x}_i^j$ 为第 j 类样本的第 i 个向量；μ 为总样本的均值；d 为样本向量的数量；$\boldsymbol{x}_i$ 为第 i 个样本向量。此时，样本的类间和类内散度矩阵见式（10）：

$$\begin{cases}\boldsymbol{S}_b=\displaystyle\sum_{i=1}^{c}(\mu_i-\mu)(\mu_i-\mu)^T\\ \boldsymbol{S}_w=\displaystyle\sum_{i=1}^{c}\sum_{i=1}^{n_j}(x_i^j-\mu_j)(x_i^j-\mu_j)^T\end{cases} \tag{10}$$

式中，$\boldsymbol{S}_b$ 为类间散度矩阵；c 为类别数；$\boldsymbol{S}_w$ 为类内散度矩阵。通过上述类间和类内散度矩阵可以对转换矩阵进行计算。此时，线性判别分析的特征变换见式（11）：

$$Y=\boldsymbol{W}_L^T\boldsymbol{X} \tag{11}$$

式中，$\boldsymbol{W}_L$ 为转换矩阵；$\boldsymbol{X}$ 为样本向量集合。为了在最大似然准则下使用线性变换矩阵对参数特征矢量进行解相关，研究对特征参数进行了最大似然线性变换。似然度计算公式见式（12）：

$$P=(2\pi)^{-Nd/2}\exp\left\{-\frac{1}{2}N[\bar{u}-u]^T\boldsymbol{\varepsilon}^{-1}[\bar{u}-u]+Tr(\boldsymbol{\varepsilon}^{-1}\bar{\boldsymbol{\varepsilon}})+\log|\varepsilon|\right\} \tag{12}$$

式中，P 为似然度；$\bar{u}$ 和 u 分别为原始样本均值和最大似然准则下的样本均值；$\boldsymbol{\varepsilon}$ 和 $\bar{\boldsymbol{\varepsilon}}$ 分别为原始样本的协方差矩阵和 ML 的协方差矩阵；Tr 为 MLLT。此时，即可根据线性变换后所得的特征参数空间构建模型，并计算似然度。但由于不同特征参数间的似然度无法直接进行比较，因此需要将其转换回原始特征空间中[12-13]。考虑到上述方法计算复杂，因此为简化计算，令线性变换矩阵为 1。此时，协方差矩阵对角化前后，模型与训练数据间的似然度计算公式见式（13）：

$$\begin{cases}\boldsymbol{P}^*(x_1^N)=(2\pi e)^{-Nd/2}|\boldsymbol{A}\bar{\boldsymbol{\varepsilon}}\boldsymbol{A}^T|^{-N/2}=|\boldsymbol{A}|^{-N}\boldsymbol{P}^*(x_1^N)\\ \boldsymbol{P}_{\text{diag}}^*(x_1^N)=(2\pi e)^{-Nd/2}|\boldsymbol{A}\bar{\boldsymbol{\varepsilon}}\boldsymbol{A}^T|=(2\pi e)^{-Nd/2}|\operatorname{diag}(\boldsymbol{\varepsilon})|^{-N/2}\end{cases} \tag{13}$$

式中，$\boldsymbol{P}^*(x_1^N)$ 为模型的似然度矩阵；$\boldsymbol{A}$ 为线性矩阵；$\boldsymbol{P}_{\text{diag}}^*(x_1^N)$ 为协方差矩阵对角化后的模型似然度；diag 为对角化。对比式（13）可知，经协方差矩阵对角化后，模型的似然度有所降低。样本协方差矩阵对角化后的似然度计算公式见式（14）：

$$\boldsymbol{P}_{\text{diag}}^*(y_1^N)=(2\pi e)^{-Nd/2}|\operatorname{diag}(\bar{\boldsymbol{\varepsilon}})|^{-N/2} \tag{14}$$

式中，$\boldsymbol{P}_{\text{diag}}^*(y_1^N)$ 为样本协方差矩阵对角化后的似然

度。对比可知，其大于对角化后的模型似然度。可见，在协方差矩阵对角化后，样本与模型间的似然度增加。此外，在对语音进行识别前，为去除个人码本对识别结果的影响，研究还利用 SAT 对说话人无关（SI，speaker independent）码本进行线性变换，以得到说话人自适应（SA，speaker adapted）码本，进而达到提升模型识别性能的目的，其变化公式见式（15）：

$$(\overline{\boldsymbol{W}},\bar{\lambda}) = \arg\max\prod_{k=1}^{K}\prod_{t\in\Omega(k)}P[o(t)\mid\lambda,\boldsymbol{W}_k] = \arg\min\Big(-\sum_{k=1}^{K}\sum_{t\in\Omega(k)}\log[P(o(t)\mid\lambda,\boldsymbol{W}_k)]\Big) \tag{15}$$

式中，$\overline{\boldsymbol{W}}$ 为最佳估计矩阵；$\bar{\lambda}$ 为参数集合的均值；$P(o(t)\mid\lambda,\boldsymbol{W}_k)$ 为输出概率函数；$o(t)$ 为输出；λ 为 SI 码本均值和方差的集合；$\boldsymbol{W}_k$ 为第 k 个说话人的变换矩阵。

2　模型构建

2.1　改进 DNN-HMM

通过上述方法虽然可以实现对语音的特征提取，但若想实现对语音指令的准确识别，还需要借助声学模型。目前的主流的声学模型为高斯混合模型－隐马尔科夫模型（GMM-HMM，Gaussian mixture model hidden Markov model），虽然具有不错的识别性能，但随着数据量的增加，模型的性能会逐渐降低。同时，当前的声学模型还存在非线性流型数据空间建模能力差的缺点[14-16]。而随着基于深度学习的语音识别技术的逐渐成熟，基于 DNN-HMM 的声学模型已开始被广泛应用。虽然相较于 GMM-HMM 模型，DNN-HMM 不需要对声学特征所服从的分布进行假设，且可以采用连续的拼接帧作为输入，有效提高了上下文信息的利用效率。但 DNN-HMM 模型对于复杂语音的建模能力较差，难以实现对复杂语音指令的准确识别。鉴于此，研究利用 DBM 对 DNN-HMM 进行了改进。改进 DNN-HMM 与传统 DNN-HMM 的结构类似，均包含 4 个隐藏层和输入层、输出层，不同的是改进 DNN-HMM 的输入层与隐藏层之间是无向图全连接 DBM 模型。DBM 与深度置信网虽然都是由受限玻尔兹曼机（RBM，restricted Boltzmann machine）堆叠而成，但不同的是 DBM 连接层的连接均是无向的。RBM 公式见式（16）：

$$E(v,h) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}h_iv_j - \sum_{j=1}^{m}b_jv_j - \sum_{i=1}^{n}c_ih_i \tag{16}$$

式中，$E(v,h)$ 为可视节点与隐藏节点间的能量函数；ω_{ij} 为权重；n 和 m 分别为隐藏节点和可视节点的数量；h_i 和 v_j 分别为隐藏节点和可视节点；b_j 和 c_i 分别为可视节点和隐藏节点的偏置。此时，可视节点和隐藏节点的联合概率见式（17）：

$$p(v\mid h) = \frac{e^{-E(h,v)}}{\sum_{v,h}e^{-E(h,v)}} \tag{17}$$

式中，$p(v\mid h)$ 为可视节点和隐藏节点的联合概率。根据统计学理论，低能量的发生概率最大，因此通过自由能量函数可将联合概率最大化，此时联合概率可表示为式（18）：

$$\ln p(v) = -FreeEnergy(v) - \ln Z \tag{18}$$

式中，$p(v)$ 为似然函数；$FreeEnergy(v)$ 为网络的自由能力和；Z 为归一化因子。值得注意的是，对于改进 DNN-HMM 模型而言，其训练一般通过误差反向传播进行，即利用误差反向传播算法来计算模型的参数梯度，进而实现网络参数的更新。现将代价函数关于神经元输入的偏导数作为误差项，其计算公式见式（19）：

$$\delta^{(l)} = \frac{\partial J(W,b;x,y)}{\partial z^{(l)}} = \frac{\partial J(W,b;x,y)}{\partial z^{(l+1)}}\cdot\frac{\partial z^{(l+1)}}{\partial a^{(l+1)}}\cdot\frac{\partial a^{(l+1)}}{\partial z^{(l)}} = f_l^{z^{(l)}}\odot[(W^{(l)})^T\delta^{(l+1)}] \tag{19}$$

式中，$\delta^{(l)}$ 为第 l 层神经元的误差项；$\boldsymbol{W}$ 为权重矩阵；b 为偏置；x 和 y 分别为输入和输出；$z^{(l)}$ 为第 l 层的输入；$a^{(l)}$ 为第 l 层神经元的激活输出；$f_l[z^{(l)}]$ 为第 l 层的激活函数。在得到误差项后，对每层参数的偏导数进行计算，即可实现参数的更新。误差反向传播公式见式（20）：

$$\delta^{(l)} = f_i^{\prime z^{(l)}}\odot[(W^{(l)})^T\delta^{(l+1)}] \tag{20}$$

由于 DNN-HMM 的训练是通过反向传播进行的，因此其损失函数一般为交叉熵函数。交叉熵函数见式（21）：

$$F_{\mathrm{CE}} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u}\log y_{ut}(s_{ut}) \tag{21}$$

式中，F_{CE} 为预测状态标签和参考状态标签间的交叉熵函数；y_{ut} 为状态概率输出；s_{ut} 为 t 时刻的状态。此时，输出层的梯度函数见式（22）：

$$\begin{cases}\dfrac{\partial F_{\mathrm{CE}}}{\partial \boldsymbol{a}_{ut}(s)} = -\dfrac{\partial \log y_{ut}(s_{ut})}{\partial \boldsymbol{a}_{ut}(s)} = y_{ut}(s) - (\delta_s, s_{ut})\\ (\delta_s, s_{ut}) = \begin{cases}1, & s = s_{ut}\\ 0, & s \neq s_{ut}\end{cases}\end{cases} \tag{22}$$

式中，$\boldsymbol{a}_{ut}(s)$ 为输入向量；δ_s 为克罗内克函数。通过上述方法即可构建出基于改进 DNN-HMM 的声学模型，对于该模型，其参数则利用误差反向传播法进行调整。

2.2　基于 DNN 及谐波增强的语音增强模型

通过上述特征提取及识别模型，虽然能对语音指令进行准确识别，但受环境噪声的影响，可能导致语音指令的可懂度降低[17-19]。鉴于此，为了避免环境噪声对语音指令的干扰，研究提出了基于 DNN 的语音增强方法。在进行语音增强前，首先需要对噪声的自回归模型

参数进行估计。基于 DNN 的自回归模型参数预测如图 2 所示。

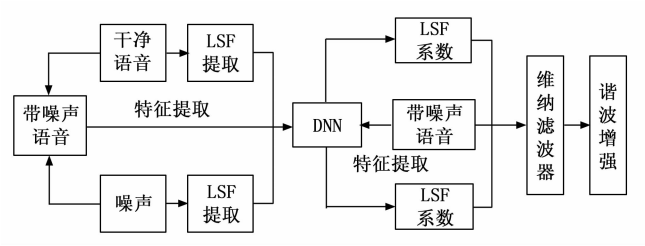

图 2 基于 DNN 的自回归模型参数预测

由图 2 可知，首先将待噪声语音分为纯净语音和噪声，并分别对噪声及纯净语音的线谱频率（LSF，line spectrum frequency）参数进行提取，并将两种 LSF 参数进行融合，以作为 DNN 的训练目标。随后利用 Adam 算法对 DNN 进行优化，以获取 LPS 至 LSF 的映射函数。利用 Adam 算法优化 DNN 时，首先选取一个 batch 的样本，并计算 DNN 的权值梯度。接着计算梯度的有偏一阶矩和有偏二阶矩估计，计算公式见式（23）：

$$\begin{cases} s \leftarrow \rho_1 s + (1-\rho_1)\boldsymbol{g} \\ r \leftarrow \rho_2 r + (1-\rho_2)\boldsymbol{g}\cdot\boldsymbol{g} \end{cases} \tag{23}$$

式中，s 为梯度的有偏一阶矩估计；ρ_1 和 ρ_2 均为平滑系数；$\boldsymbol{g}$ 为梯度向量；r 为梯度的有偏二阶矩。在得到有偏一阶矩和有偏二阶矩估计后，对其偏差进行修正即可计算 DNN 参数的更新。DNN 参数更新公式见式（24）：

$$\Delta\theta = -\varepsilon\frac{\hat{s}}{\sqrt{\hat{r}}+\delta} \tag{24}$$

式中，θ 为 DNN 的参数；ε 为步长；δ 为用于数值稳定的常数。最后，即可利用 LSF 参数构建维纳滤波器，并通过谐波增强技术实现语音增强。DNN 的代价函数见式（25）：

$$J(w,b) = \frac{1}{M_b}\sum_{i=1}^{M_b}\left[d^{(i)} - h_{w,b}(x^{(i)})\right]^2 \tag{25}$$

式中，$J(w,b)$ 为 DNN 的代价函数；w 为 DNN 的权值；b 为 DNN 的偏置；M_b 为批量的大小；$d^{(i)}$ 为 DNN 的期望输出；$h_{w,b}$ 为 DNN 的映射函数；$x^{(i)}$ 为 DNN 的输入。在训练好 DNN 后，即可利用所得的权值进行线上测试。在进行线上测试时，首先需要对 LPS 特征进行归一化，以作为 DNN 的输入特征；而其输出则为语音及噪声的 LSF 系数，其计算公式见式（26）：

$$\begin{cases} \hat{p}_j^x = h_{w,b}^j(x^{(i)}), & 0 \leqslant j \leqslant 9 \\ \hat{p}_j^w = h_{w,b}^j(x^{(i)}), & 10 \leqslant j \leqslant 19 \end{cases} \tag{26}$$

式中，$\hat{p}_j^x$ 为输出中的语音 LSF 系数；$h_{w,b}^j$ 为第 j 维的 DNN 映射函数；$\hat{p}_j^w$ 为输出中的噪声 LSF 系数。此时，维纳滤波器计算公式见式（27）：

$$WF(k) = \frac{\dfrac{\hat{g}_x}{|A_x(k)|^2}}{\left[\dfrac{\hat{g}_x}{|A_x(k)|^2} + \dfrac{\hat{g}_w}{|A_w(k)|^2}\right]} \tag{27}$$

式中，WF 为维纳滤波器；$\hat{g}_x$ 为语音的自回归增益；$\dfrac{1}{|A_x(k)|^2}$ 为语音的自回归谱的形状；$\hat{g}_w$ 为噪声的自回归增益；$\dfrac{1}{|A_w(k)|^2}$ 为噪声的自回归谱的形状。在通过上述方法对自回归参数进行估计后，即可利用谐波增强算法去除噪声，以实现语音增强。在进行语音增强时，首先需要选出峰值的对应频率，并确认该频率是否为谐波频率，其判断准则见式（28）：

$$20\log_{10}[A(w_a)] - 20\log_{10}(max\{A(w_i)\}) > 8\ \text{dB} \tag{28}$$

式中，A 为幅值；w_a 为峰值的对应频率；w_i 为旁瓣频率。若某一频率满足式（28），则表明该频率为谐波频率。峰值频率与基音频率的判别关系见式（29）：

$$\frac{|w_a - lw_o|}{lw_0} < 0.1 \tag{29}$$

式中，w_0 为基音频率；l 为距峰值频率与基音频率之比最近的整数。通过上述方法仅能对谐波频率进行大致判断，若想确认该频率为谐波频率，还需要对谱平坦度进行判别，其计算公式见式（30）：

$$min\left\{\left[\left(10\log_{10}\frac{n\sqrt[n]{x_1x_2\cdots x_i}}{x_1+x_2+\cdots+x_i}\right)/50\right],1\right\} > 0.1,$$
$$x_i \in \left(w_c + \frac{w_0}{2}, w_c + 3\frac{w_0}{2}\right) \tag{30}$$

式中，x_i 为第 i 个频率。对于功率谱而言，其越不平坦，则越接近语音谱，反之则越接近白噪声。若某一频率的功率谱平坦度满足上式，则确认该频率为谐波频率。此时，将该谐波频率赋值于上一区间的谐波频率，并对下一频率进行判别，直至达到频谱边界[20-21]。通过上述方法，即可找到频谱中的所有谐波频率，并以此构建梳状滤波器。梳状滤波器的计算公式见式（31）：

$$H_l(w_k) = \begin{cases} WF(w_k)\cdot e^{-2(w_k-w_a)/\sigma}, & w_k \in \left(w_a - \dfrac{w_0}{2}, w_a + \dfrac{w_0}{2}\right) \\ WF(w_k), & \text{otherwise} \end{cases} \tag{31}$$

式中，$H_l(w_k)$ 为梳状滤波器；w_k 为频谱中的频率；σ 为标准差。通过谐波增强算法能有效去除谐波间的噪声，同时辅以逆傅里叶变换即可实现对语音信号的增强。融合语音增强后的语音识别模型如图 3 所示。

由图 3 可知，首先语音指令进行预处理，并对其进行语音增强。然后对增强后的语音的特征参数进行提取。

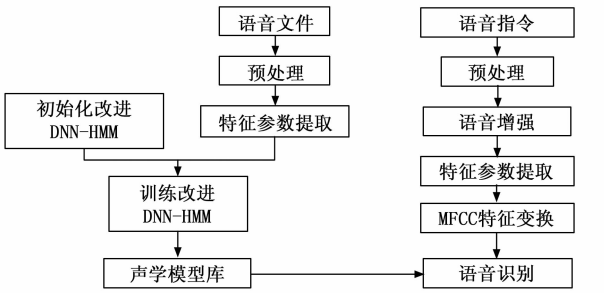

图 3　语音识别模型

在获得特征参数后，对 MFCC 特征进行变换，得到新的语音特征。最后，通过训练好的改进 DNN-HMM 模型进行语音识别。

3　实验结果及分析

为了验证研究提出的语音指令识别方法的性能，研究分别测试了其语音增强性能和语音识别性能，并搭建了一个简单的轮式移动机器人语音控制平台，用以进行指令识别测试。

为了验证研究提出的语音识别算法对语音指令的识别效果，研究在不同噪声环境下，分别以音素和孤立词作为识别单位，测试了轮式移动机器人语音控制平台对指令的识别结果。识别指令包括前进、转弯等简单指令和“前进 2 m 并向左转弯”等复杂指令。DNN 隐藏层数量及神经元数量分别为 3 和 512，批量大小为 128。不同算法的分段信噪比（SSNR，segmented signal-to-noise ratio）如表 1 所示。

表 1　不同算法的 SSNR

噪声		MIC	SHMM	DNNHE
babble	−5 dB	11.4	10.1	12.7
	0 dB	9.7	8.5	10.7
	5 dB	8.4	6.6	9.1
	10 dB	6.6	4.4	7.8
factory	−5 dB	15.4	12.6	16.0
	0 dB	13.5	10.2	14.4
	5 dB	11.6	8.1	13.0
	10 dB	9.4	5.8	10.5
white	−5 dB	17.0	13.2	18.5
	0 dB	14.0	9.2	14.3
	5 dB	11.5	6.3	12.4
	10 dB	8.9	3.2	10.6

由表 1 可知，基于乘法迭代准则（MIC，multiplication iteration criterion）的语音增强方法和基于稀疏隐马尔科夫（SHMM，sparse hidden Markov model）的语音增强方法在不同输入信噪比及噪声类型下的 *SSNR* 始终小于 DNNHE。以 factory 噪声为例，在 5 dB 的输入信噪比下，三者的 *SSNR* 分别为 11.6、8.1 和 13.0。上述结果表明，DNNHE 能有效实现不同噪声环境中的语音增强。不同算法的语音质量感观评价（PESQ，perceptual evaluation of speech quality）如表 2 所示。

表 2　不同算法的 PESQ

噪声		MIC	SHMM	DNNHE
babble	−5 dB	1.5	1.7	1.7
	0 dB	1.9	2.0	2.1
	5 dB	2.3	2.4	2.5
	10 dB	2.7	2.8	2.8
factory	−5 dB	1.9	1.8	2.2
	0 dB	2.4	2.4	2.6
	5 dB	2.7	2.6	2.8
	10 dB	3.0	2.9	3.0
white	−5 dB	2.0	1.0	2.1
	0 dB	2.2	1.3	2.4
	5 dB	2.4	1.7	2.7
	10 dB	2.7	2.2	2.8

由表 2 可知，不同噪声类型及输入信噪比下，DNNHE 的 *PESQ* 均高于 MIC 和 SHEE。以 factory 噪声为例，在 −5 dB的输入信噪比下，MIC、SHMM 和 DNNHE 的 *PESQ* 分别为 1.9、1.8 和 2.2。上述结果表明，经过语音增强后，DNNHE 的语音质量更优秀。为了验证研究提出的语音识别方法的性能，研究首先对不同语音特征下的性能进行了测试。不同语音特征下的字错误率如图 4 所示。

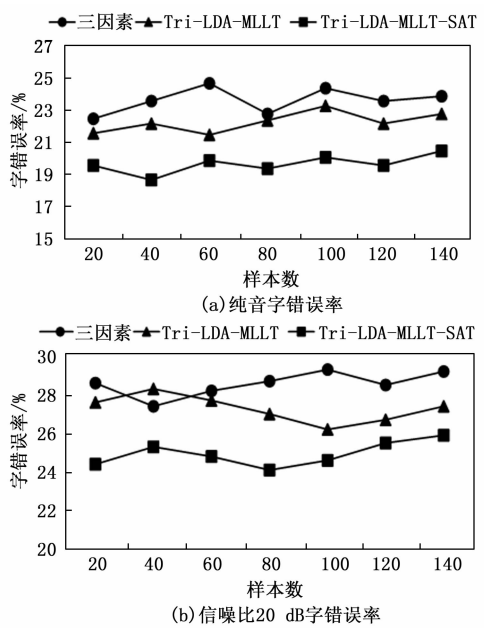

图 4　不同语音特征下的字错误率

由图 4（a）可知，相较于三因素特征和 Tri-LDA-MLLT 特征，研究提出的 Tri-LDA-MLLT-SAT 特征的

纯音字错误率更低。其中三因素特征和 Tri-LDA-MLLT 特征的纯音字评价错误率分别为 23.6% 和 22.3%，而 Tri-LDA-MLLT-SAT 特征的平均错误率仅为 19.7%。由图 4（b）可知，三因素特征、Tri-LDA-MLLT 特征和 Tri-LDA-MLLT-SAT 特征在信噪比 20 dB 下的平均字错误率分别为 28.6%、27.3% 和 24.9%，其中 Tri-LDA-MLLT-SAT 特征的错误率最低。上述结果表明，研究提出的 Tri-LDA-MLLT-SAT 特征能有效降低语音识别的错误率。不同识别模型的损失值如图 5 所示。

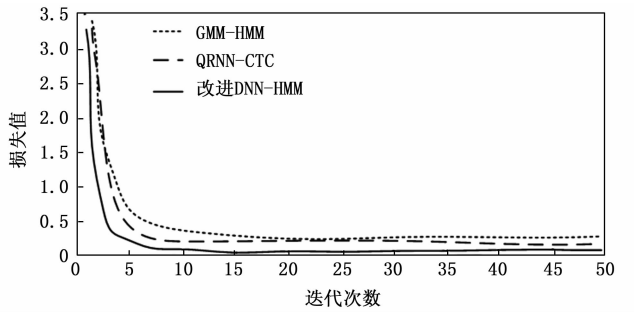

图 5 不同识别模型的损失值

由图 5 可知，相较于其他识别模型，改进 DNN-HMM 的损失值更低，且收敛速度更快。GMM-HMM、准循环神经网络和连接时序主义（QRNN-TC，quasi recurrent neural network connectionist temporal classification）分别在迭代 9 次和 7 次后收敛，二者的损失值分别为 0.4 和 0.2。而改进 DNN-HMM 在迭代 5 次后开始收敛，且损失值仅为 0.08。不同识别模型的字错误率如图 6 所示。

由图 6（a）可知，相较于其他模型，改进 DNN-HMM 的纯音字错误率更低。GMM-HMM、QRNN-TC 模型和改进 DNN-HMM 的平均纯音字错误率分别为 20.5%、23.1%和 17.9%。由图 6（b）可知，在信噪比20 dB 下，GMM-HMM、QRNN-CTC 和改进 DNN-HMM 的平均字错误率分别为 25.3%、26.3% 和 22.1%，其中 DNN-HMM 的字错误率最低。上述结果表明，改进 DNN-HMM 具有更高的识别准确率。为了进一步探究改进前后 DNN-HMM 模型的性能，研究不同滤波器组数量的声学模型进行了测试，实验结果如图 7 所示。

由图 7（a）可知，随着滤波器数量的增加，DNN-HMM 和改进 DNN-HMM 的句错误率均先降低后升高，但改进 DNN-HMM 的句错误率始终低于 DNN-HMM。以 30 组滤波器为例，改进 DNN-HMM 和 DNN-HMM 的句错误率分别为 20.7%和 21.7%。由图 7（b）可知，

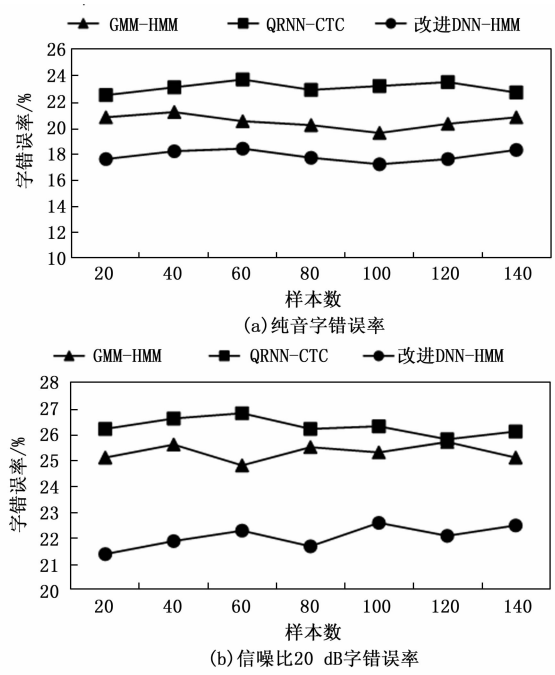

图 6 不同识别模型的字错误率

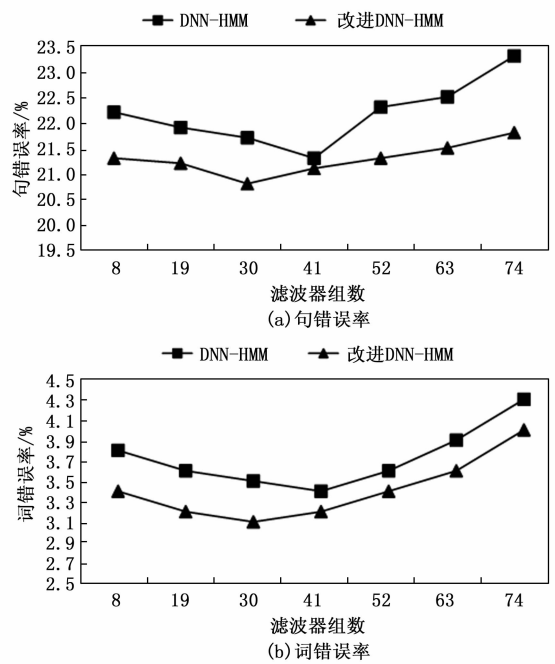

图 7 不同滤波器组数量的声学模型性能

随着滤波器数量的增加，DNN-HMM 和改进 DNN-HMM 的词错误率同样先降低后升高，且改进 DNN-HMM 的词错误率始终低于 DNN-HMM。当滤波器数量为 30 组，改进 DNN-HMM 的词错误率最低，仅为 3.1%低于 DNN-HMM。上述结果表明改进 DNN-HMM 的识别性能优于 DNN-HMM，且能通过增加滤波器组数提高模型的识别性能。但值得注意的是，滤波器的组数最好不超过 30 组。

4　结束语

在构建机器人语音控制系统时，考虑到环境噪声的干扰可能导致语音指令的清晰度降低。因此，为了提高语音控制的准确性，研究提出了基于 MFCC 特征及 DNN 的语音增强及识别方法。实验结果显示，DNNHE 能增强不同类型及信噪比下的语音，提高其 *SSNR* 和 *PESQ*。以 factory 噪声为例，在 5 dB 的输入信噪比下 DNNHE 的 *SSNE* 和 *PESQ* 分别为 13.0 和 2.8，均高于其他算法。此外，利用 Tri-LDA-MLLT-SAT 特征进行识别的字错误率更低，以 20 dB 信噪比为例，其字错误率仅为 24.9%，低于其他特征。同时，改进 DNN-HMM 的平均字错误率均低于其他声学模型，例如改进 DNN-HMM 的纯音字错误率仅为 17.9%。同时，相较于 DNN-HMM，改进 DNN-HMM 的句错误率和词错误率均更低。以 30 组滤波器为例，DNN-HMM 和改进 DNN-HMM 的句错误率分别为 21.7%和 20.7%，词错误率分别为 3.5%和 3.1%。上述结果表明，研究提出的基于 MFCC 特征提取和 DNN 的语音识别方法能准确识别噪声环境中的语音，实现对机器人的准确控制。但由于 DNN 的训练数据较为庞大，导致模型的训练时间偏长。因此，将尽可能地减少训练所需数据量，以降低模型的训练时间。

参考文献：

[1] 陶加贵，陈清淼，宋思齐，等. 基于 SDK 的 ABB 机器人语音控制方法研究 [J]. 微型电脑应用，2023，39 (10)：73－75.

[2] 梁　立，刘　宁，徐　炜，等. 语义理解下的地铁服务型机器人语音控制系统设计 [J]. 信息技术，2022，46 (11)：130－135.

[3] BOULAL H，HAMIDI M，ABARKAN M，et al. Amazigh CNN speech recognition system based on Mel spectrogram feature extraction method [J]. International Journal of Speech Technology，2024，27 (1)：287－296.

[4] 吕坤儒，吴春国，梁艳春，等. 融合语言模型的端到端中文语音识别算法 [J]. 电子学报，2021，49 (11)：2177－2185.

[5] KUMAR A，MITTAL V. Hindi speech recognition in noisy environment using hybrid technique [J]. International Journal of Information Technology，2021，13 (2)：483－492.

[6] 王雨佳. 基于语音合成的机器翻译机器人设计 [J]. 自动化与仪器仪表，2023 (4)：185－190.

[7] 蒙倩霞，余　江，常　俊，等. 基于 MFCC 特征的 Wi-Fi 信道状态信息人体行为识别方法 [J]. 计算机应用与软件，2022，39 (12)：125－131.

[8] 陶媛媛，陶　丹. 基于 DNN 与规则学习的机器翻译算法研究 [J]. 计算机测量与控制，2021，29 (1)：150－158.

[9] 杨淑莹，李　欣. 用于流式语音识别的轻量化端到端声学架构 [J]. 模式识别与人工智能，2023，36 (3)：268－279.

[10] 应　娜，吴顺朋，杨　萌，等. 基于小波散射变换和 MFCC 的双特征语音情感识别融合算法 [J]. 电信科学，2024，40 (5)：62－72.

[11] 胡　翔，杨　洋，蒋长江，等. 一种基于深度神经网络的电力系统调度控制语音识别模型 [J]. 电子器件，2023，46 (1)：90－95.

[12] 张晓艳，张天骐，葛宛营，等. 联合深度神经网络和凸优化的单通道语音增强算法 [J]. 声学学报，2021，46 (3)：471－480.

[13] 王小莉. 多语音和深度学习的对话机器人语音增强技术研究 [J]. 自动化与仪器仪表，2023 (12)：173－177.

[14] DEUERLEIN C，LANGER M，SENER J，et al. Human-robot-interaction using cloud-based speech recognition systems [J]. Procedia CIRP，2021，97 (2)：130－135.

[15] 张天骐，罗庆予，张慧芝，等. 复谱映射下融合高效 Transformer 的语音增强方法 [J]. 信号处理，2024，40 (2)：406－416.

[16] KATHRYN B W，RYAN W M，ELIZABETH A W. Hearing thresholds，speech recognition，and audibility as indicators for modifying intervention in children with hearing aids [J]. Ear and Hearing，2023，44 (4)：787－802.

[17] DUA M，BHAGAT B，DUA S，et al. A review on Gujarati language based automatic speech recognition (ASR) systems [J]. International Journal of Speech Technology，2024，27 (1)：133－156.

[18] 周　坤，陈文杰，陈伟海，等. 基于三次样条插值的扩展谱减语音增强算法 [J]. 北京航空航天大学学报，2023，49 (10)：2826－2834.

[19] JIANG Y C，JONG S Y，LAU W F，et al. Exploring the effects of automatic speech recognition technology on oral accuracy and fluency in a flipped classroom [J]. Journal of Computer Assisted Learning，2023，39 (1)：125－140.

[20] VEISI H，HOSSEINI H，MOHAMMADAMINI M，et al. Jira：a central Kurdish speech recognition system，designing and building speech corpus and pronunciation lexicon [J]. Language Resources and Evaluation，2022，56 (3)：917－947.

[21] 张德辉，董安明，禹继国，等. 融合门控循环单元及自注意力机制的生成对抗语音增强 [J]. 计算机科学，2023，50 (s2)：350－358.