CSLT TECHNICAL REPORT-20160010 [Thursday 5th January, 2017]

基于说话人信道相关的录音重放检测若干方法探究

Lantian Li[1]
, Yixiang Chen[1]
and Dong Wang[1*]

*Correspondence: wang-dong99@mails.tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

如何防止假冒者闯入，是说话人识别研究中的重要课题之一。声音模仿、语音合成、语音转换和录音重放均是假冒者闯入说话人识别系统的若干手段。本文针对录音重放闯入开展了一系列探索。本文从信道的角度分析录音重放检测任务，提出了基于说话人信道相关的录音重放检测方法。首先，利用F-ratio准则，统计原始语音与录音重放语音在各个频带的区分性；并以此实现了频带加权和频带弯折等算法。此外，本文还提出了基于卷积神经网络的特征学习方法，提取原始语音与重放语音的区分性特征，完成录音重放检测任务。

Keywords: 录音重放检测; 频带弯折; 频带加权; 卷积神经网络

1 Introduction

说话人识别是根据语音中反映说话人生理和行为特征的语音参数，来识别待测语音说话者身份的技术。作为语音识别的一个分支，说话人识别在公安侦察、医疗诊断、电子金融业务、声纹控制等方面有着广泛的应用前景。

从安全系统应用角度出发，说话人识别系统急需解决的问题是如何防止各种假冒手段闯入说话人系统的行为。如果一个说话人识别系统能够对假冒语音进行拒绝或警告，那么就可以大大提高该系统的可应用性。根据闯入手段的不同，说话人识别系统的假冒闯入手段主要可分为以下四种 [1]：声音模仿、语音合成、语音转换和录音重放。本文将重点研究如何通过对录入系统的语音文件进行录音信道的检测。

如图1所示，若说话人A的语音被企图闯入者预先窃取录制得到，继而将录制的语音重新播放至说话人识别系统，系统通常将难以区分，从而导致闯入者成功地以说话人A的录音闯入说话人A的模型，对说话人A带来极大的安全隐患。

从信号处理的角度来看，即便假冒闯入者使用高保真的录音设备录制说话人的语音信号，由于原始语音信号在进入说话人识别系统之前必定经过一个相同的或不

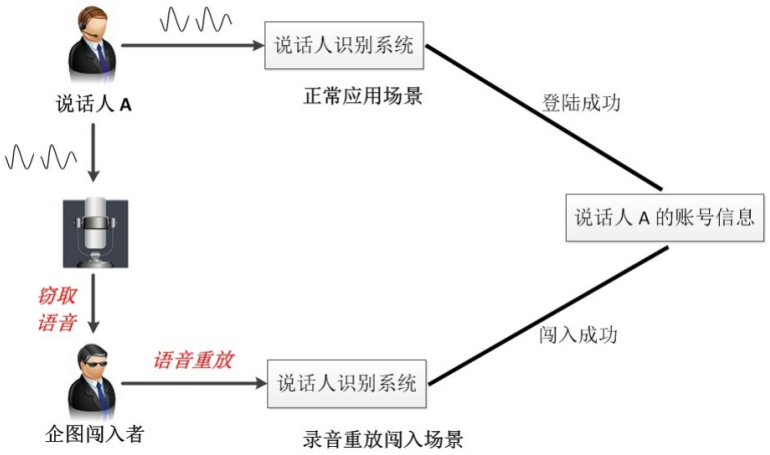

Figure 1 录音重放闯入场景

同的录音和放音系统，这两个额外的系统就会对原始语音带来额外的频谱损伤，所得到的语音必定与原始语音信号存在差异。如图 2所示，原始语音在经过录音重放后，其在信号频谱上发生了改变，这使得录音重放检测成为可能。

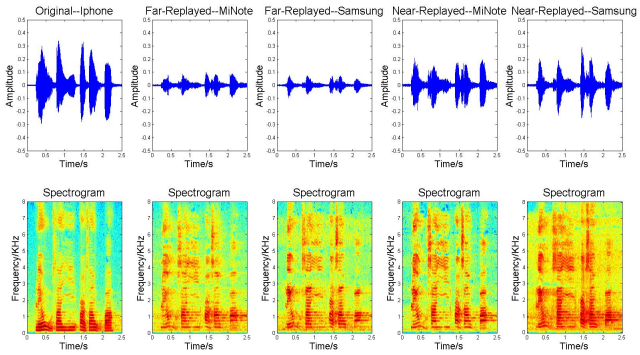

Figure 2 语音经录音重放前后的频谱图

为此，本文提出了一系列录音重放检测的方法。在现有说话人识别系统的前端增加一个录音重放检测模块，有效地检测并解决录音重放闯入的问题。

2 Related work

Villalba等人 [2]通过研究智能手机扬声器的频响特性来确定录音重放对各个频带的影响。如图 3所示，录音重放设备对各个频带的响应程度不同，也就表明语音信号经过录音重放在各个频带上的失真程度不同。为此，作者提出了一系列防录音重放鲁棒性特征，并取得了不错的检测效果。

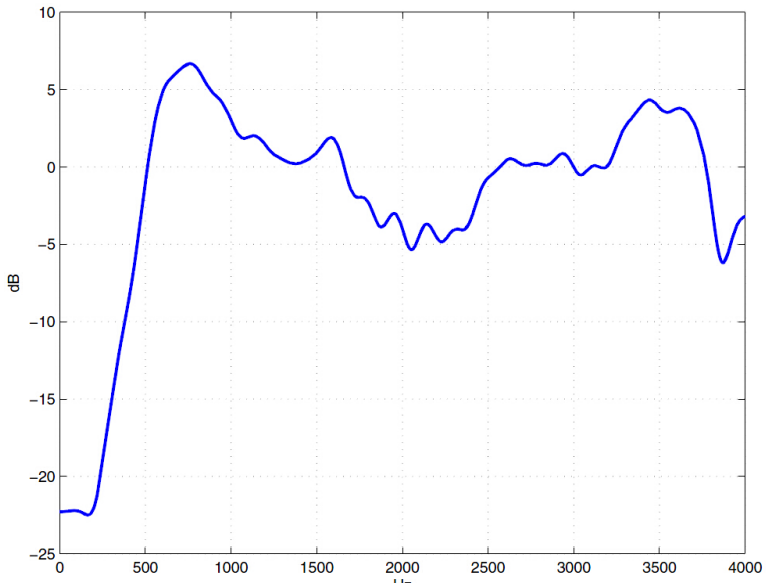

Figure 3 典型的手机扬声器频响特性

今年来，针对录音重放检测的相关研究也越来越多 [3, 4, 5]。本文结合前人工作，分别从信号域、特征域和模型域开展了一些探究。

3 Theory

如图 3所示，手机扬声器对各个频带的保真度并不相同。换言之，不同频带对录音重放的敏感度不同。为此，一个简单的思路是在特征提取过程中，依据频带的敏感度分布，对不同频带赋予不同的权重。

以现今在说话人识别中最为常用的特征MFCC为例，MFCC特征是参照人耳听觉特性，将原始频域转换为梅尔域，其转换关系如下：

$$Mel = 1127 \times \log(1 + \frac{freq}{700})$$

图4为梅尔空间与线性空间的对比图。

MFCC特征提取流程如图 5所示。针对MFCC，如何依据频带的敏感度分布，对不同频带赋予不同的权重。我们提出了两种思路：1、在三角滤波器进行频带弯折时，对不同频带上的三角滤波器乘以不同的权重，该方法称为频带加权；2、针对不同频带的敏感度，按不同比例对不同频带分配三角滤波器数，该方法称为频带弯折。

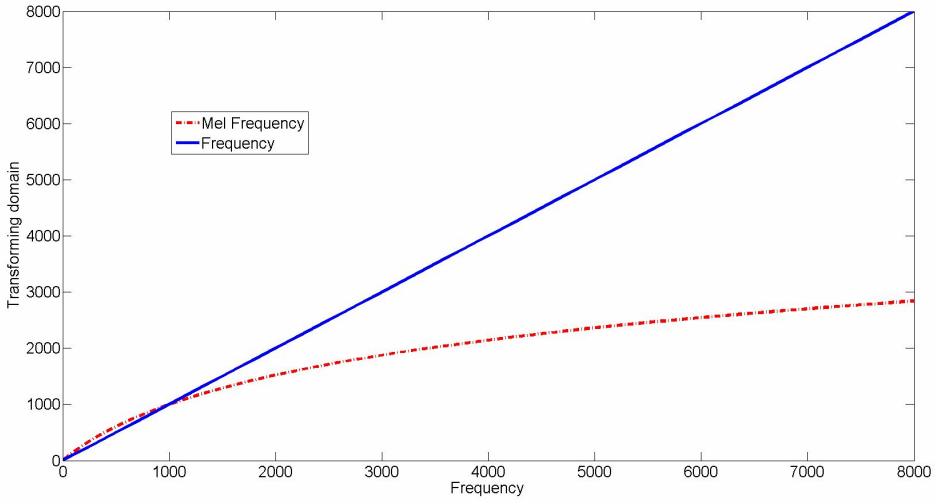

Figure 4 梅尔空间与线性空间的比较

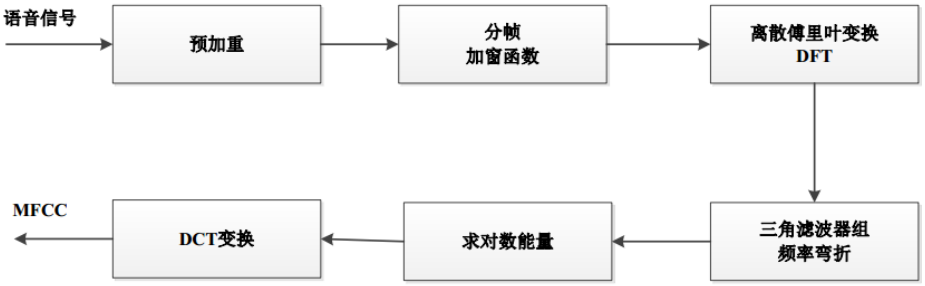

Figure 5 MFCC特征提取流程

4 Experiment

4.1 实验数据

为了更好地研究录音重放检测任务，实验数据在采集过程中使用了不同型号的六部手机。首先由这6部手机获取原始语音信号，然后采用六部手机两两互放的方式获取重放语音数据(如1号手机中的原始录音通过扬声器播放至2号手机，则2号手机录制的语音即为1号手机对2号手机的重放语音)。具体数据分布如表 1所示。

训练集由原始语音与重放语音两部份组成，其中原始语音为声密保UBM训练语音，共包含5,126 句；重放语音为从表 1中随机抽取出于声密保UBM训练语音等量的重放语音，共包含3,400句。训练集语音用于i-vector模型中的UBM和T矩阵训练，以及CNN特征学习中的网络参数训练。

本文采用基于说话人信道相关的录音重放检测方法，其特点在于预先假定每个说话人与其录音手机之间存在一一对应关系；而后系统从每个说话人的录音手机

手机ID	手机型号	录音人数	原始语音数量(句)	重放语音数量(句)
1	小米	64	1,611	7,064
2	中兴	63	2,264	6,252
3	魅族	65	1,638	8,160
4	三星	63	1,521	8,599
5	TCL	63	2,262	6,091
6	苹果	61	2,191	6,786
ALL	ALL	65	11,487	42,952

Table 1 实验数据统计表

中抽取5条该说话人原始正常语音，用于训练说话人信道模型；在测试过程中，若测试语音与说话人信道模型匹配，则判决该测试语音为原始正常语音(Target)；反之系统认定其为录音重放语音(Nontarget)。表 2给出了基于说话人信道相关的各个手机测试集。本文采用等错误率(Equal Error Rate, EER)衡量系统录音重放检测性能。

手机ID	手机型号	测试人数	Num of Target	Num of Nontarget
1	小米	64	1,291	6,936
2	中兴	63	1,949	6,070
3	魅族	65	1,313	8,160
4	三星	63	1,206	8,355
5	TCL	63	1,947	5,901
6	苹果	61	1,886	6,422
ALL	ALL	65	9,592	41,844

Table 2 实验数据统计表

下文实验中的基于性能驱动，以及基于F-ratio准则的录音重放检测都使用MFCC特征及i-vector的整体框架，为比较模型性能，使用相同实验配置，实验配置如下

语音数据的采样率	16000
每帧的时长	25
高频上界	8000
低频下界	0
提取GMM-UBM模型时高斯分布数量	1024
i-vector维度	50

4.2 基于性能驱动的录音重放检测

结合前文理论分析，我们首先采用性能驱动的方法，探究不同频带下录音重放检测的性能。图6和图7为不同频带下，在梅尔域和线性域上的录音重放检测性能。

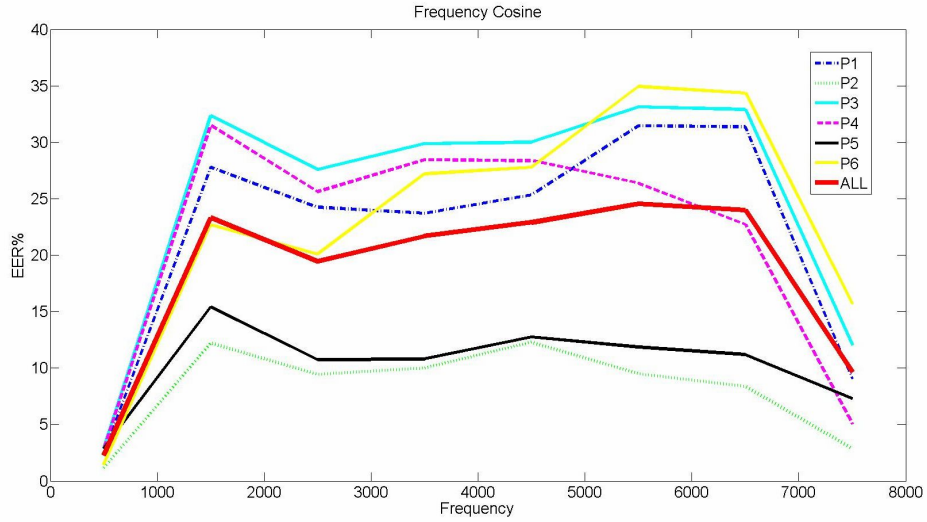

Figure 6 线性域下的各个子频带录音重放检测性能分布

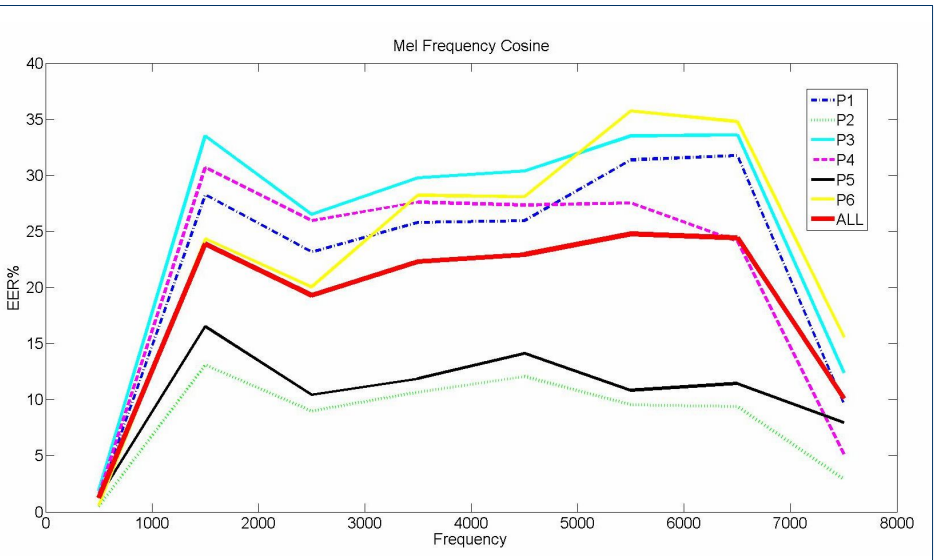

Figure 7 梅尔域下的各个子频带录音重放检测性能分布

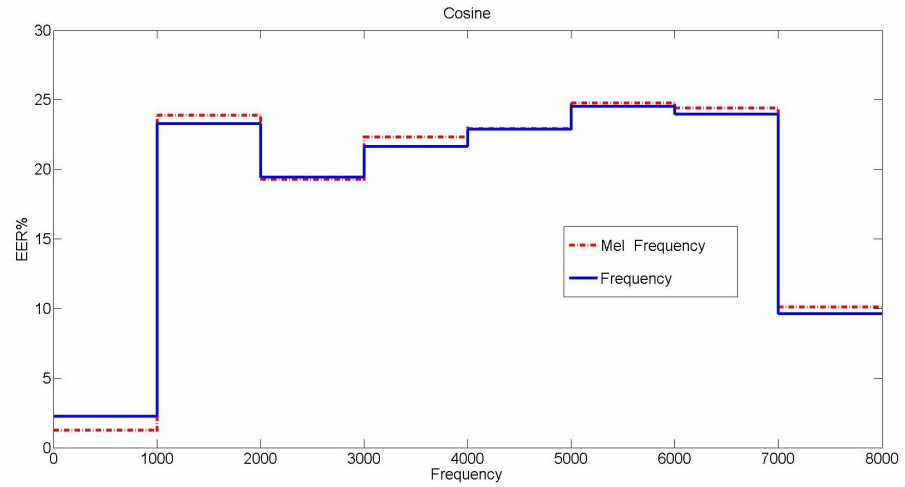

Figure 8 ALL测试集下的各个子频带检测性能分段函数图

可以看出6部手机在各个频带下的录音重放检测性能十分相似。假设每个频带的敏感度在当前频带范围内是稳定的，则可绘制一条各个子频带检测性能的分段函数图，如图8所示。

为了提升原始语音与录音重放语音之间的区分性，在特征提取过程中，我们根据各个子频带检测性能的不同，对相应子频带赋予不同的权重。显然，EER越小的频带，权重应越大，反之亦然。各个子频带的权重计算如下：

$$Weight_i = \frac{Max(EER)}{EER_i}$$

其中$Max(EER)$为录音重放检测性能最差的子频带所对应的EER值。经过权重单位化（一阶norm归一为8），梅尔域和线性域各个子频带权重如下：

F-Weight (Mel): 5.479, 0.291, 0.36, 0.311, 0.303, 0.28, 0.285, 0.687

F-Weight (Linear): 4.337, 0.424, 0.508, 0.456, 0.431, 0.402, 0.412, 1.025

在全频带范围下，对各个子频带进行频带加权，录音重放检测结果如下图9所示。如图所示，梅尔域相比于线性域的检测效果更优，我们认为其原因在采用梅尔刻度在对原始频域进行弯折时，其为低频段分配了更多的频域范围，侧面上加强了对低频段的特征提取。而从图6和图7中可发现，在0-1,000Hz频带范围内，原始语音与重放语音的区分性最大。因此，梅尔刻度上的频域弯折规律恰好加重了区分性最大的低频段，使得其检测性能相比于线性域更优。此外，在线性域上，采用该性能驱动的频带加权方法，的确使系统检测性能有所提高，表明了该基于性能驱动频带加权方法的有效性。

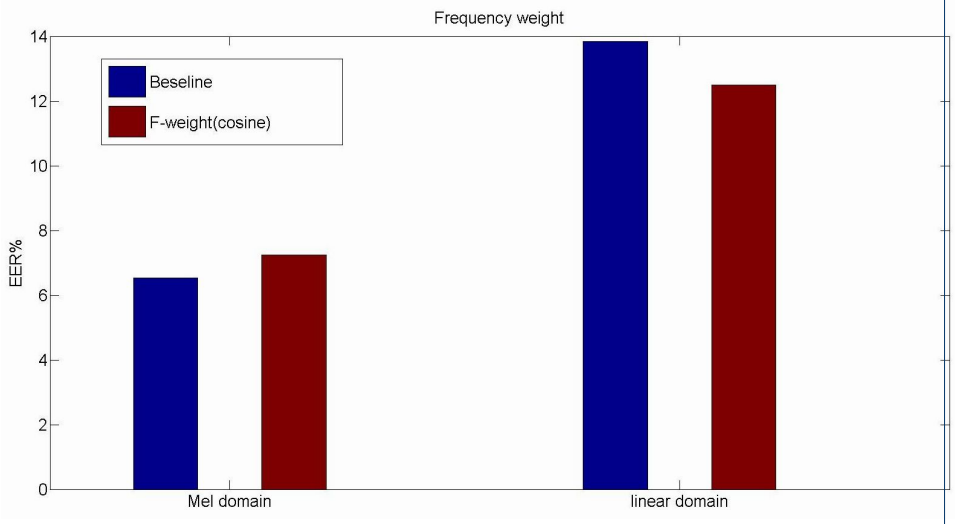

Figure 9 基于性能驱动的录音重放检测

4.3 基于F-ratio准则的录音重放检测

4.2节基于性能驱动的子频带加权方法，首先是将全频带分为若干段，进行分段性能检测；而后利用其检测结果计算不同子频带的权重；最终将权值乘在对应滤波器上。该方法虽略微有效，但其先验假设在于每个子频带的敏感度在当前频带范围内是稳定的，这使得连续的频域空间离散化，如图8所示。为此，本节讨论了另一种对频带处理的方式，基于F-ratio准则的频带加权和频带弯折。与 4.2节性能驱动的子频带加权方法的不同在于，该方法是从滤波器的角度出发，基于F-ratio 准则统计原始语音与重放语音在不同滤波器上的区分性，并计算其对应的权重。该方法的优点在于从滤波器的角度出发，统计频域空间上不同滤波器对录音重放检测的敏感度，其对频域的处理是连续的；且计算得到的F-ratio参数更加准确地反映了原始语音和重放语音在不同频带上的区分性。

基于F-ratio准则对第i个滤波器敏感度计算公式如下：

$$F-ratio_i = \frac{(\mu_{oi}-\mu_{ri})^2}{\frac{1}{N_o}\sum_{i=1}^{N_o}(a_i-\mu_{oi})\frac{1}{N_r}\sum_{j=1}^{N_r}(b_j-\mu_{ri})}$$

公式中μ_{oi}和μ_{ri}分别代表原始语音与重放语音在第i个滤波器下的特征中心，N_o和N_r代表原始语音与重放语音的数量，a_i和b_i分别表示第i个滤波器下原始语音与重放语音的特征采样。计算得到各个滤波器所对应的F-ratio后，对其进行单位化，使得其数值总和等于滤波器个数。

图 10和图 11中反映了不同滤波器下的F-ratio值变化趋势。在线性域上，取40个滤波器，通过DCT后降到20维；梅尔域取23个滤波器，通过DCT后降到13维。

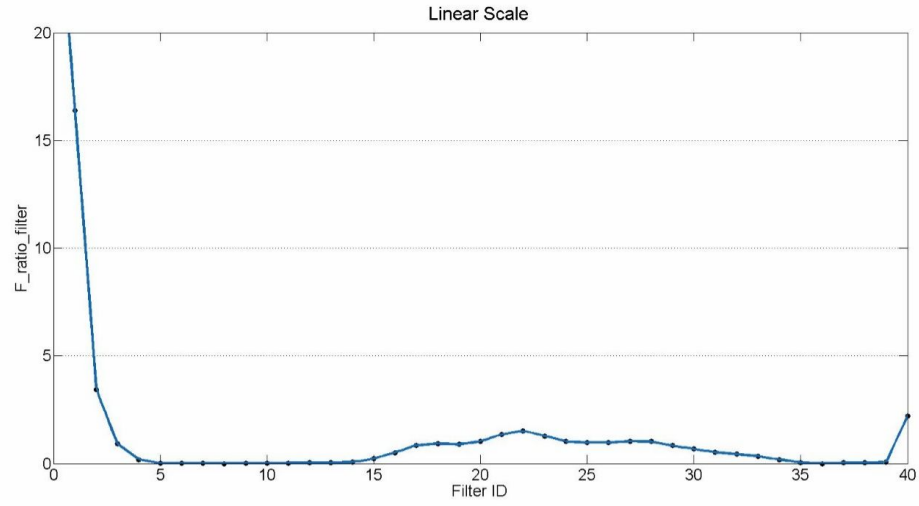

Figure 10 线性域上的F-ratio变化

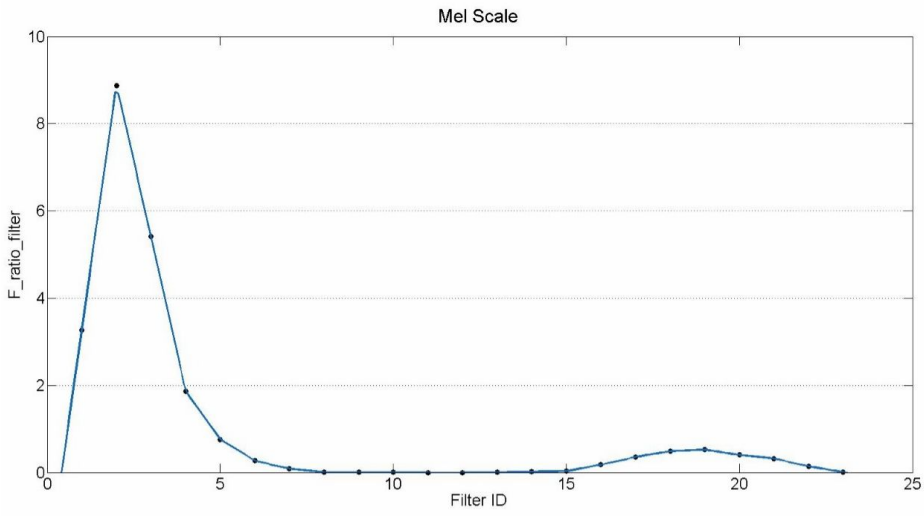

Figure 11 梅尔域上的F-ratio变化

根据图 10和图 11中各个滤波器的区分度权值，使用频带加权和频带弯折两种方式对特征提取过程进行改进，提高特征对录音重放的鲁棒性。

在实现方面来说，对于标准的MFCC特征提取过程，计算各滤波器能量过程中三角滤波器是通过对FFT点的不同加权求和来计算总能量的，以梅尔域为例计算三角滤波器能量算法参见Algorithm 1

上述算法中首先通过MelScale函数确定所求频带在梅尔域下的范围，再求出梅尔域下对应滤波器的大小，即为mel_freq_delta。之后计算每个FFT点所对应权重$weight$，从而算出滤波器能量，参数对应如图12。

Algorithm 1 计算三角滤波器能量

Input:
低频下界:low_freq, 高频上界:high_freq, 滤波器个数:num_bins, 滤波器中FFT点数:num_fft_bins;
Output:
the list of energy for bin: energy_bin;

```
1: mel_low_freq ← MelScale(low_freq)
2: mel_high_freq ← MelScale(high_freq)
3: mel_freq_delta ← (mel_high_freq − mel_low_freq)/(num_bins + 1)
4: for bin = 0 to num_bins − 1 do
5:     left_mel ← mel_low_freq + bin ∗ mel_freq_delta
6:     center_mel ← mel_low_freq + (bin + 1) ∗ mel_freq_delta
7:     right_mel ← mel_low_freq + (bin + 2) ∗ mel_freq_delta
8:     for i = 0 to num_fft_bins do
9:         freq ← fft_bin_width ∗ i
10:        mel ← MelScale(freq)
11:        if mel > left_mel and mel < right_mel then
12:            if mel <= center_mel then
13:                weight ← (mel − left_mel)/(center_mel − left_mel)
14:            else
15:                weight ← (right_mel − mel)/(right_mel − center_mel)
16:            end if
17:            weight_list append weight
18:        end if
19:    end for
20:    energy ← Calculate Energy by weight_list and label of bin
21:    energy_bin append energy
22: end for
23: return energy_bin
```

以基于F-ratio准则为例，频带方法对各个滤波器进行加权，需要将滤波器权重对上述代码中的$weight$再次加权，如图13。

即计算权值时再乘以录音重放任务下滤波器的权重，改进Algorethm 1中第17行代码为：

weight_list append $weight * filter\_weight$

另一种改进方式频带弯折方法需要改变MelSacale函数，从而改变频带映射后的空间，新的Sacale函数参见Algorithm 2

Algorithm 2 经过频带弯折后新的尺度函数Scale()

Input:
freq, list of filter_weight: weight, mel_freq_delta;
Output:
The frequency of the new scales;

```
1: mel ← 1127 ∗ logf(1 + freq/700)
2: num_parts ← int(mel/delta)
3: for i = 0 to num_parts − 1 do
4:     warping_freq ← warpint_freq + mel_freq_delta ∗ weight[i]
5: end for
6: surplus ← mel − num_parts ∗ mel_freq_delta
7: return warping_freq + surplus ∗ weight[num_parts]
```

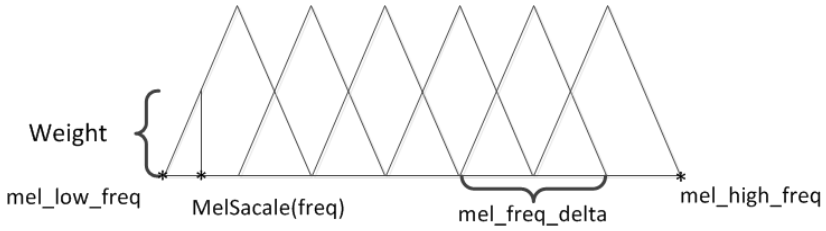

Figure 12 滤波器能量计算示意

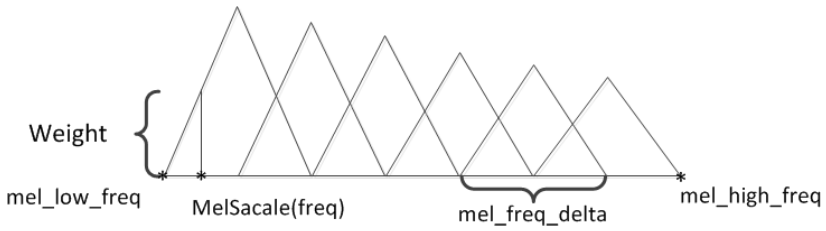

Figure 13 加权后滤波器能量

Sacale函数输入变量weight指的是对相应滤波器对应频带的弯折程度，函数先将线性域映射到梅尔域，再按照弯折程度进行弯折，如图14，从而达到对各个滤波器根据其对录音重放任务敏感性的不同而进行调整。

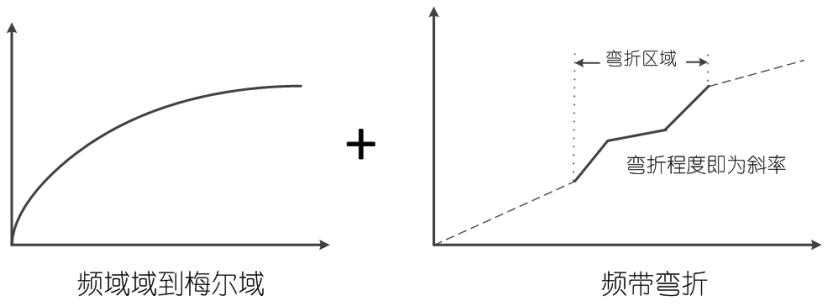

Figure 14 梅尔尺度后的频带弯折

不同改进算法对录音重放检测任务的结果如表 3所示。从表中可以看出，基于F-ratio的频带加权和频带弯折的检测性能均好于基线系统。此外，对于频带弯折方法均好于频带加权，这是因为频带加权并未改变滤波器的分布，仅是在不同滤波器上赋予不同的权值。而频带弯折则是将频域空间的滤波器进行了重新分配，充分利用各个滤波器，对区分性大的频带分配更多滤波器，而区分性小的频带分配更少甚至不分配滤波器。

4.4 基于深度卷积神经网络(CNN)的录音重放检测

	Mel	Linear
Baseline	9.83	17.30
Weighting	4.49	15.43
Warping	0.66	1.61

Table 3 基于F-ratio准则的频带加权与弯折(EER%)

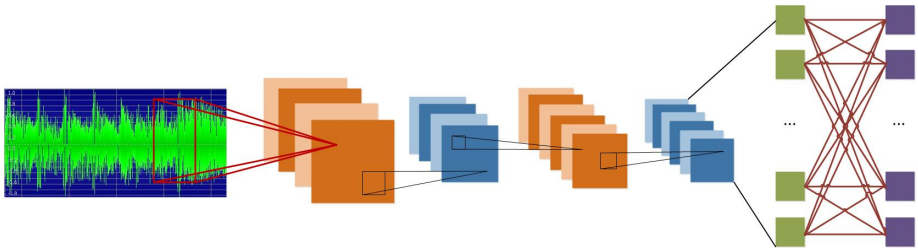

Figure 15 基于CNN的特征学习示意图

结果如下：

	C1	C2
SGD	0.131	0.157
NG-SGD	0.042	0.042

Table 4 基于CNN区分性特征学习的录音重放检测结果

注：C1为基于说话人信道相关的录音重放检测，C2为录音重放检测(无说话人信道模型)。

5 Conclusions

总结前文所述的各个录音重放检测方法，其性能对比如表5所示。可以看出，基于F-ratio准则的频带加权和频带弯折算法均提升了录音重放检测的性能，且频带弯折算法效果更优。此外，基于卷积神经网络，通过数据驱动迭代训练的方式，自动地学习出原始语音与重放语音的区分性特征，并取得了不俗的检测效果。

该工作的相关代码已开源，详见http://lilt.cslt.org/codes/replay_src.zip。

Method	EER(%)
Linear Baseline	17.30
Linear Weighting	15.43
Mel Baseline	9.83
Mel Weighting	4.49
Linear Warping	1.61
Mel Warping	0.66
CNN SGD	0.131
CNN NG-SGD	0.042

Table 5 各个录音重放检测方法性能对比

Acknowledgement

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. Evans Nicholas, Kinnunen Tomi, and Yamagishi Junichi, "Spoofing and countermeasures for automatic speaker verification," *INTERSPEECH 2013*, pp. 925–929, 2013.
2. Villalba Jesús and Lleida Eduardo, "Detecting replay attacks from far-field recordings on speaker verification systems," Springer Berlin Heidelberg, 2011, pp. 274–285.
3. Wu Zhizheng, Gao Sheng, Chngz Eng, Siong, and Li Haizhou, "A study on replay attack and anti-spoofing for text-dependent speaker verification," *APSIPA 2014*, pp. 1–5, 2014.
4. Korshunov Pavel, Marcel Sébastien, Muckenhirn Hannah, A. R. Goncalves, Mello A. G. Souza, and et al R. P. Velloso, Violato, "Overview of btas 2016 speaker anti-spoofing competition," *Idiap*, 2016.
5. Sriskandaraja Kaavya, Sethu Vidhyasaharan, Le Phu, Ngoc, and Ambikairajah Eliathamby, "Investigation of sub-band discriminative information between spoofed and genuine speech," *INTERSPEECH 2016*, pp. 1710–1714, 2016.