Computer Science and Application 计算机科学与应用, 2025, 15(4), 394-405
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154112

Hans汉斯

基于语音活动检测的阵列信号测向研究

李纪元[1]，田益民[2*]，赵乾曜[1]，孙兆永[2]

[1]北京印刷学院信息工程学院，北京
[2]北京印刷学院基础部，北京

收稿日期：2025年3月25日；录用日期：2025年4月23日；发布日期：2025年4月30日

摘　要

为研究长时信号中对具有特定特征的声音来源方向进行检测的问题，本课题提出一种基于多特征自适应的语音信号活动检测对长时阵列信号进行检测，将结合多子空间拟合(MUSIC)算法与语音活动检测(VAD)技术，提出一种新型的信号处理方法，旨在提高对特征明显且目标具有特定属性的信号源的检测精度和定位准确性。通过语音信号MFCC特征和语音信号能量特征来设置自适应阈值，对特定声源的特征进行语音活动检测，以提高语音活动检测的准确性。再通过检测到的语音信号活动片段进行阵列信号测向，通过MUSIC算法实现对长时信号中不同时段不同来源方向的特定声源进行检测。

关键词

语音活动检测，阵列信号测向，MUSIC算法，MFCC

Research on Direction of Arrival Based on Voice Activity Detection

Jiyuan Li[1], Yimin Tian[2*], Qianyao Zhao[1], Zhaoyong Sun[2]

[1]Faculty of Information Engineering, Beijing Institute of Graphic Communication, Beijing
[2]Foundation Department, Beijing Institute of Graphic Communication, Beijing

Received: Mar. 25th, 2025; accepted: Apr. 23rd, 2025; published: Apr. 30th, 2025

Abstract

To investigate the problem of detecting the direction of sound sources with specific features in long-term signals, this project proposes a voice signal activity detection method based on multi feature adaptation for detecting long-term array signals. By combining the Multi Subspace Fitting (MUSIC) algorithm with Voice Activity Detection (VAD) technology, a new signal processing method is

*通讯作者。

文章引用: 李纪元, 田益民, 赵乾曜, 孙兆永. 基于语音活动检测的阵列信号测向研究[J]. 计算机科学与应用, 2025, 15(4): 394-405. DOI: 10.12677/csa.2025.154112

proposed to improve the detection and localization accuracy of signal sources with obvious features and specific target attributes. By setting adaptive thresholds based on the MFCC features and energy features of voice signals, voice activity detection can be performed on specific sound source features to improve the accuracy of voice activity detection. Then, the direction of arrival is determined by detecting active voice signal segments, and the MUSIC algorithm is used to detect specific sound sources in different time periods and source directions in long-term signals.

Keywords

Voice Activity Detection (VAD), Direction of Arrival (DOA), MUSIC Arithmetic, MFCC

1. 引言

随着语音信号处理技术的迅速发展，语音信号的阵列信号测向成为了一个备受关注的领域。在通信、语音识别、音频处理等领域中具有广泛的应用。随着智能音频设备的广泛应用和语音控制技术的兴起，对于特定语音信号的检测与定位需求日益增长[1]。例如，在智能家居系统中，需要准确地识别家庭成员的指令，并根据不同声源的位置调节相应设备；在语音会议系统中，需要将不同方向的说话人识别并进行语音定位，以提升会议的交流效率。在语音信号处理中，对具有特定特征的目标进行检测和测向具有重要的意义和实际应用价值。特别是针对嘈杂环境下的语音信号或特定人声言语信号，如语音识别、语音定位等任务，对有特定特征的目标进行准确的测向和检测至关重要。

传统的语音信号处理方法往往无法有效应对嘈杂环境下的语音信号，特别是在多个声源混叠的情况下，对特定目标进行准确的测向成为了一项具有挑战性的任务。因此，如何结合现代信号处理技术和阵列信号测向算法，实现对具有特殊特征语音信号的检测和测向，成为当前研究的热点之一。本文旨在通过探讨语音信号的阵列信号测向理论与实践，致力于为有特殊特征的语音目标的检测提供新的方法和技术支持，从而推动语音信号处理领域的发展和应用。

语音活动检测(Voice Activity Detection, VAD)是语音处理领域中的一个重要研究课题，关于基于语音活动检测的研究主要有：1) 传统的语音活动检测方法通常使用能量门限、短时能量、短时过零率等基本特征来判断语音信号和非语音信号的区别。刘思伟[2]等通过对每一帧的过零率和能量进行分析，通过过零率的变化和能量的大小来判断是否为语音。但过零率和能量两个特征不能用于特定特征的目标上。雷静[3]等通过信噪比对噪音进行分类来设置判决条件以提高对噪音判断的准确性，在对高信噪比情况下采用四个特征中任意两种小于阈值的判别法，低信噪比情况下采用任意三种小于阈值的判别法。在每一帧对阈值参数进行实时更新。赵新燕[4]等，通过估计语音短时信噪比，设置 MFCC 倒谱距离乘数，再根据信噪比设置检测门限进行检测，但是单一特征可能导致检测结果不准确。2) 基于统计模型算法 Bao [5]等使用基于高斯混合模型的隐马尔可夫模型来对语音信号进行 VAD 检测，但是使用统计模型不能做到实时性。3) 基于深度学习的算法。Kang [6]等采用深度学习网络对语音片段和非语音片段进行训练，训练出能够区别语音非语音的模型，以达到语音端点检测的目的，但是使用深度学习等方法需要大量的语音数据和非语音数据用来训练，而特定特征的应用场景不具备这样的条件，所以不方便使用。

语音活动检测也被用于语音信号增强，语音信号定位，语音信号测向等方面，一般被用作语音信号

的前段处理，用于计算语音信号的信噪比，语音降噪等处理[3]。黄毅伟等[7]将语音活动检测用于阵列信号声源定位上，在声源等位之前对声音进行预处理，区别出语音活动段和静音段用于后面的定位，但是在面对多个定位声源的时候不能对单一具有特定特征的目标进行定位。Catic [8]等通过在双耳助听器上运用 VAD 技术，区分收到的音频信息的语音和噪音并对噪音进行降噪等语音增强处理。Zhehui Zhu [9]等采用多特征结合外加自适应技术对语音信号进行 VAD 操作并且对 VAD 结果进行平滑操作提出一种软 VAD 的方法，以便于更好的对语音信号进行语音增强操作，但对于特定特征目标测向软 VAD 技术不可以明确分辨目标。

针对语音信号的阵列信号测向技术已经取得了一定的进展。其中，一些研究提出了基于声源空间特征、波束形成等方法来改进测向算法的性能；另外，一些研究借鉴机器学习和深度学习技术，提出了更加高效的信号处理方法。然而，在特定特征目标检测领域，仍存在一些挑战，例如对于背景噪声的抑制、目标定位精度的提高等方面需要进一步研究。

在阵列信号测向技术上使用语音活动检测技术已经是非常成熟，一般用于预处理，为了减少噪音对阵列信号测向的干扰，Küçük 等[10]用 VAD 技术处理音频后再运用于实时卷积神经网络的 DOA 估计，以便于减少噪音对 DOA 估计的影响。Varzandeh 等[11]利用空间特征与称为周期度(PD)的听觉启发的周期性特征的组合作为卷积神经网络(CNN)的输入特征来替代 VAD 使用进行 DOA 估计。在处理具有特定特征的目标声源测向时，本课题选取对于阵列信号目标测向技术中存在的缺少对具有特定特征的目标进行测向的问题，本课题利用多特征自适应 VAD 检测技术，对目标的特定特征进行提取，通过 VAD 技术确定目标声源的语音活动，再对检测到的语音活动段进行 MUSIC 算法的 DOA 估计，来实现对具有特定特征的声源目标进行测向的目的。

本文提出一种基于语音活动检测的目标测向技术，其中使用语音信号 MFCC 特征和能量谱特征来对特定特征目标进行语音活动检测，对两种特征进行加权融合，并设置自适应阈值进行检测，再对检测结果提取进行下一步的基于 MUSIC 算法语音测向。解决在特定场合对具有特定特征的目标进行语音测向的问题。

2. 阵列信号模型

通过立体阵列传声器采集到的语音信号为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \left[a(\theta_1) a(\theta_2) \cdots a(\theta_D) \right] \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_D \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix}$$

其中 X 为阵列传声器接受信息，a 为对应的导向矢量，F 为传入 X 的信号，W 为传入 X 的噪声。

取出语音信号中的一个通道，以便于进行语音活动检测，对取出的语音信号进行预加重，预加重是在语音信号处理之前对语音信号进行处理之前，对语音信号进行高通滤波，突出高频部分，减少低频部分对处理的影响，提高信号信噪比。预加重的原理是通过对语音信号进行一阶滤波，增强高频信号的能量。预加重通过以下公式来实现：

$$Y[n] = X[n] - \alpha X[n-1]$$

其中，$X[n]$ 是输入的语音信号，$Y[n]$ 是经过预加重后的输出信号，α 是预加重系数，通常取值在 0.9 到 1 之间。

对预加重过后的语音信号进行分帧处理，将连续的语音信号分成小段，以便于后面的语音信号特征提取，每一帧长度设置为 882 个点，帧移为一帧的一半为 441 个点。

在分帧过后对每一帧语音信号进行加窗处理，将每一帧信号与窗函数相乘，使信号在帧的边界处逐渐减小为零，以减少频谱泄漏。加窗的公式表示为：

$$X_W(n)=Y(n)\cdot w(n)$$

其中，$X_W(n)$是加窗后的信号，$X(n)$是原始信号，$w(n)$是加窗函数。

对分帧加窗后的语音信号进行快速傅里叶变换将信号时域转化成频域，以便于更好的观察和处理。

对进行分帧过后的语音信号计算每一帧的语音信号短时能量，得到每一帧的语音信号能量。短时能量的计算公式为：

$$E=\sum_{i=0}^{N-1}x(i)^2$$

Mel 频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)是一种用于语音信号处理和语音识别的特征提取方法，是一种模拟人耳的听觉系统对于语音信号频率的感应得到的特征。将频谱图转换为梅尔频率域(Mel Frequency Domain)，通常使用梅尔滤波器组对频谱图进行滤波。在梅尔频率域上定义一组 Mel 三角窗，每个 Mel 三角窗对应一个梅尔滤波器的频率响应。将每个 Mel 三角窗与梅尔频率域上的频谱图进行卷积，得到每个频率范围内的能量信息。

通过梅尔频率设置三角窗滤波器，设置 24 个三角窗，对应 24 个梅尔频率段，其中中心频率点响应值为 1，两端频率点响应值为 0。

对每一帧信号的能量谱通过三角窗滤波器计算出语音信号的 mel 特征。

$$Y_t(m)=\sum_{k=1}^{N}H_M(k)\left|X_t(k)\right|^2$$

其中$H_M(k)$为三角窗滤波器。

对每一帧的 mel 特征进行对数能量运算，然后再进行离散余弦变换(Discrete Cosine Transform, DCT)，得到最终的 MFCC 系数。离散余弦变换的计算公式为：

$$C_{dct}(k)=\sum_{n=0}^{N-1}C(n)\cos\left(\frac{\pi}{N}k\left(n+\frac{1}{2}\right)\right)$$

3. 基于语音活动检测的阵列信号测向算法

3.1. 自适应的语音活动检测算法

提取语音信号前十帧 MFCC 特征，计算前十帧平均值记作噪音特征，通过计算语音信号每一帧的 MFCC 特征与噪音特征的欧氏距离记作欧式距离。欧式距离计算公式为：

$$X_{mfcc}=\sqrt{\sum_{i=0}^{N}x(i)^2-y(i)^2}$$

对欧式距离进行归一化，以便于更好的和能量进行加权求和计算，欧式距离归一化公式为：

$$X_0=\frac{X_{mfcc}}{X_{max}}$$

其中$X_{\max}$为预估的X_{mfcc}中的最大值。

将语音信号短时能量进行归一化，语音信号短时能量进行归一化公式为：

$$E_0=\frac{E}{E_{\max}}$$

其中 $E_{\max}$ 为预估的 E 中的最大值。

对归一化欧式距离和归一化语音信号短时能量进行加权求和得出多特征权值曲线，通过多特征权值曲线和每一次曲线的变化大小设置自适应阈值对其进行语音活动检测。自适应阈值计算公式为：

$$W = 0.7X_0 + 0.3E_0$$

$$T = P^*\left(0.5W_{\max}\right) + \left(1-P\right)*W$$

通过多特征权值曲线进行语音端点检测。在检测结束后对检测结果进行修正，将检测为非语音活动的片段且片段长度小于 10 帧的重新转换为语音部分，以便于减少语音间歇造成的语音片段分散，使检测结果看起来更加完整，记录检测结果为，其中语音段为 $X_{vad}=1$，噪音段 $X_{vad}=0$。

3.2. MUSIC 阵列信号测向算法

MUSIC 算法利用了信号源信号子空间与噪声信号子空间的正交性，通过特征值分解得到的空间谱估计可以准确估计信号源的方向。

获取语音活动检测的结果中语音段的第一段，并进行阵列信号测向，记录立体阵坐标，对语音信号进行分解，分解公式为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \left[a\left(\theta_1\right)a\left(\theta_2\right)\cdots a\left(\theta_D\right)\right]\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_D \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix}$$

$$X = AF + W$$

其中 X 为阵列传声器接受信息，a 为对应的导向矢量，F 为传入 X 的信号，W 为传入 X 的噪声。

接收到的信号协方差矩阵公式表示为：

$$\begin{aligned} R &= E\left[XX^H\right] \\ &= AE\left[SS^H\right]A^H + \sigma^2 I \\ &= AR_sA^H + \sigma^2 I \end{aligned}$$

对信号协方差矩阵进行特征值分解，特征值分解公式为：

$$\begin{aligned} R &= \sum_{i=1}^{M}\lambda_i e_i e_i^H \\ &= e_s\Lambda_s e_s^H + e_n\Lambda_n e_n^H \end{aligned}$$

分解完的特征值由大到小进行排列，其对应的特征向量分别组成信号子空间 $E_s=\left[e_1,\cdots,e_p\right]$ 和噪音子空间 $E_n=\left[e_{p+1},\cdots,e_M\right]$。

MUSIC 算法利用了阵列方向矢量在噪声空间中的投影特性。由于受到噪声干扰、协方差计算误差等因素的影响，使得信号源对应的阵列导向矢量在噪声子空间的投影很小，而噪声信号对应的导向矢量投影较大。因此，当将投影结果倒置后，我们会在信号源方向得到一个尖峰。这样得到的空间谱估计可以准确地指示信号源的方向[12]。其公式为：

$$P_{music}\left(\theta\right) = \frac{1}{a^H\left(\theta\right)E_nE_n^Ha\left(\theta\right)}$$

其中 $P_{music}\left(\theta\right)$ 的几个峰值为信号的波达方向。

4. 实验结果

4.1. 语音信号采集

语音信号来源于自设实验，第一段语音信号采集实验由一段人声语音信号和城市噪声语音信号分别各播放一分钟再同时播放一分钟所得，第二段语音信号采集实验由一段人声语音信号和坦克噪声语音信号分别各播放一分钟再同时播放一分钟所得，本次采集采用 4 元立体阵列，4 个阵列的坐标标定如下表 1 所示，传声器 1、2、3 号在水平面上，传声器 4 号在垂直杆上，测试场景如图 1 所示。

Table 1. Coordinates of the elemental stereoscopic array
表 1. 元立体阵坐标

传声器编号	坐标	实测值(mm)
1	x	406.1363
	y	231.1628
	z	178.4329
2	x	−439.7724
	y	222.4280
	z	176.5127
3	x	−17.0835
	y	−502.2046
	z	168.5859
4	x	−13.7957
	y	−14.8519
	z	261.8348

Figure 1. Test scene diagram
图 1. 测试场景图

4.2. 语音信号活动检测结果

人声城市噪声音频分为三段，第一段是人声语音信号，第二段是城市噪声信号，第三段是语音信号和城市噪声混合部分。四个图表分别为语音信号的时域图像，语音信号的归一化能量特征曲线图像，语音信号的归一化 MFCC 特征曲线图像和语音信号多特征权值曲线图像。

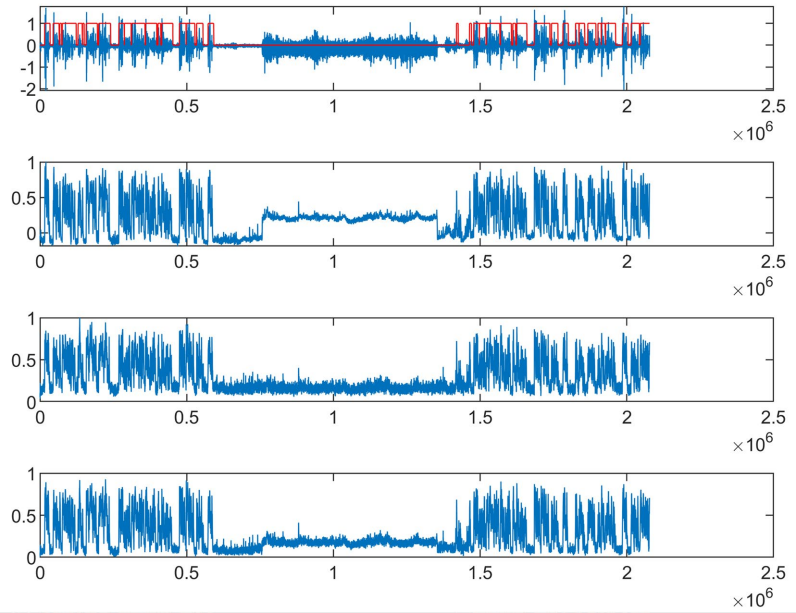

Figure 2. Results of human voices and urban noise
图 2. 人声城市噪声结果

从图 2 中可以看出改进的多特征自适应语音活动检测结果，在能量特征曲线和 MFCC 特征曲线能够很明显地看出人声部分和城市噪声部分区别很大，加权求和过后的特征区别也很大，对语音活动的检测结果很清晰准确。

对语音活动检测中人声部分进行测向结果为 30 度，如图 3 所示：

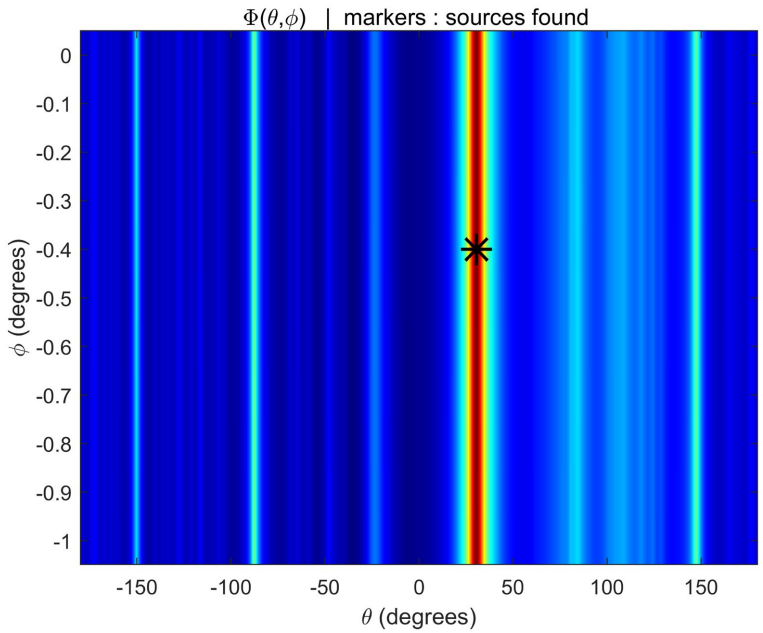

Figure 3. Directional results of the human voice component in urban noise
图 3. 人声城市噪声的人声部分测向结果

对语音活动检测中城市噪声部分进行测向，结果为 110 度，如图 4 所示：

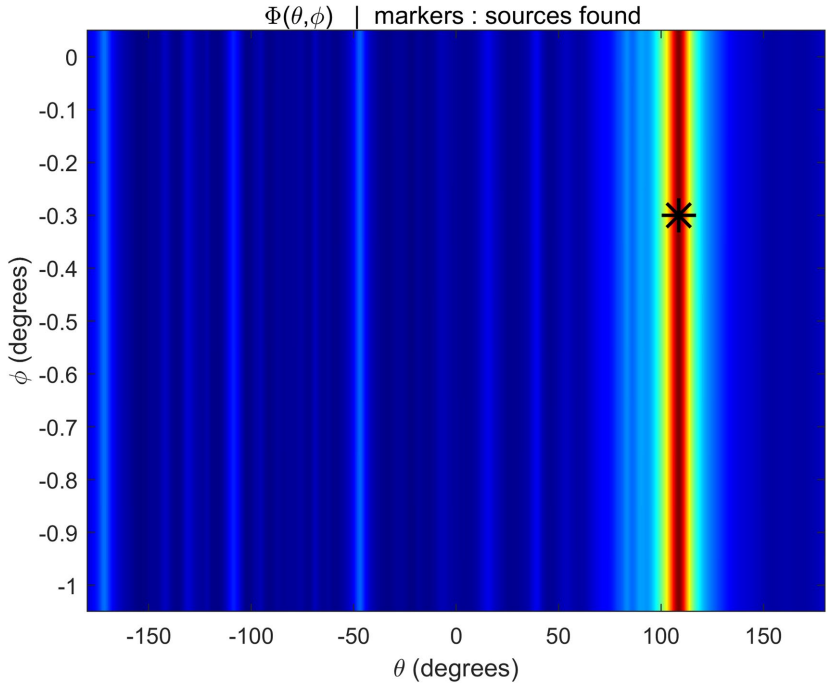

Figure 4. Directional results of the urban noise component in human voices and urban noise
图 4. 人声城市噪声的城市噪声部分测向结果

在人声与城市噪声混合的部分我们需要对人声进行检测，其检测结果为 30 度。与对人声单独检测的结果一致。如图 5 所示：

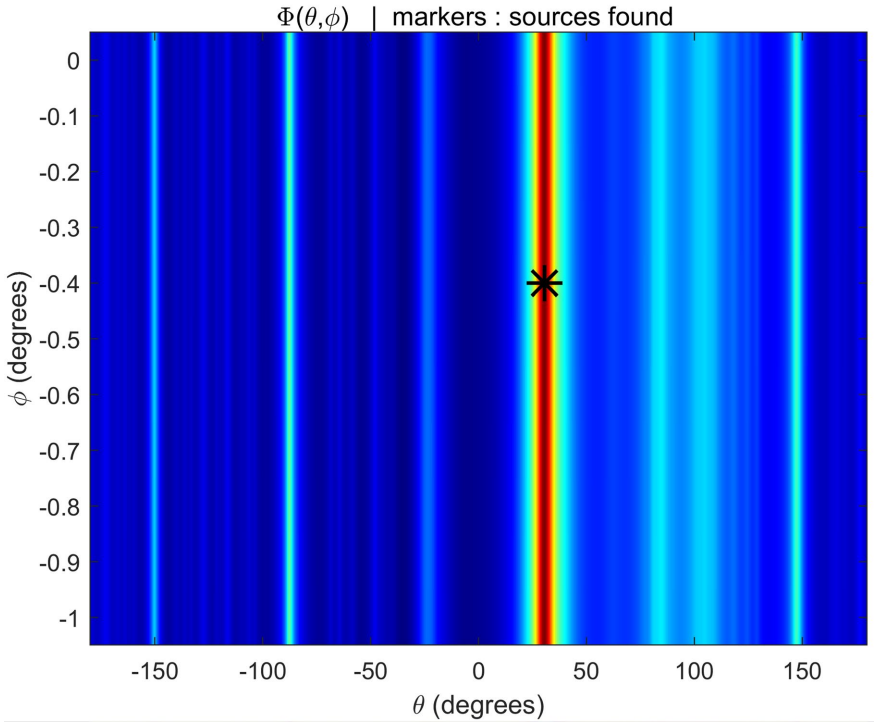

Figure 5. Directional results of the mixed component of human voice and urban noise
图 5. 人声城市噪声的人声城市噪声混合部分测向结果

人声坦克噪声音频分为三段，第一段是人声语音信号，第二段是坦克噪声信号，第三段是语音信号和坦克噪声混合部分。四个图表分别为语音信号的时域图像，语音信号的归一化能量特征曲线图像，语音信号的归一化 MFCC 特征曲线图像和语音信号多特征权值曲线图像。

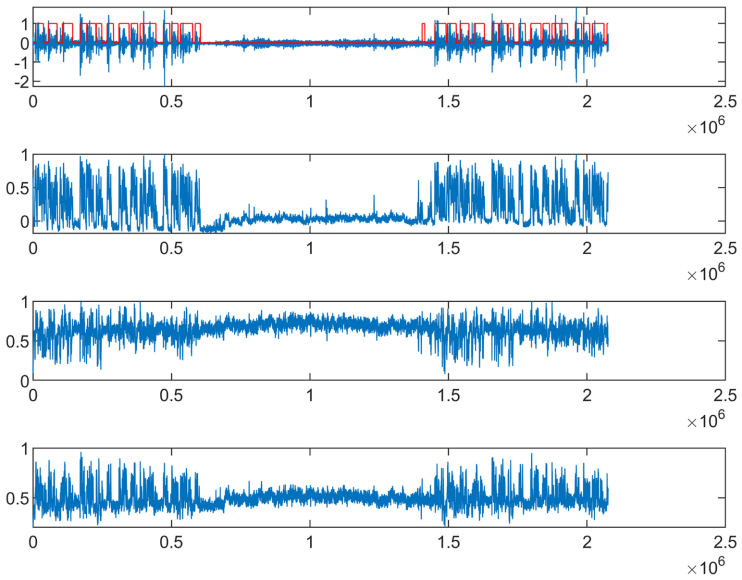

Figure 6. Results of human voice and tank noise
图 6. 人声坦克噪声结果

从图 6 中可以看出改进的多特征自适应语音活动检测结果，在能量特征曲线和 MFCC 特征曲线能够很明显地看出人声部分和坦克噪声部分区别很大，加权求和过后的特征区别也很大，对语音活动的检测结果很清晰准确。

对语音活动检测中人声部分进行测向结果为 30 度，如图 7 所示：

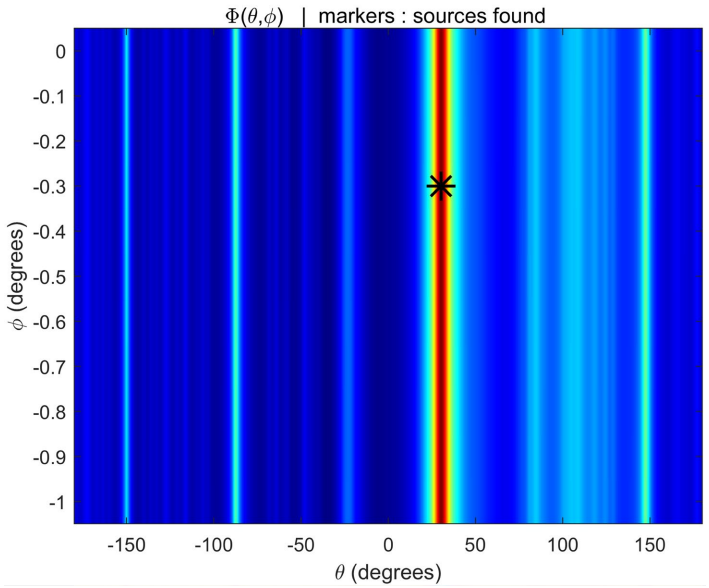

Figure 7. Directional results of the human voice component in tank noise
图 7. 人声坦克噪声的人声部分测向结果

对语音活动检测中城市噪声部分进行测向结果为 110 度，如图 8 所示：

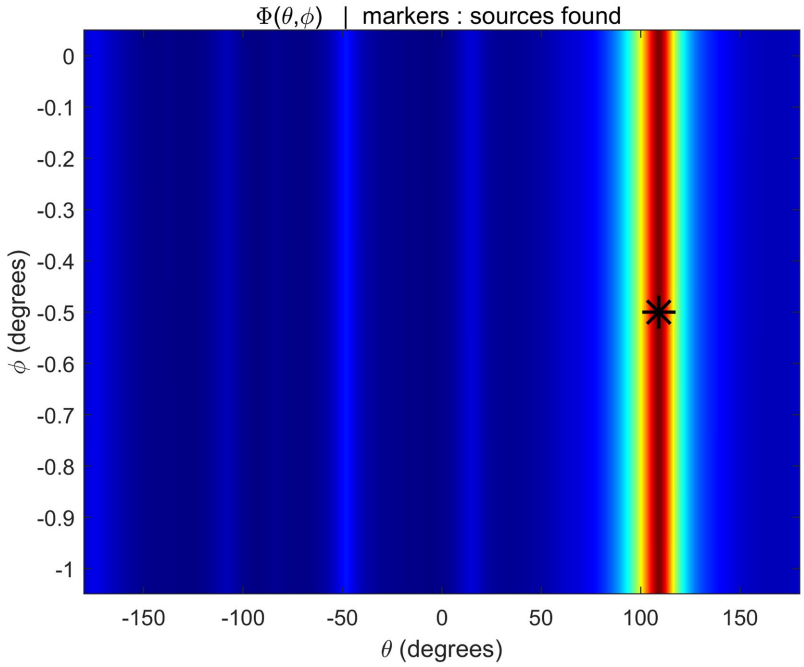

Figure 8. Directional results of the tank noise component in human voice and tank noise
图 8. 人声坦克噪声的坦克噪声部分测向结果

在人声与城市噪声混合的部分我们需要对人声进行检测，其检测结果为 30 度。与对人声单独检测的结果一致。如图 9 所示：

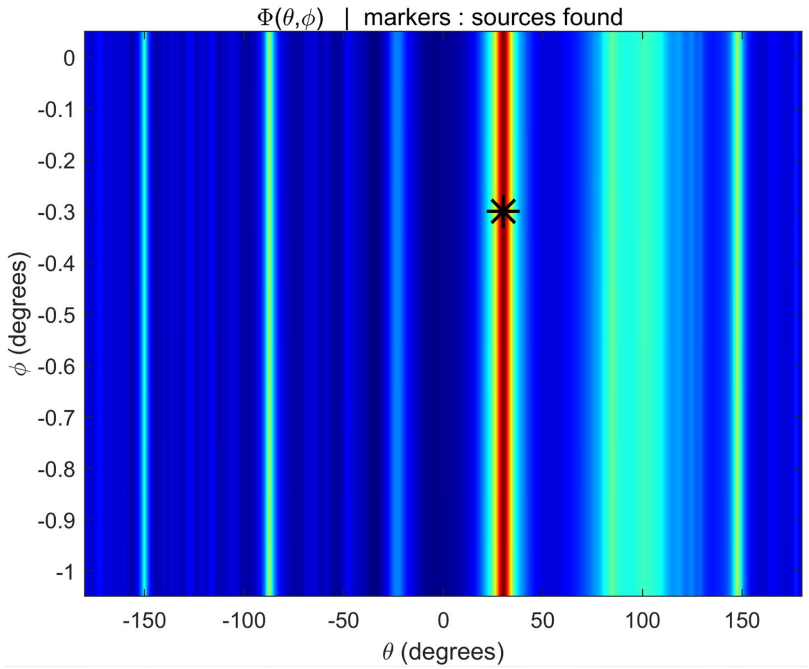

Figure 9. Directional results of the mixed component of human voice and tank noise
图 9. 人声坦克噪声的人声坦克噪声混合部分测向结果

4.3. 本文算法性能对比实验结果

选取 TIMIT 数据集中的随机挑选 20 条语音信号，进行性能对比实验，对实验用到的语音信号进行手动标注，以提高实验结果的准确性，选取两种经典算法能量过零率双门限阈值和能量 MFCC 双门限阈值，一种文献 9 中的 mvad 算法和本文的自适应阈值算法进行对比，对实验语音信号添加白噪声，使语音信号信噪比分别为−5, 0, 5, 10, 20，用每种算法分别对 20 条信号进行检测，得到结果计算出平均正确率结果如图 10 所示：

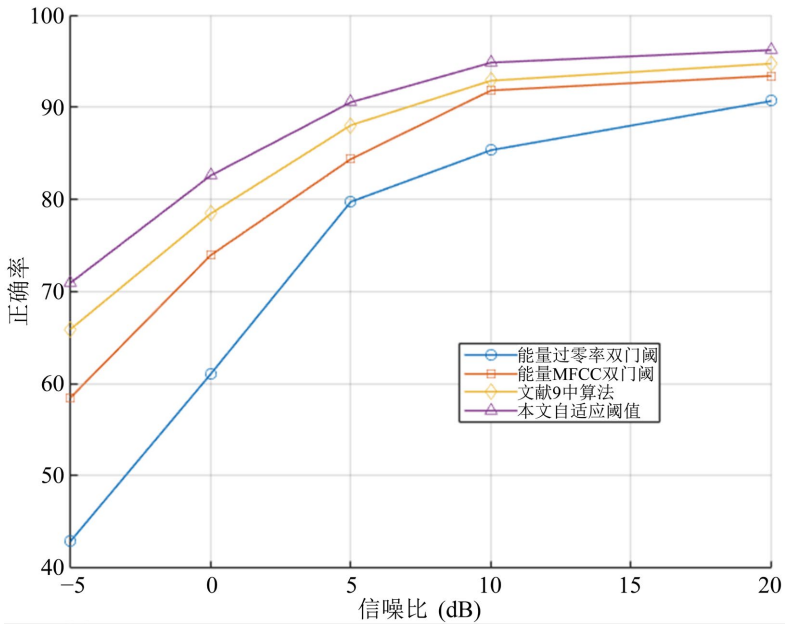

Figure 10. Comparative experimental results
图 10. 对比实验结果

由图 10 可知，本文中的自适应算法正确率相对于其他几种算法有所提升，在信噪比为−5 的情况下正确率仍能达到 70%。

基金项目

本研究得到以下两个项目支持：杜云飞，北京印刷学院基础教育学院，北京 102600，项目：北京市教育委员会科技一般项目(KM202110015001)；北京印刷学院重点教学改革项目——工程认证背景下的工科数学教学改革对大学生创新思维与创业能力培养的研究与实践。

参考文献

[1] 伦向敏, 王乃英. 基于特定频率声源的实时定位系统研究[J]. 仪表技术, 2014(8): 32-34, 37.

[2] 刘思伟, 吕海波, 慕德俊. 基于 G.729 的自适应实时语音活动检测方法研究[J]. 计算机工程与应用, 2007, 43(34): 57-60.

[3] 雷静, 何培宇, 徐自励. 低信噪比下多参数融合的自适应语音端点检测[J]. 信号处理, 2020, 36(8): 1205-1211.

[4] 赵新燕, 王炼红, 彭林哲. 基于自适应倒谱距离的强噪声语音端点检测[J]. 计算机科学, 2015, 42(9): 83-85, 117.

[5] Bao, X. and Zhu, J. (2012) A Novel Voice Activity Detection Based on Phoneme Recognition Using Statistical Model. *EURASIP Journal on Audio, Speech, and Music Processing*, **2012**, Article No. 1. https://doi.org/10.1186/1687-4722-2012-1

[6] Kang, T.G. and Kim, N.S. (2016) DNN-Based Voice Activity Detection with Multi-Task Learning. *IEICE Transactions on Information and Systems*, **99**, 550-553. https://doi.org/10.1587/transinf.2015edl8168

[7] 黄毅伟. 基于分布式传声器网络的声源定位研究[D]: [博士学位论文]. 北京: 中国科学院声学研究所, 2021.

[8] Catic, J., Dau, T., Buchholz, J. and Gran, F. (2010) The Effect of a Voice Activity Detector on the Speech Enhancement Performance of the Binaural Multichannel Wiener Filter. *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, **2010**, Article ID: 840294. https://doi.org/10.1186/1687-4722-2010-840294

[9] Zhu, Z., Zhang, L., Pei, K. and Chen, S. (2023) A Robust and Lightweight Voice Activity Detection Algorithm for Speech Enhancement at Low Signal-to-Noise Ratio. *Digital Signal Processing*, **141**, Article ID: 104151. https://doi.org/10.1016/j.dsp.2023.104151

[10] Kucuk, A., Ganguly, A., Hao, Y. and Panahi, I.M.S. (2019) Real-Time Convolutional Neural Network-Based Speech Source Localization on Smartphone. *IEEE Access*, **7**, 169969-169978. https://doi.org/10.1109/access.2019.2955049

[11] Varzandeh, R., Doclo, S. and Hohmann, V. (2024) Speech-aware Binaural DOA Estimation Utilizing Periodicity and Spatial Features in Convolutional Neural Networks. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **32**, 1198-1213. https://doi.org/10.1109/taslp.2024.3356987

[12] 张远驰, 胡进. 一种基于 MUSIC 算法的宽带信号 DOA 估计[J]. 电声技术, 2023, 47(10): 97-99.