基于波束形成的长短时记忆网络语音分离算法研究

兰朝凤* 刘 岩 赵宏运 刘春东

(哈尔滨理工大学测控技术与通信工程学院 哈尔滨 150080)

摘 要：在利用深度学习方式进行语音分离的领域，常用卷积神经网络(RNN)循环神经网络进行语音分离，但是该网络模型在分离过程中存在梯度下降问题，分离结果不理想。针对该问题，该文利用长短时记忆网络(LSTM)进行信号分离探索，弥补了RNN网络的不足。多路人声信号分离较为复杂，现阶段所使用的分离方式多是基于频谱映射方式，没有有效利用语音信号空间信息。针对此问题，该文结合波束形成算法和LSTM网络提出了一种波束形成LSTM算法，在TIMIT语音库中随机选取3个说话人的声音文件，利用超指向波束形成算法得到3个不同方向上的波束，提取每一波束中频谱幅度特征，并构建神经网络预测掩蔽值，得到待分离语音信号频谱并重构时域信号，进而实现语音分离。该算法充分利用了语音信号空间特征和信号频域特征。通过实验验证了不同方向语音分离效果，在60°方向该算法与IBM-LSTM网络相比，客观语音质量评估(PESQ)提高了0.59，短时客观可懂(STOI)指标提高了0.06，信噪比(SNR)提高了1.13 dB，另外两个方向上，实验结果同样证明了该算法较IBM-LSTM算法和RNN算法具有更好的分离性能。

关键词：语音分离；超指向波束形成；长短时记忆网络算法

中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2022)07-2531-08

DOI: 10.11999/JEIT210229

Research on Long Short-Term Memory Networks Speech Separation Algorithm Based on Beamforming

LAN Chaofeng LIU Yan ZHAO Hongyun LIU Chundong

(*College of Measurement and Communication Engineering, Harbin University of Science and Technology, Harbin* 150080, *China*)

Abstract: In the field of speech separation using deep learning, the Recurrent Neural Network (RNN) is commonly used for speech separation, but the network model has a gradient descent problem in the separation process, and the separation result is not ideal. Considering this problem, this paper uses Long Short-Term Memory (LSTM) network to explore the signal separation, which makes up for the deficiency of RNN network. The separation of multi-channel vocal signals is more complicated, and most of the separation methods used at this stage are based on the spectrum mapping method, and the spatial information of the voice signal is not effectively used. In response to this problem, this paper combines the beamforming algorithm and the LSTM network to propose a beamforming LSTM algorithm. The voice files of three speakers are randomly selected from the TIMIT voice library, and the super-pointing beamforming algorithm is used to obtain beams in three different directions. The spectral amplitude characteristics in each beam are extracted, and a neural network is constructed to predict the masking value. The to-be-separated speech signal spectrum is obtained. and the time-domain signal is constructed, and the speech separation is realized. The algorithm makes full use of the spatial characteristics of the speech signal and the signal frequency domain characteristics. The effect of speech separation in different directions is verified through experiments. Compared with the IBM-LSTM network, at 60-degree direction, this algorithm improves Perceptual Evaluation of Speech Quality (PESQ) by 0.59, Short-Time Objective Intelligibility (STOI) index by 0.06, and Signal to Noise Ratio (SNR) by 1.13 dB. At the other two reverse directions, the experimental results also prove that the algorithm has better separation performance than the IBM-LSTM algorithm and the RNN algorithm.

Key words: Speech separation; Super-directional beamforming; Long Short-Term Memory (LSTM) algorithm

收稿日期：2021-03-22；改回日期：2021-07-20；网络出版：2021-07-28

*通信作者： 兰朝凤 lanchaofeng@hrbust.edu.cn

基金项目：国家自然科学基金青年基金(11804068)，黑龙江省自然科学基金(LH2020F033)

Foundation Items: The National Natural Science Youth Foundation of China (11804068), The Natural Science Foundation of Heilongjiang Province (LH2020F033)

1 引言

语音信号分离问题最早起源于鸡尾酒会问题，该问题致力于解决在嘈杂环境中分离出重点关注的语音信号。经过学者不断努力，解决该问题的方法不断被创新，语音信号分离速度和分离质量都有所提高[1]。随着社会进步和智能家居的发展，语音信号处理知识被广泛应用于日常生活之中，对信号处理速度和质量提出了更高要求[2,3]。语音分离问题的解决方法主要可以归结为两个大类，分别为基于信号变换的传统方式和近年来流行的深度学习方式，传统分离方法主要是通过数字信号处理方式，对混合语音信号矩阵进行数学变化，使分离后语音信号彼此之间达到最大独立性来完成信号分离[4]。该方法为语音信号分离领域做出了一定贡献，但是其往往需要对混合语音信号施加限制条件，如ICA施加的是弱正交约束，最终得到一个具有分布式的信号表征从而实现数据降维目的，矢量量化模型对观测信号施加一种强约束，将数据拟合成两种彼此相互排斥模型，最终达到语音数据聚类目的[5,6]。但是在实际生活中，这些限制条件并不容易满足，因而在实际应用过程中，使用该方法的分离效果还有待提高。

随着计算机技术的不断发展，计算机运算速度逐渐提高，运算成本逐渐下降，基于深度学习的语音信号处理方式被众多学者提出并加以研究，在语音信号处理领域取得了一定成果[7–10]。深度学习网络(Deep Neural Network, DNN)结构是较早用于语音分离的网络，并且取得了一定进展[11]。Wang等人[12]最先提出将DNN应用于语音分离领域，并结合理想软模板和理想二值模板完成了语音分离任务，并对两种模板的分离结果做出了具体阐述分析。DNN具有多层次结构，可以从训练数据中抽取出更加抽象的特征并具有非常强大的非线性数据处理能力，但是其训练过程中存在大量参数计算，从而导致其模型收敛所需要时间更长的问题。随后，有学者利用卷积神经网络(Convolution Neural Network, CNN)模型探究了语音信号分离问题。Huang等人[13]将DNN和RNN模型结合起来应用于该问题，并在模型中加入了模板计算方法，该方法在模型中被称为确定层，通过确定层实现了对误差函数最小化操作，通过对误差函数优化和网络模型优化得到比DNN更好的语音分离结果。Hui等人[14]提出一种基于CMNN的结构，该结构结合理想幅值掩蔽(Ideal Ratio Mask, IRM)和maxout激活函数，实现对语音分离问题的建模，实验结果表明，相对于传统的CNN语音分离效果具有较大提升。Chandna等人[15]通过构建一种深度卷积网络模型，成功分离单通道低延迟的混合语音信号，其分离语音信号中包含鼓声、贝斯和随歌曲变化的其他种类乐器，在实验中研究人员还对该提出模型和多层感知器模型进行了对比实验，实验结果表明该模型无论从信号分离效果还是分离速度上都优于多层感知器。2014年，有学者提出了深度堆叠网络(Deep Stacking Network, DSN)用于语音信号分离任务，该网络是由多个神经网络堆叠而成的，并且后一层网络输入包含上一层网络输出和原始输入。Nie等人[16]给出了一种层级堆叠神经网络，并通过该网络对语音短时动态信息进行分析，此类网络提高了原始信号的估计精确度，但是其对于语音信号每一个频带估计过程中相互间是独立操作，没有考虑到频带相关性。其后有学者将循环神经网络(Recurrent Neural Network, RNN)应用于语音分离实践中，相较于卷积神经网络只关心数据局部信息特征，而忽略了语音信号前后联系的情况，RNN是一种时序模型，其在某一时刻的输出可以在下一个时刻作用其自身，因其结构具有循环链接特性，所以常用于时序信号的短时动态信息建模，并且其更加适用于语音信号这种与数据出现次序有关的信息处理，在语音分离领域取得了巨大成功。单层RNN因只有单个隐层，层级结构的缺乏令其在学习语音信号深层结构信息时具有缺陷性。随后有学者针对该问题提出了基于深层循环神经网络(Deep Recurrent Neural Network, DRNN)的语音信号分离方法，但是DRNN中还存在梯度消失问题有待解决[17]。

综上所述，深度学习方式解决语音分离问题主要依靠频域特征，没有对语音信号空间特征进行有效利用。针对深度学习中RNN梯度消失问题，本文提出一种基于长短时记忆网络(Long Short-Term Memory networks, LSTM)的语音分离方法，该方法既考虑到了语音信号时序相关性，又克服了传统RNN算法梯度消失问题。当前深度学习模型都是对语音信号进行频域特征提取，之后对该特征进行训练得到输入特征和关注语音信号频谱特征间非线性映射关系，从而解决语音分离问题，但是该方法的不足之处在于其分离依据是目标语音于干扰语音间频谱结构差异，若二者结构相似则其分离结果较差。针对该问题，本文结合波束形成算法和LSTM网络，提出了改进算法，充分利用了语音信号的空间特征和频谱特征并在具体实验中对分离结果进行验证。

2 神经网络模型

2.1 神经网络算法流程及算法基础介绍

利用深度学习方式更好对输入和输出特征进行

非线性拟合，相对于浅层网络，其更加具有优势。一般来说，监督性语音分离系统流程如图1所示。

图1给出了监督性学习实现步骤，主要分为5个子模块，首先通过时频分解模块将语音时域信号转换成2维时频信号；而后对语音信号进行特征提取操作，常用方法包括短时傅里叶变换谱、梅尔频率倒谱系数等；第3个模块是确定分离目标，后续分离过程中将利用此分离目标并结合观测信号分离出多路原始信号。分离目标选择和深度学习最终任务有关，常用分离目标有目标语音幅度谱估计和时频掩蔽目标等；第4个部分为模型训练过程，通过大量观测信号和纯净语音数据之间进行非线性映射，训练过程中动态调整神经网络参数，使其达到更好拟合效果；分离系统最后一个阶段是语音信号波形合成阶段，该阶段利用训练得到的分离模型对观测信号进行处理，而后通过傅里叶逆变换得到目标语音波形信号[18]。

RNN模型可以利用所有时刻的输入信息，并将其映射到不同输入单元中，对于语音信号等具有上下文关系的信息处理具有积极意义。但是RNN神经网络存在梯度消失问题，即某一时刻输出无法长时间对下一时刻造成影响，随着网络传播，作用效果越来越小，导致网络中单元只受到其附近单元影响，因而其并不适合处理具有长期依赖性的问题。

为解决RNN梯度消失问题，有学者提出了一种LSTM网络，该网络和RNN具有相同组织形式，但是相较于RNN，其神经元内部结构有所不同。LSTM的一个标准神经元包括了输出门、遗忘门和输入门。3个门相互配合使得信息可以长时间保存在网络中并进行上下文信息传递。当网络中输入门关闭时，就不会有新网络输入影响LSTM状态，那么可以将较为靠前的序列信息传递到序列后端，从而解决了梯度消失和梯度爆炸问题[19–22]。一个标准的LSTM网络梯度信息保存如图2所示。

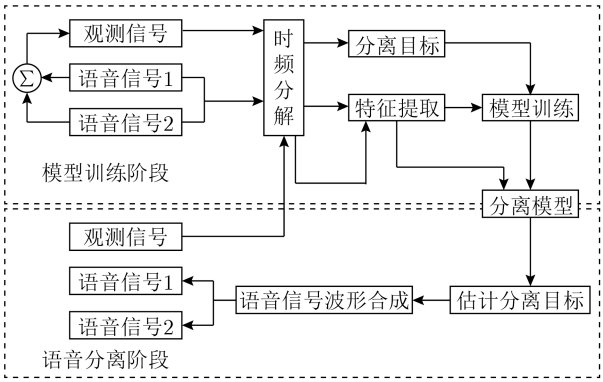

图 1 监督性语音分离系统流程图

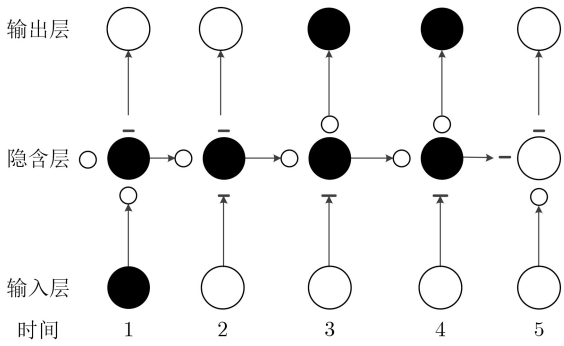

图 2 LSTM梯度信息保存示意图

如图2所示，在图2中每一个神经元左侧代表遗忘门状态，下方代表输入门状态，上方代表输出门状态，在图2中○代表开关打开，_代表开关关闭。在时刻1，网络输入门打开，新数据信息被输入到网络中，而后在时刻2, 3, 4输入门保持关闭，遗忘门保持打开，前一个时刻信息被传递到后一个时刻中，并且因为输入门关闭，当前LSTM神经网络状态不会被新网络输入所覆盖，所以位置较靠前的上下文信息被传递到了网络后端，解决了RNN梯度消失问题，在时刻3和时刻4，输出门保持打开状态，当前网络信息反映到输出层中。

2.2 神经网络算法流程及算法基础介绍

2.1小节中给出了对神经网络进行语音分离的具体流程，并介绍了常用于语音分离的神经网络，说明了LSTM网络相对于RNN具有的优势，本节将针对LSTM网络内部结构给出具体说明，并给出一种基于理想二值掩码的LSTM神经网络。典型LSTM网络记忆块如图3所示。

如图3所示，一个标准LSTM网络记忆块具有1个记忆单元和3个门控单元，网络利用图3中3个门控单元来控制记忆单元状态。门控单元分别表示输入门。输出门和遗忘门。某一时刻t，LSTM记忆块利用门状态的改变来对记忆块状态进行更新。更新过程由遗忘门状态更新、记忆单元状态更新和输出更新组成。

如图3中①所示，遗忘门在t时刻输出m_t由前一

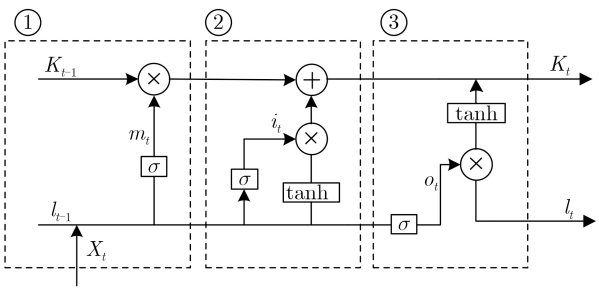

图 3 LSTM网络记忆块

个时刻输出l_{t-1}和当前时刻输入X_t决定，m_t可以表示为

$$m_t = \sigma(W_f \times [l_{t-1}, X_t] + b_f) \tag{1}$$

其中，W_f代表遗忘门权重；b_f代表遗忘门偏置；σ代表sigmod激活函数。遗忘门通过l_{t-1}, X_t状态来确定一个0～1的值，通过这个值来确定上一时刻学习到的信息K_{t-1}作用于下一时刻的比例。

如图3中②所示，其功能是对记忆单元状态进行更新。更新内容包含两部分内容，第1部分是通过式(2)决定哪些值需要更新，另一部分为通过式(3)生成新候选值。最后通过式(4)完成记忆单元输出，运算过程为

$$i_t = \sigma(W_i \times [l_{t-1}, X_t] + b_i) \tag{2}$$

其中，i_t代表输入门输出结果，通过sigmod激活函数来决定哪些值用来进行更新；W_i代表输入门权重；b_i代表输入门偏置

$$\hat{K}_t = \tanh(W_k \times [h_{t-1}, X_t] + b_k) \tag{3}$$

其中，$\hat{K}_t$表示通过tanh函数生成的新候选值，W_k代表记忆单元权重，b_k代表记忆单元偏置，

$$K_t = m_t * K_{t-1} + i_t * \hat{K}_t \tag{4}$$

通过式(4)完成记忆单元输出，其主要由式(2)、式(3)两部分构成，其中$m_t * K_{t-1}$代表遗忘门对上一时刻信息的遗忘，$i_t * \hat{K}_t$代表新加入状态信息，将两部分作和，最终得到了K_t，即当前时刻记忆单元状态，也是当前时刻t的输出。

如图3中③所示，该过程代表输出门控制的输出信息，需要输出的记忆单元状态信息通过sigmod激活函数来进行控制更新，可表示为

$$o_t = \sigma(W_o \times [l_{t-1}, X_t] + b_o) \tag{5}$$

其中，o_t代表输出门输出结果；W_o代表输出门权重；b_o代表输出门偏置。通过tanh函数来对当前时刻记忆单元状态进行，最后输出结果l_t是两部分乘积，可表示为

$$l_t = o_t * \tanh(K_t) \tag{6}$$

利用二值掩码结合LSTM网络来进行语音信号分离训练有望达到比RNN更好的分离效果。IBM-LSTM网络结果如图4所示。

图4中，在训练阶段，网络输入是混合语音信号时频谱，通过与纯净语音信号时频谱进行非线性映射，形成一个通过LSTM网络训练得到的二值掩码模型；在分离阶段通过该模型估计出对应语音信号的二值掩码，而后通过短时傅里叶逆变换得到原始语音信号，完成语音分离任务。

3 超指向波束形成算法

随着当前深度学习领域和人工智能快速发展，通过语音信号实现智能交互过程已经成为现实，其对语音拾音系统也提出了新要求，利用单个麦克风进行信号采集已经无法适应当前环境，通过麦克风阵列进行语音信号采集变成了一种必然趋势。相对于单个麦克风语音采集情况，麦克风阵列可以有效采集到发声场原始语音信号空间信息，通过特定波束形成算法可以实现盲通道辨识、语音信号增强、盲源信号分离等多种目标[23,24]。

波束形成算法是针对麦克风阵列提出的一种信号处理算法，其可以实现声源信号定位和定向、语音信号增强和分离等操作[25–27]。其最基本原理是利用麦克风阵列得到的语音信号空间信息，建立一个增益随方位角和距离变化的空域滤波器。常用的波束形成方法有延迟求和波束形成算法、差分波束形成、超指向波束形成算法等[28–31]。本文就波束形成基本原理进行叙述，进而引出本文中使用的超指向波束形成算法，该算法相对于其他波束形成算法，对来自非导向方向语音信号抑制作用更强，更适用于本文所提出的语音分离模型。

波束形成过程如图5所示，其中包含两个过程，分别代表滤波和信号叠加。波束形成器由不同通道所对应的滤波器共同组成，其大多数是在频域进行设计，通过短时傅里叶变换方式实现。

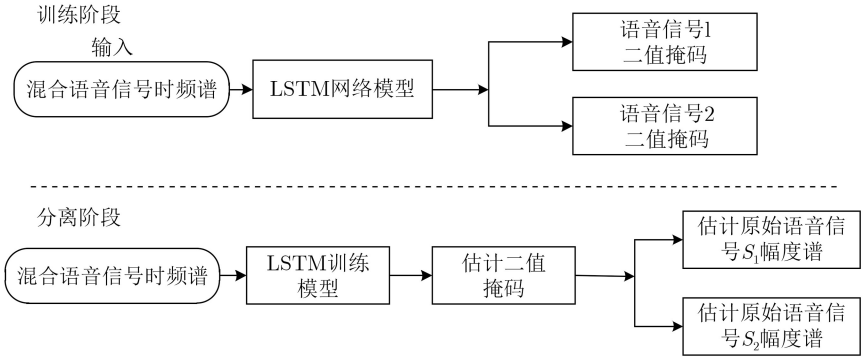

图 4 波束形成频域求和结构示意图

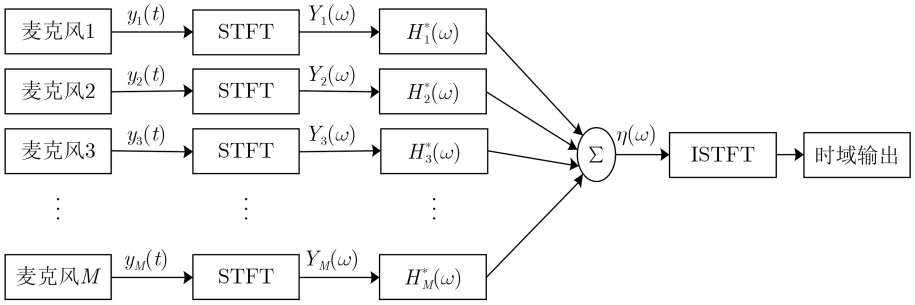

图 5 波束形成频域求和结构示意图

波束形成过程将给定时刻和频带上的阵列观测信号叠加变成一个向量，该向量可以表示为

$$\boldsymbol{y}(\omega)=[Y_1(\omega),Y_2(\omega),\cdots,Y_M(\omega)]^{\mathrm{T}} \tag{7}$$

其中，M代表麦克风阵列中包含的麦克风数量；ω代表频率。

经过上述过程后，波束形成算法输出表示为

$$\eta(\omega)=\sum_{m=1}^{M}\boldsymbol{H}_m^*(\omega)Y_m(\omega)=\boldsymbol{h}^{\mathrm{H}}(\omega)y(\omega) \tag{8}$$

其中，$*$代表共轭；H代表矩阵的共轭转置；$\boldsymbol{h}(\omega)=[H_1(\omega),H_2(\omega),\cdots,H_M(\omega)]^{\mathrm{T}}$代表当前阵列的波束形成滤波器，并且其是一个与$\boldsymbol{y}(\omega)$具有相同维度的复向量，在$h(\omega)$确定的前提下，麦克风阵列空域响应可以表示为

$$\boldsymbol{\alpha}(\omega,\vartheta)=\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{d}(\omega,\vartheta) \tag{9}$$

其中，ϑ代表方向，其由方位角和俯仰角共同决定，$\boldsymbol{d}(\omega,\vartheta)$代表阵列导向矢量，其与阵列几何结构、方向等有关。式(9)显示了波束形成算法对不同方向信号的响应特性，空域响应在期望语音信号方向具有最大幅度值，并且会对其他方向信号产生一定程度的衰减。

空域响应幅度随方向变化的曲线被称为波束图，如果设置0°方向为主瓣，其他方向为旁瓣，在这种情况下，波束形成器性能可以通过指向性因子进行衡量，形成器主瓣窄旁瓣低时性能较好，指向性因子可表达为

$$\boldsymbol{\beta}(\omega)=\frac{\left|\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{d}(\omega,\vartheta_0)\right|^2}{\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{h}(\omega)} \tag{10}$$

其中，ϑ_0代表所期望的导向方向；$\boldsymbol{\Gamma}(\omega)$代表各向同性噪声协方差矩阵。

为实现对非导向方向语音信号最大抑制效果，需要使波束形成器指向性越高越好。实现该目标最简单的方式是极大化指向性因子，通过这种方式设计出来的波束形成器称为超指向波束形成器，超指向波束形成算法在期望声源方向无失真约束条件下，可以将问题简化为

$$\min_{\boldsymbol{h}(\omega)}\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{h}(\omega) \tag{11}$$

在满足$\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{d}(\omega,\vartheta_0)=1$条件下求出上式的最小值。此条件下通过推导可得到当前阵列的超指向波束形成滤波器$\boldsymbol{h}_{\mathrm{SD}}(\omega)$，表示为

$$\boldsymbol{h}_{\mathrm{SD}}(\omega)=\frac{\boldsymbol{\Gamma}^{-1}(\omega)\boldsymbol{d}(\omega,\vartheta_0)}{\boldsymbol{d}^{\mathrm{H}}(\omega,\vartheta_0)\boldsymbol{\Gamma}^{-1}(\omega)\boldsymbol{d}(\omega,\vartheta_0)} \tag{12}$$

超指向波束形成实质是在保证期望声源信号不失真的前提下，尽量对来自其他方向的语音信号进行抑制的过程，因此将其和上文中提到过的LSTM神经网络相结合，可以在神经网络基础上引入语音信号方向信息，有望进一步提高神经网络模型语音信号分离效果。

4　算法分离流程

传统深度学习模型仅仅利用观测信号频谱信息和原始信号频谱信息进行非线性映射操作，而忽略了语音信号空间信息。针对该问题，本文利用麦克风阵列，提出一种超指向波束形成算法和LSTM网络结合模型，通过波束形成算法得到3个不同方向语音波束信号，而后该算法提取的特征是每一个波束中的频谱幅度特征，通过本文构建的LSTM网络预测掩蔽值，通过掩蔽值得到待分离语音信号频谱并重构出时域信号，实现语音分离。分离算法流程图如图6所示。

由图6可见，利用合适的麦克风阵列对3路语音信号进行采集，通过超指向波束形成算法，得到3个不同方向的指向性波束。对3个波束信号进行频

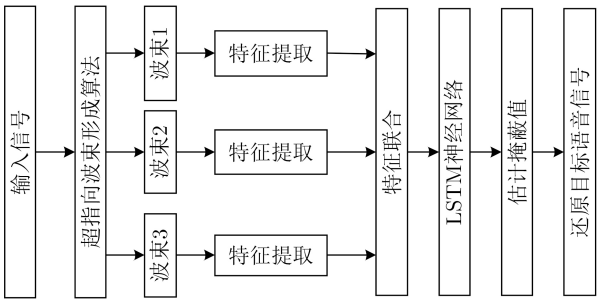

图 6 分离算法流程图

谱幅度特征提取进而得到联合特征，根据数据每个维度上特征值的平均值和标准差对联合特征进行标准化操作。当前联合特征作为LSTM网络输入，根据目标语音信号频谱幅度特征，通过网络训练得到掩蔽值，根据掩蔽值得到目标语音频谱，进行语音信号重构得到原始时域目标语音信号，完成语音信号分离工作。

5 实验数据采集及设置

实验过程中，通过TIMIT语音库随机选择3名说话人语音信号，并且说话人年龄和性别均保持随机抽取，在进行语音录制前将3段语音信号裁剪成相同时间长度并进行幅度归一化操作，阵列布放及声源位置如图7所示，本文通过图7的布放方式进行语音信号录制。

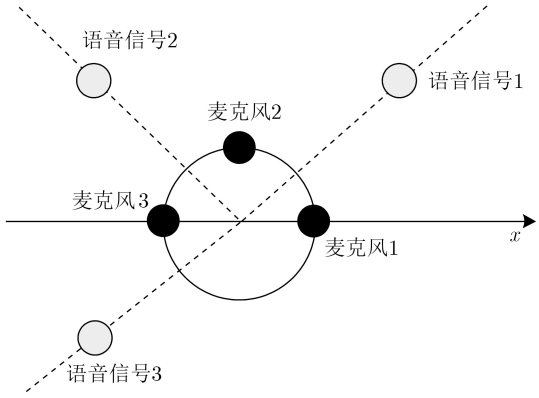

图 7 阵列布放及声源位置

如图7所示，采用3个麦克风组成麦克风阵列，阵列中每个麦克风距离圆形中心距离为50 mm，并且保持在同一水平面上，图中所示箭头方向作为0°方向，将圆形按照逆时针方向均匀划分成6个区域，每个区域60°，3个说话人保持和麦克风在同一水平面上，距离圆彩中心位置2.5 m，方向分别为60°, 120°, 240°，同时播放3段语音信号，通过麦克风录制实验所需要的训练数据，本文共录制了接近7000条语音数据用于神经网络训练过程。

本文所用神经网络结构如图8所示，神经网络结构由1层掩蔽层、3层LSTM层和1层全连接层构成。其中全连接层中包含600个节点，训练中用到的损失函数为mean_squared_error函数，即最小均方误差(Mean Squared Error, MSE)函数，在训练过程中，损失函数值越小，说明神经网络和训练集拟合性越好，匹配度越高。

实验中语音信号采样频率统一为16 kHz，帧长设置为512点，采用3层LSTM网络结构，每层由600个神经元组成。在60°, 120°, 240°方向进行超指向波束形成，得到不同方向的语音波束信号，进而得到不同波束频谱幅度特征，将特征拼接起来得到联合特征，将联合特征作为网络输入，结合纯净语音信号频谱幅度特征，得到一个二值掩蔽训练模型。分离阶段通过模型得到混合语音信号对应的二值掩蔽，将其作用于混合语音信号幅度谱得到待分离语音信号幅度谱，重构原始语音信号，达到语音信号分离的目的。由上文理论分析可知，相对于传统LSTM网络，该方法不仅利用了观测信号频谱信息，还通过波束形成算法利用了观测信号空间信息，有望得到较好的分离结果。

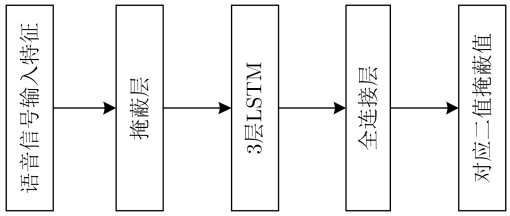

图 8 LSTM神经网络结构

6 实验结果及分析

为量化实验结果，本文通过客观语音质量评估(Perceptual Evaluation of Speech Quality, PESQ)、短时客观可懂(Short-Time Objective Intelligibility, STOI)、信噪比(Signal to Noise Ratio, SNR)指标对分离结果进行评价，对提出的波束形成LSTM网络分离效果进行测试，并将实验结果同IBM-LSTM方法进行对比，并将本文结果同RNN分离结果对比，本文以60°, 120°及240°方向语音信号的分离结果为例，不同网络分离结果如表1所示。

由表1可知，结合波束形成算法的LSTM网络，综合利用了语音信号的频谱信息和空间信息，相较于仅仅应用频谱信息的神经网络在语音分离效果上有所提高，在60°方向时，波束形成LSTM网络与IBM-LSTM网络相比，PESQ提高了0.59 dB，STOI指标提高了0.06，SNR提高了1.13。与RNN相比，PESQ提高了0.76，STOI指标提高了0.09，SNR提高了2.16 dB。在120°方向时，波束形成LSTM网络与IBM-LSTM网络相比，PESQ提高了0.56，STOI指标提高了0.05，SNR提高了1.13 dB。与RNN相比，PESQ提高了0.76，STOI指标提高了0.09，SDR提高了2.18 dB。在240°方向时，由表1可得到与上述两种角度相同的结论，即波束形成LSTM网络相较于另外两种算法在语音分离评价指标上均有所提高。实验结果表明，结合超指向波束形成的LSTM网络相较于IBM-LSTM, RNN在语

表 1　不同网络结构分离人声信号结果

评价指标 网络结构	观测信号角度(°)	PESQ	STOI	SNR (dB)
波束形成LSTM	60	3.34	0.91	6.75
IBM-LSTM		2.75	0.85	5.62
RNN		2.58	0.82	4.59
波束形成LSTM	120	3.28	0.89	6.74
IBM-LSTM		2.72	0.84	5.61
RNN		2.52	0.80	4.56
波束形成LSTM	240	3.32	0.91	6.74
IBM-LSTM		2.76	0.84	5.60
RNN		2.54	0.81	4.54

音分离领域取得了更好的分离效果，证明了本文所提算法的有效性。

7　结论

本文提出的超指向波束形成LSTM算法，通过麦克风阵列获得语音信号空间信息，定向对某一个方向上的语音信号进行波束形成，降低其他方向声源干扰，并弥补了神经网络仅通过语音信号频谱信息进行信号分离的弊端，实现了多个说话人语音混合信号分离效果的提高。实验结果表明：在60°方向上，本文算法与LSTM网络相比，PESQ提高了0.59，STOI指标提高了0.06，SNR提高了1.13 dB。与RNN相比，PESQ提高了0.76，STOI指标提高了0.09，SNR提高了2.16 dB。在120°和240°方向上结果同样优于另外两种对比算法，由此表明本文提出基于超指向波束形成LSTM算法的有效性。

参 考 文 献

[1] EPHRAT A, MOSSERI I, LANG O, *et al*. Looking to listen at the cocktail party: A speaker–independent audio–visual model for speech separation[J]. *ACM Transactions on Graphics*, 2008, 37(4): 109:1–109:11.

[2] JONES G L and LITOVSKY R Y. A cocktail party model of spatial release from masking by both noise and speech interferers[J]. *The Journal of the Acoustical Society of America*, 2011, 130(3): 1463–1474. doi: 10.1121/1.3613928.

[3] XU Jiaming, SHI Jing, LIU Guangcan, *et al*. Modeling attention and memory for auditory selection in a cocktail party environment[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018.

[4] 黄雅婷, 石晶, 许家铭, 等. 鸡尾酒会问题与相关听觉模型的研究现状与展望[J]. 自动化学报, 2019, 45(2): 234–251.
HUANG Yating, SHI Jing, XU Jiaming, *et al*. Research advances and perspectives on the cocktail party problem and related auditory models[J]. *Acta Automatica Sinica*, 2019, 45(2): 234–251.

[5] 李娟. 基于ICA和波束形成的快速收敛的BSS算法[J]. 山西师范大学学报: 自然科学版, 2018, 32(4): 52–56.
LI Juan. A fast-convergence algorithm combining ICA and beamforming[J]. *Journal of Shanxi Normal University: Natural Science Edition*, 2018, 32(4): 52–56.

[6] 陈国良, 黄晓琴, 卢可凡. 改进的快速独立分量分析在语音分离系统中的应用[J]. 计算机应用, 2019, 39(S1): 206–209.
CHEN Guoliang, HUANG Xiaoqin, and LU Kefan. Application of improved fast independent component analysis in speech separation system[J]. *Journal of Computer Applications*, 2019, 39(S1): 206–209.

[7] 王昕, 蒋志翔, 张杨, 等. 基于时间卷积网络的深度聚类说话人语音分离[J]. 计算机工程与设计, 2020, 41(9): 2630–2635.
WANG Xin, JIANG Zhixiang, ZHANG Yang, *et al*. Deep clustering speaker speech separation based on temporal convolutional network[J]. *Computer Engineering and Design*, 2020, 41(9): 2630–2635.

[8] 崔建峰, 邓泽平, 申飞, 等. 基于非负矩阵分解和长短时记忆网络的单通道语音分离[J]. 科学技术与工程, 2019, 19(12): 206–210. doi: 10.3969/j.issn.1671-1815.2019.12.029.
CUI Jianfeng, DENG Zeping, SHEN Fei, *et al*. Single channel speech separation based on non–negative matrix factorization and long short–term memory network[J]. *Science Technology and Engineering*, 2019, 19(12): 206–210. doi: 10.3969/j.issn.1671-1815.2019.12.029.

[9] 陈修凯, 陆志华, 周宇. 基于卷积编解码器和门控循环单元的语音分离算法[J]. 计算机应用, 2020, 40(7): 2137–2141.
CHEN Xiukai, LU Zhihua, and ZHOU Yu. Speech separation algorithm based on convolutional encoder decoder and gated recurrent unit[J]. *Journal of Computer Applications*, 2020, 40(7): 2137–2141.

[10] WANG Deliang and CHEN Jitong. Supervised speech separation based on deep learning: An overview[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, 26(10): 1702–1726. doi: 10.1109/TASLP.2018.2842159.

[11] 刘文举, 聂帅, 梁山, 等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报, 2016, 42(6): 819–833.
LIU Wenju, NIE Shuai, LIANG Shan, *et al*. Deep learning based speech separation technology and its developments[J]. *Acta Automatica Sinica*, 2016, 42(6): 819–833.

[12] WANG Yuxuan, NARAYANAN A, and WANG Deliang. On training targets for supervised speech separation[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(12): 1849–1858. doi: 10.1109/TASLP.2014.2352935.

[13] HUANG P S, KIM M, HASEGAWA–JOHNSON M, *et al*. Deep learning for monaural speech separation[C]. 2014 IEEE International Conference on Acoustics, Speech and

Signal Processing (ICASSP), Florence, Italy, 2014: 1562–1566.

[14] HUI Like, CAI Meng, GUO Cong, *et al*. Convolutional maxout neural networks for speech separation[C]. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, United Arab Emirates. 2015: 24–27.

[15] CHANDNA P, MIRON M, JANER J, *et al*. Monoaural audio source separation using deep convolutional neural networks[C]. The 13th International Conference, Grenoble, France, 2017: 258–266.

[16] NIE Shuai, ZHANG Hui, ZHANG Xueliang, *et al*. Deep stacking networks with time series for speech separation[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014: 6667–6671.

[17] GERS F A, SCHMIDHUBER J, and CUMMINS F. Learning to forget: Continual prediction with LSTM[J]. *Neural Computation*, 2000, 12(10): 2451–2471. doi: 10.1162/089976600300015015.

[18] 梁尧, 朱杰, 马志贤. 基于深度神经网络的单通道语音分离算法[J]. 信息技术, 2018, 42(7): 24–27.
LIANG Yao, ZHU Jie, and MA Zhixian. A monaural speech separation algorithm based on deep neural networks[J]. *Information Technology*, 2018, 42(7): 24–27.

[19] 李文杰, 罗文俊, 李艺文, 等. 基于可分离卷积与LSTM的语音情感识别研究[J]. 信息技术, 2020, 44(10): 61–66.
LI Wenjie, LUO Wenjun, LI Yiwen, *et al*. Speech emotion recognition based on separable convolution and LSTM[J]. *Information Technology*, 2020, 44(10): 61–66.

[20] WESTHAUSEN N L and MEYER B T. Dual–signal transformation LSTM network for real–time noise suppression[EB/OL]. https://arxiv.org/abs/2005.07551,2020.

[21] GREZES F, NI Zhaoheng, TRINH V A, *et al*. Combining spatial clustering with LSTM speech models for multichannel speech enhancement[EB/OL]. https://arxiv.org/abs/2012.03388,2020.

[22] LI Xiaofei and HORAUD R. Online monaural speech enhancement using delayed subband LSTM[EB/OL]. https://arxiv.org/abs/2005.05037, 2020.

[23] 潘超, 黄公平, 陈景东. 面向语音通信与交互的麦克风阵列波束形成方法[J]. 信号处理, 2020, 36(6): 804–815.
PAN Chao, HUANG Gongping, and CHEN Jingdong. Microphone array beamforming: An overview[J]. *Journal of Signal Processing*, 2020, 36(6): 804–815.

[24] 朱训谕, 潘翔. 基于麦克风线阵的语音增强算法研究[J]. 杭州电子科技大学学报: 自然科学版, 2020, 40(5): 30–33, 72.
ZHU Xunyu and PAN Xiang. Research on speech enhancement algorithm based on microphone linear array[J]. *Journal of Hangzhou Dianzi University*: *Natural Science*, 2020, 40(5): 30–33, 72.

[25] KIM H S, KO H, BEH J, *et al*. Sound source separation method and system using beamforming technique[P]. USA Patent. 008577677B2, 2013.

[26] ARAKI S, SAWADA H, and MAKINO S. Blind speech separation in a meeting situation with maximum SNR beamformers[C]. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing–ICASSP'07, Honolulu, USA, 2007, 1: I–41–I–44.

[27] SARUWATARI H, KURITA S, TAKEDA K, *et al*. Blind source separation combining independent component analysis and beamforming[J]. *EURASIP Journal on Advances in Signal Processing*, 2003, 2003: 569270. doi: 10.1155/S1110865703305104.

[28] WANG Lin, DING Heping, and YIN Fuliang. Speech separation and extraction by combining superdirective beamforming and blind source separation[M]. NAIK G and WANG Wenwu. Blind Source Separation. Heidelberg: Springer, 2014: 323–348.

[29] XENAKI A, BOLDT J B, and CHRISTENSEN M G. Sound source localization and speech enhancement with sparse Bayesian learning beamforming[J]. *The Journal of the Acoustical Society of America*, 2018, 143(6): 3912–3921. doi: 10.1121/1.5042222.

[30] QIAN Kaizhi, ZHANG Yang, CHANG Shiyu, *et al*. Deep learning based speech beamforming[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018: 5389–5393.

[31] HIMAWAN I, MCCOWAN I, and LINCOLN M. Microphone array beamforming approach to blind speech separation[C]. The 4th International Workshop, Brno, The Czech Republic, 2007: 295–305.

兰朝凤：女，1981年生，博士，博士生导师，研究方向为智能语音人机交互、图像处理、噪声控制等.

刘 岩：男，1995年生，硕士，研究方向为语音信号处理、语音分离、深度学习.

赵宏运：男，1994年生，硕士，研究方向为语音识别、深度学习.

刘春东：女，1996年生，硕士，研究方向为语音增强、深度学习.

责任编辑：余 蓉