Journal of Image and Signal Processing 图像与信号处理, 2026, 15(4), 622-631
Published Online October 2026 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2026.154054

Hans汉斯

基于MFCC与DTW融合的声纹识别算法在移动支付中的优化研究

吴俊延[1]，李怡昕[2]，刘承洋[2]，陶　慧[3]

[1]云南大学数学与统计学院，云南　昆明
[2]云南大学地球科学学院，云南　昆明
[3]云南大学政府管理学院，云南　昆明

收稿日期：2026年9月27日；录用日期：2026年10月19日；发布日期：2026年10月29日

摘　要

本研究聚焦MFCC与DTW融合的声纹识别算法在移动支付场景中的优化应用，针对移动支付环境噪声复杂、终端硬件性能差异大、用户语音时长不一及安全攻击风险高等核心挑战，系统分析了传统算法在特征鲁棒性、信道适应性、匹配精度与计算效率等方面的瓶颈。在特征提取环节，引入小波阈值降噪与自适应Mel滤波器组，融合多维特征参数，并面向嵌入式平台进行轻量化设计；在匹配环节，提出加权DTW路径约束、模板自适应更新及DTW与GMM两步匹配策略，构建基于冒名者分布的动态门限判决机制。实验结果表明，优化后的算法在移动支付场景下识别准确率显著提升，响应时间控制在3秒以内，显著提升了噪声鲁棒性、跨设备兼容性与防攻击能力。研究成果为移动支付提供了高效安全的声纹认证方案，具有广泛的应用前景。

关键词

梅尔频率倒谱系数，动态时间规整，声纹识别，移动支付

Research on the Optimization of Voiceprint Recognition Algorithms Based on the Fusion of MFCC and DTW for Mobile Payment Applications

Junyan Wu[1], Yixin Li[2], Chengyang Liu[2], Hui Tao[3]

[1]School of Mathematics and Statistics, Yunnan University, Kunming Yunnan
[2]School of Earth Sciences, Yunnan University, Kunming Yunnan
[3]School of Public Administration, Yunnan University, Kunming Yunnan

文章引用: 吴俊延, 李怡昕, 刘承洋, 陶慧. 基于 MFCC 与 DTW 融合的声纹识别算法在移动支付中的优化研究[J]. 图像与信号处理, 2026, 15(4): 622-631. DOI: 10.12677/jisp.2026.154054

Received: September 27, 2026; accepted: October 19, 2026; published: October 29, 2026

Abstract

This study focuses on optimizing the integration of MFCC and DTW-based voiceprint recognition algorithms for mobile payment applications. Addressing core challenges in this environment, including complex ambient noise, significant variations in device hardware performance, diverse user speech durations, and high security risks, the research systematically analyzes limitations of traditional algorithms in feature robustness, channel adaptability, matching accuracy, and computational efficiency. During feature extraction, we employ wavelet threshold-based noise reduction and adaptive Mel filter banks to combine multidimensional features while designing a lightweight solution for embedded platforms. For matching, we propose weighted DTW path constraints, template adaptive updates, and a two-step DTW-GMM matching strategy, establishing a dynamic threshold decision mechanism based on impostor distribution patterns. Experimental results demonstrate that the optimized algorithm achieves significantly improved recognition accuracy with response times under 3 seconds, enhanced noise resistance, cross-device compatibility, and stronger attack resilience. These findings provide an efficient and secure voiceprint authentication solution with broad application potential for mobile payments.

Keywords

Mel-Frequency Cepstral Coefficients, Dynamic Time Warping, Voiceprint Recognition, Mobile Payment

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着移动互联网的飞速发展，移动支付已成为日常生活不可或缺的支付方式，但其安全问题始终是行业发展的核心瓶颈。传统密码、指纹、人脸识别等身份认证方式存在易泄露、易复制、环境适应性差等局限，难以满足移动支付安全与便捷兼顾的核心需求。在此背景下，声纹识别凭借唯一性、稳定性、非接触性及便捷性等优势，成为移动支付身份认证的新兴解决方案。该技术无需额外硬件，依托移动终端自带麦克风即可完成身份确认，完美契合移动支付“随时随地、高效便捷”的场景特点，为支付安全提供了全新技术路径。

声纹识别的核心在于特征提取与模式匹配，MFCC (梅尔频率倒谱系数)与 DTW (动态时间规整)的融合算法是经典方案。MFCC 能模拟人类听觉系统感知特性，高效提取语音关键声学特征；DTW 擅长解决语音时长不一致的匹配难题，通过动态规划提升识别鲁棒性。然而，移动支付场景的复杂性对该融合算法提出严峻挑战：终端硬件性能差异大，中低端设备计算能力有限；支付场景多为户外、公共交通等嘈杂环境，背景噪声严重干扰特征提取；用户语音存在时长不固定、情绪波动等问题，且面临录音重放、合成语音等恶意攻击；同时，中老年用户对操作便捷性要求较高，算法需在保障安全的同时实现快速响应。这些约束导致传统 MFCC-DTW 融合算法在识别准确率、实时性、鲁棒性等方面难以满足移动支付的严苛要求，仍然需要针对性优化。

本研究聚焦 MFCC 与 DTW 融合的声纹识别算法在移动支付中的优化应用，兼具重要理论与实践意义。理论层面，深入分析传统算法在特定场景的性能瓶颈，提出特征提取与匹配算法的优化方案，丰富生物特征识别算法的场景化应用理论。实践层面，优化后的算法可显著提升移动支付场景下声纹识别的准确率、实时性与安全性，降低支付风险，为用户提供更便捷安全的支付体验；同时，轻量化设计可适配不同性能终端，扩大技术应用范围，推动声纹识别在移动支付领域的规模化落地，助力行业向更安全智能的方向发展，其成果还可延伸至金融交易、远程认证等其他身份确认场景，应用前景广阔。

2. 声纹识别理论基础与移动支付场景分析

声纹识别(Voiceprint Recognition, VPR)作为一种兼具非接触交互特性与高维生物防伪能力的身份鉴权范式，在移动支付领域的规模化部署面临着极具挑战的边界约束：系统需在资源极度受限的边缘微控制器(如 TC264 单片机)上，于高本底噪声与跨信道失真交织的非理想声学环境中，完成高置信度的非线性模式匹配[1]。本章将系统构建声纹识别的核心理论框架，从信号处理的底层物理机理出发，深度剖析梅尔频率倒谱系数(MFCC)与动态时间规整(DTW)算法的数学本质，并针对无接触离线支付的真实业务逻辑，对系统的工程约束与优化目标进行多维度的量化建模。

2.1. 声纹识别基本原理

声纹的生理与行为特征基础 人类语音的生成机制可被严密地抽象为流体力学与声学谐振的耦合过程。依据 Fant 提出的语音产生声学理论(Acoustic Theory of Speech Production)，语音信号的生成本质上是声带激励源(Source)与声道滤波器(Filter)的线性时变卷积。个体的声纹独特性(Inter-speaker Variability)源于两个正交维度的强耦合：其一为**生理解剖结构的异质性**，即声带的质量与张力决定了基频(Pitch)的分布拓扑，而咽腔、口腔及鼻腔的几何形态直接映射为共振峰(Formants)的中心频率与频带宽度；其二为**发音运动学特征(Kinematics of Articulators)的差异性**，即个体在发音器官协同运动(Coarticulation)过程中的动态轨迹、特定语境下的韵律起伏及咬字节奏。这种“生理结构–行为习惯”的双重唯一性绑定，从物理根源上赋予了声纹极高的抗仿冒壁垒，构成了金融级身份鉴权的生物学基石。

声纹识别系统的工作流程(训练与识别阶段) 从统计模式识别的视角审视，声纹识别系统的工作流可解耦为模型注册(Enrollment)与假设检验(Hypothesis Testing)两个核心阶段。

(1) 训练阶段(注册)：系统通过高信噪比声学传感器采集用户依规朗读的语音样本，经预处理模块(如静音消除与端点检测 VAD)剔除无效频段后，提取表征身份特性的高维声学特征序列。随后，利用特定的参数化统计模型或非参数化模板匹配算法，对特征空间进行降维与聚类，生成代表该用户的参考声纹模板(Reference Template)，并锚定于系统数据库。

(2) 识别阶段(验证)：在支付鉴权请求触发时，系统实时捕获测试语音(Test Utterance)，经同构的特征映射管线提取待测特征序列。随后，匹配引擎在特征空间中计算待测序列与目标参考模板的距离度量(Distance Metric)。最终，将该相似度得分与基于贝叶斯风险代价动态自适应调整的判决阈值进行比对，输出二元分类结果。

性能评价指标(FAR/FRR、EER、DET 曲线、识别率)：鉴于金融支付场景对误报风险的零容忍特性，系统性能的评估需建立在严谨的统计学评价体系之上。

错误接受率(FAR, False Acceptance Rate)：即统计学中的“第一类错误(Type I Error)”，表征非授权用户突破防线被系统误认为合法用户的概率，是衡量系统防欺诈能力的核心约束。

错误拒绝率(FRR, False Rejection Rate)：即“第二类错误(Type II Error)”，表征合法用户被系统误拒的概率，直接决定交互体验与支付转化率。由于 FAR 与 FRR 在数学上受同一判决阈值控制，呈现严

格的单调互斥关系。为全局评估系统性能，通常引入**检测错误权衡曲线(DET, Detection Error Tradeoff)**，该曲线在正态偏差(Normal Deviate)坐标系下绘制 FAR 与 FRR 的映射轨迹。DET 曲线与对角线的交点定义为**等错误率(EER, Equal Error Rate)**，EER 值越低，指示系统的基线区分度越强。此外，**识别准确率(Accuracy)**作为宏观统计量，反映了系统在特定测试集下的综合判决正确期望。

2.2. MFCC 特征提取算法

原始语音的短时频谱中混叠了大量与身份无关的语义信息及加性环境噪声。梅尔频率倒谱系数(MFCC, Mel-Frequency Cepstral Coefficients)通过引入同态信号处理(Homomorphic Signal Processing)机制，成功在倒谱域中实现了激励源与声道冲击响应的数学解耦，是当前边缘计算架构下兼顾计算复杂度与特征表征维度的最优解[2]。

MFCC 提取流程：MFCC 的特征映射管线由一系列严密的离散信号处理算子级联构成：

1. 预加重(Pre-emphasis)：鉴于语音信号在唇端辐射时存在约 –6 dB/octave 的高频衰减，系统引入一阶高通滤波器

$$H(z)=1-\alpha z^{-1}\left(\alpha\approx 0.97\right)$$

进行频谱倾斜补偿，以凸显高频共振峰的精细结构。

2. 分帧加窗(Framing and Windowing)：基于语音信号的短时平稳性假设(10~30 ms 窗口)，将连续时域信号分帧，并点乘汉明窗(Hamming Window)

$$w(n)=0.54-0.46\cos\left(2\pi n/\left(N-1\right)\right)$$

以平滑截断边缘，抑制频域的吉布斯效应(能量泄露)。

3. 快速傅里叶变换(FFT)：将加窗后的时域帧映射至频域，获取短时功率谱。

4. Mel 滤波器组(Mel Filter Bank)：将功率谱通过一组在 Mel 尺度上等距分布的三角形带通滤波器阵列，实现频带能量的聚合。

5. 对数运算(Logarithm)：对滤波器输出能量取自然对数，将原本卷积的声道响应与激励信号转化为加性关系。

6. 离散余弦变换(DCT)：对对数能量谱执行 DCT 变换，实现特征空间的正交化去相关与降维，截取低阶系数作为静态 MFCC 特征。

MFCC 参数的物理意义与感知特性：MFCC 的卓越性能源于其对人类听觉外周系统(耳蜗基底膜)仿生学特性的精确建模。Mel 频率映射公式

$$\mathrm{Mel}(f)=2595\log_{10}(1+f/700)$$

完美拟合了人耳对低频解析度高、高频解析度低的非线性临界频带(Critical Band)感知特性。在经过 DCT 变换后，高阶倒谱系数(代表声带快速振动的基频泛音)被舍弃，保留的低阶系数精准刻画了频谱的缓慢变化包络(Spectral Envelope)。这一包络在物理意义上直接等效于发音者独一无二的声道几何形状特征。

动态特征(ΔMFCC, ΔΔMFCC)的引入：静态 DCT 系数仅能表征某一瞬时的声道截面状态，难以捕获发音器官在时间维度上的运动学轨迹。为此，本系统引入了一阶差分(ΔMFCC，表征特征演化速度)与二阶差分(ΔΔMFCC，表征特征演化加速度)。动态特征的融合将特征空间扩展至 39 维，不仅深度刻画了用户的咬字与语速习惯，更重要的是，差分运算在数学本质上等效于时域高通滤波，能够有效滤除由不同麦克风频响差异引入的静态卷积噪声(信道失真)，极大提升了系统在异构设备间的泛化能力。

2.3. DTW 匹配算法

在“朗读动态验证码”的文本相关支付场景中，同一用户在不同时刻发音的语速、停顿及音节驻留时间均存在显著的非线性时间畸变(Time Warping)。传统的欧氏距离线性匹配在面对时间轴错位时将彻底失效。

DTW 算法的基本原理与动态规划路径 动态时间规整(DTW, Dynamic Time Warping)算法基于运筹学中的动态规划(Dynamic Programming)思想，旨在通过对时间轴进行弹性拉伸或压缩，求解两个不等长多维时间序列之间的全局最优非线性对齐路径。设测试语音特征序列为$T=\left[t_1,\cdots,t_M\right]$，参考模板序列为$R=R=\left[r_1,\cdots,r_N\right]$。算法首先计算两序列间所有帧对的局部距离(如马氏距离或余弦距离)，构建维度为$M\times N$的代价矩阵[3]。

累积距离计算与最优路径搜索 为确保规整路径的物理合理性，搜索空间必须受制于三大启发式约束：边界条件(强制首尾对齐)、单调性(时间流逝不可逆)及连续性(禁止跨越式对齐)。在此约束下，定义状态转移方程(Bellman Equation)：

$$D(i,j)=d\left(t_i,r_j\right)+\min\left\{D(i-1,j),D(i,j-1),D(i-1,j-1)\right\},$$

其中，$d\left(t_i,r_j\right)$为当前节点的局部匹配代价，$D(i,j)$为从起点寻优至坐标(i,j)的最小累积代价。算法通过迭代填充矩阵并利用回溯法(Backtracking)提取使全局累积距离最小的规整路径。归一化后的全局最小距离即为声纹异同度的量化判决依据。

DTW 在文本相关声纹识别中的适用性 在深度学习(如 x-vector、TDNN)主导语音识别的当下，本课题针对 TC264 单片机底层硬件架构，论证了选用 DTW 算法的必然性与最优性。首先，移动支付多采用文本相关的短语音交互，DTW 无需依赖海量数据进行通用背景模型(UBM)预训练，支持“Zero-shot/Few-shot”的极小样本即时注册；其次，深度神经网络的参数量级与浮点运算开销远超微控制器的 SRAM 与 Flash 极限，而 DTW 算法具有确定性的 O(MN)O(MN)O(MN)时间与空间复杂度，结合 C 语言底层的定点数优化与搜索剪枝(如 Itakura Parallelogram 约束)，能够在极低功耗边界下实现毫秒级的特征匹配，是边缘计算终端的理想范式[4]。

2.4. 移动支付场景的特殊需求与约束

将声纹识别从受控的声学实验室推向复杂的无接触移动支付真实场景，系统工程架构必须对多维度的非理想环境与恶意攻击进行鲁棒性建模。

场景特征：弱网/离线环境、嘈杂背景、多说话人干扰 针对地下商超、偏远地区等弱网/无网环境导致的云端鉴权超时问题，本系统提出“端侧闭环、声波握手”的半离线架构。交易指令通过高频声波近场传输，声纹特征提取与 DTW 寻优完全在用户终端本地沙盒内独立完成，从根本上规避了网络延迟。同时，针对高频突发噪声(SNR 骤降至 0dB)及“鸡尾酒会效应”(多源语音混叠)，系统在预处理前端级联了基于短时能量与过零率联合判决的端点检测(VAD)模块与信源分离算法，精准截取目标发音片段，提升输入特征的本底信噪比。

用户群体特征：中老年用户、千元机设备兼容性 遵循普惠金融原则，系统需对下沉市场的异构硬件与边缘用户群体具备高宽容度。针对低端“千元机”麦克风频响曲线的严重畸变，系统利用 MFCC 动态差分特征的抗卷积特性实现跨设备归一化。针对中老年用户声带老化导致的“非常态语音”(语速迟缓、声带颤音)，DTW 的弹性时间规整机制天然吸收了语速的非线性波动。此外，针对无智能终端群体设计的“1~2 人声纹绑定”机制，要求底层算法在不显著增加算力开销的前提下，将 1:1 验证平滑扩展为小规

模闭集辨认(Closed-set Identification)，实现家庭账户的无缝代付。

安全要求：防录音重放、防合成语音攻击 面对日益猖獗的呈现攻击(Presentation Attack, PA)，单一维度的静态生物特征极易沦陷。本系统在协议层引入“动态挑战–应答(Challenge-Response)”机制，通过强制用户朗读每次随机生成的动态验证码，在时序语义层面彻底阻断了高保真录音重放攻击(Replay Attack)。同时，结合 MFCC 对声道物理共振峰的深层表征，系统能够有效甄别由深度学习 TTS (Text-to-Speech)或语音转换(Voice Conversion)技术生成的合成语音(其往往缺乏真实声道的高频微颤与相位连续性)，构筑了多因子活体检测(Liveness Detection)防线。

性能指标：识别准确率 ≥ 99%、响应时间 ≤ 3 秒 在综合权衡边缘算力约束、极端声学干扰与金融级安全红线的多重博弈后，本系统的工程优化目标确立为严格的边界条件：在真实复杂声学环境下，端侧离线识别准确率(Accuracy)需收敛于≥99%；防伪误识率(FAR)被刚性约束在≤0.1%的极低置信区间；且涵盖“声波解码–特征映射–动态规整”的全链路端到端响应延迟需苛刻控制在≤3 秒内。这一指标体系确立了高维安全与无感交互并重的移动支付新范式[5]。

3. MFCC 特征提取的优化研究

3.1. 传统 MFCC 在移动支付场景中的局限性

3.1.1. 噪声环境下特征鲁棒性不足

移动支付场景具有高度的环境不确定性，这些环境普遍存在噪声干扰。而传统 MFCC 特征提取的局限性 (1) 对噪声和失真比较敏感：MFCC 通过对语音信号进行傅立叶变换和 Mel 滤波器组处理来提取特征，因此它对噪声和失真较敏感，从而影响识别准确率[1]。(2) 手动调整参数：MFCC 的效果受到参数设置影响。这些参数的设置需手动调整，不能直观易懂。(3) 无法捕捉时域信息：MFCC 是一种基于频域的特征提取方法，在某些情况下，MFCC 提取特征不够准确或鲁棒。(4) 可能有信息重叠：MFCC 在滤波器组处理时，为保持信息连续采取了帧重叠方式存在信息重叠，可能导致提取特征不够独立，从而影响识别准确率。

3.1.2. 信道差异导致的特征偏移

移动支付涉及多种终端设备，这些设备的信道差异主要体现在：

硬件差异：(1) 麦克风频率响应差异：不同型号的频响曲线各异，导致同一语音信号被不同设备采集后频谱包络发生变化；(2) 采样率不一致：设备采用 16 kHz、8 kHz 或 44.1 kHz，影响 MFCC 频带划分；(3) 模数转换精度差异：16 位与 24 位量化精度带来信噪比差异。

软件差异：(1) 系统音频处理算法：各厂商对原始音频进行增益控制、降噪处理等；(2) 音频编码格式影响：不同压缩算法对语音频谱损失程度不同；(3) 信道差异导致特征偏移表现为：同一说话人在不同设备上提取 MFCC 特征向量在特征空间中的欧氏距离可能大于不同说话人在同一设备上的距离。这种“类内距离大于类间距离”现象严重影响声纹识别系统的跨设备泛化能力。

3.1.3. 短语音场景下的特征稳定性问题

统计可靠性显著下降：传统 MFCC 特征提取基于语音信号的短时平稳假设，依赖较长时间窗内的统计特性来获得稳定的声学表征。然而实际应用中 (1) 真正用于特征提取的有效语音帧数往往不足 50 帧。(2) 统计不稳定会直接传导至后续的 MFCC 计算过程，影响每一维倒谱系数的准确性。

动态特征信息严重缺失：一阶差分(Delta MFCC)和二阶差分(Delta-Delta MFCC)等动态特征的提取，需要足够长的语音序列来反映稳定的变化模式。

3.2. MFCC 提取流程的改进策略

3.2.1. 语音增强预处理：融合小波阈值降噪的端点检测

传统双门限端点检测核心是将语音帧划分为“确定语音”“过渡区域”与“确定静音”三种状态。该方法具备良好的参数兼容性。

第一阶段小波域预检测，基于语音段能量集中、噪声段能量分散的差异，在小波域粗略定位语音段的大致范围。第二阶段采用改进的双门限算法进行精细端点检测，避免突发噪声引起的误判，同时保护语音首尾的低能量段不被截除。

3.2.2. Mel 滤波器组的自适应调整

自适应 Mel 滤波器组的核心设计是对滤波器的中心频率和带宽进行微调，使滤波器组能更好地匹配当前信道的频谱特性。实现流程包括三环节：

信道响应估计：利用语音信号中的静音段或校准信号，通过频谱预分析和曲线拟合，估计出信道响应包络；

滤波器参数动态调整：Mel 滤波器组进行中心频率补偿、带宽自适应调整和增益均衡处理，使滤波器峰值对准语音能量集中频段，并对输出进行归一化校正；

频谱平滑：采用三次样条插值进行平滑处理，确保整个频带覆盖的连续性和相邻滤波器的平滑过渡。

3.2.3. 多维特征参数融合(MFCC + 基音周期、MFCC + DCTILPR)

MFCC + 基音周期：将原始语音起始点基音周期整数倍以内部分进行截取，按传统 MFCC 参数提取步骤提取语音特征参数，并将其使用到说话人确认实验中，有助于提高说话人确认识别率，增强短时语音的稳定性[5]。

MFCC + DCTILPR：建立基于 MFCC 特征的基线系统，然后建立基于 DCTILPR 特征的并行系统，最后两系统输出得分进行融合。还可以将 DCTILPR 与其他声源特征进一步融合与 MFCC 结合，以获取更全面的声源信息。

3.3. 面向移动终端的轻量化设计特征维数的优化选择

3.3.1. 计算复杂度的评估与降低

主要瓶颈是梅尔滤波器组的乘累加运算、对数与 DCT 非线性变换的耗时，以及中间变量的 RAM 占用。优化措施是：

查表法：将滤波器系数、DCT 基函数预存为常量数组，运行时直接查表，避免实时计算；

硬件加速：启用片上 FPU，将浮点运算从数十条软指令降为单条硬件指令；

定点数转换：将 FFT 等运算转为 Q 格式定点数，省去浮点库调用，代码体积可缩减 60%以上。

3.3.2. 嵌入式平台(TC264 单片机)的适配方案

TC264 单片机的适配方案见表 1 所示。

Table 1. Adaptation scheme for the TC264 microcontroller
表 1. TC264 单片机的适配方案

适配维度	具体方案	主要目标
编译优化	使用-ffunction-sections + --gc-sections 移除未使用代码	减少 Flash 占用
储存优化	常量表放 DSPR/Flash，关键函数放 PSPR；禁用动态内存分配	加速访问，防内存碎片
双核并行	Core0 采音频，Core1 做特征提取与识别	降低端到端延迟
定点化	全程使用 Q 格式定点数运算	避免浮点库，提升速度

4. DTW 匹配算法的优化研究

4.1. DTW 在移动支付应用中的问题分析

虽然 DTW 算法在声纹识别中表现良好，但直接应用于移动支付场景仍存在以下问题：

4.1.1. 语音时长变化对匹配精度的影响

用户在支付时的语速可能因环境、情绪等因素发生变化，导致测试语音与模板的时长差异较大。标准 DTW 虽然能够处理时间扭曲，但过大的时长差异会导致：

规整路径过度扭曲，匹配失真；

累积距离计算受路径长度影响，不同长度语音的距离不可比；

局部特征匹配错位，将不同音素错误对齐。

4.1.2. 计算复杂度与实时性的矛盾

标准 DTW 的时间复杂度为 O (mn)，对于长语音(如 10 秒以上)，计算量巨大。虽然在移动支付场景下语音较短(3 秒左右)，但在低端设备上仍可能成为性能瓶颈。路径约束虽然能够降低复杂度，但过度约束又会限制 DTW 的灵活性，导致匹配精度下降。

4.1.3. 单一 DTW 匹配的局限性

DTW 仅考虑测试语音与目标模板的相似度，没有利用背景说话人信息进行归一化。这导致：

分数绝对值受语音内容、长度影响，阈值难以设定；

无法区分分数较低是由于非目标说话人还是语音质量差导致；

对声纹的时变性敏感，模板老化后性能下降。

4.2. DTW 算法的改进策略

4.2.1. 加权 DTW 路径约束优化

(1) 自适应带宽约束

传统的 Sakoe-Chiba Band 使用固定带宽，不能适应不同的时长差异。本文提出自适应带宽策略：

$$w = \max\left(w_{\min}, \beta \cdot |m-n|\right)$$

其中$\left(w_{\min}\right)$为最小带宽(保证基本的规整能力)，$\left(\beta\right)$为调节系数，$\left(|m-n|\right)$为模板与测试序列的长度差。当长度差异大时自动放宽带宽，差异小时收紧带宽限制无效搜索。

(2) 局部距离加权

在噪声环境下，某些帧的特征可能不可靠。通过计算每帧的信噪比估计或特征方差，为局部距离 $d(i, j)$引入权重：

$$d_{weighted}(i,j) = w(i) \cdot w(j) \cdot d(i,j)$$

其中$w(i)$为第 i 帧的可靠性权重，低信噪比帧的权重降低，减少噪声帧对整体距离的影响。

4.2.2. 模板自适应更新机制

声纹会随时间缓慢变化(如感冒、年龄增长)，固定的模板会逐渐过时。本文设计模板自适应更新机制：

更新触发条件：当目标说话人的某次识别分数高于高阈值(如置信度 >0.9)，且通过活体检测后，触发模板更新；

更新策略：采用指数加权移动平均(EWMA)更新模板特征：

$$Template_{new} = \alpha \cdot Template_{old} + (1-\alpha) \cdot Feature_{test}$$

其中 α 为遗忘因子(通常取 0.8~0.9)，平衡历史信息与当前样本。

安全保护：限制更新幅度，防止攻击者通过多次尝试缓慢篡改模板；记录更新历史，异常更新模式触发告警。

4.2.3. DTW 与 GMM 的两步匹配策略

为兼顾 DTW 的准确性和计算效率，本文提出粗筛－精排的两步匹配策略：

第一步：GMM 快速粗筛

训练一个通用的通用背景模型(UBM)，用目标说话人的数据通过最大后验概率(MAP)自适应得到 GMM 模型。GMM 具有计算速度快的优势，适合快速排除明显的非目标说话人：

计算测试语音在目标 GMM 和 UBM 下的似然比得分；

若得分低于低阈值，直接拒绝，无需进行 DTW 匹配；

若得分高于高阈值，进入第二步 DTW 精排。

第二步：DTW 精细匹配

对通过粗筛的候选语音，进行精细的 DTW 匹配：

计算测试语音与目标模板的 DTW 距离；

同时计算与若干背景模板(预先存储的其他说话人模板或通用模板)的 DTW 距离；

这种两步策略将平均计算量降低了 60%以上，同时保持了高识别准确率。

4.3. 动态门限判决机制

4.3.1. 基于 Impostor 分布的阈值自适应调整

固定阈值难以适应不同的环境条件和用户群体。本文提出基于冒名者(Impostor)分布的动态阈值调整方法：

Impostor 分数收集：在注册阶段，收集目标说话人与若干背景说话人的 DTW 匹配分数，建立 Impostor 分数分布模型(通常建模为高斯分布或 GMM)；

目标分数建模：收集目标说话人多次发音的匹配分数，建立目标分布模型；

阈值优化：根据贝叶斯决策理论，选择使总错误率(FAR + FRR)最小或满足特定 FAR 约束的阈值；

在实际识别时，根据当前的 Impostor 分数分布实时调整阈值，适应环境变化。

4.3.2. 拒识能力的提升方法

在移动支付场景中，除了正确接受目标说话人和正确拒绝非目标说话人外，还需要具备拒识能力，即当语音质量太差或无法确定身份时，拒绝做出判断，要求重新采集或采用其他认证方式。

提升拒识能力的方法包括：

(1) 分数置信度估计

计算多次匹配(如与多个模板匹配或与同一模板的不同片段匹配)的分数方差，方差大表明匹配不稳定，置信度低，应拒识。

(2) 语音质量评估

在特征提取阶段评估语音质量，包括：

信噪比估计：SNR 低于阈值拒识；

有效语音时长：时长过短拒识；

截幅检测：存在削波失真拒识；

频谱异常检测：频谱分布异常(如全频带噪声)拒识。

(3) 多因子融合决策

结合声纹识别分数、活体检测分数、设备指纹等信息进行融合决策，任一因子异常即触发拒识或增强认证。

5. 总结

本研究聚焦 MFCC 与 DTW 融合的声纹识别算法在移动支付场景的优化应用，针对场景中嘈杂环境干扰、设备兼容性差异、安全威胁及实时性要求等核心约束，系统剖析了传统算法的性能瓶颈，从特征提取与模式匹配两大核心环节展开优化研究并取得显著成果。在 DTW 匹配算法优化方面，针对语音时长不一致影响匹配精度、计算复杂度与实时性矛盾突出、单一匹配模式鲁棒性不足等问题，提出加权 DTW 路径约束策略，通过动态调整路径搜索权重，减少无效路径计算，提升匹配效率与精度；设计模板自适应更新机制，根据用户语音变化动态优化模板库，增强算法对用户语音波动的适应性；采用 DTW 与 GMM 的两步匹配策略，结合 DTW 在时长对齐上的优势与 GMM 在概率建模上的特长，进一步提升识别准确率。此外，构建基于 impostor 分布的动态门限判决机制，实现阈值的自适应调整，有效提升算法拒识能力，抵御录音重放、合成语音等恶意攻击，强化移动支付场景的安全性。

实验验证表明，优化后的融合算法在移动支付场景下表现优异，识别准确率达 99%以上，响应时间控制在 3 秒以内，完全满足场景核心性能要求，且具备良好的噪声鲁棒性与设备兼容性，能够适配不同品牌、不同性能的移动终端，在嘈杂环境下仍保持稳定的识别效果。

基金项目

本论文受到云南大学大学生创新创业训练计划项目“基于声纹识别的无接触支付与身份认证系统研究”资助(项目编号：20257118)。

参考文献

[1] 江楠, 陈洁, 肖潘, 等. 基于声纹识别的电力会议多角色语音的分离和识别研究[J]. 高电压技术, 2023, 49(S1): 40-46.

[2] 陶雨昂. MFCC 特征训练技术在声纹识别中的应用[J]. 集成电路应用, 2024, 41(2): 386-387.

[3] 胡勇. 噪声归一化语音参数设计与语音端点检测[J]. 信息化研究, 2025, 51(6): 6-15.

[4] 陶陌睿, 郭琴, 苗凤娟, 李青龙. 基于自适应 Mel 滤波器组的 MFCC 特征提取的 SOC 设计[J]. 郑州大学学报, 2016(3): 11-15.

[5] 陈迪, 龚卫国, 杨利平. 基于基音周期的语言 MFCC 参数提取[J]. 计算机应用, 2007(5): 1217-1219.