关于自动说话人确认系统的反语音合成欺骗攻击的研究

A Study on the Countmeasures of the Automatic Speaker Verification System Against Synthetic Speech

工程领域：计算机技术
作者姓名：刘畅
指导教师：魏建国 教授
企业导师：王林 高级工程师

答辩日期	2020年6月20日		
答辩委员会	姓名	职称	工作单位
主席	方强	副研究员	中国社会科学院语言研究所
委员	冯卉	副教授	天津大学外国语言与文学学院
	侯庆志	副教授	天津大学建筑工程学院
	孙提	高级工程师	浪潮通用软件有限公司

天津大学国际工程师学院
二〇二〇年五月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 天津大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：刘畅 签字日期：2020 年 6 月 21 日

学位论文版权使用授权书

本学位论文作者完全了解 天津大学 有关保留、使用学位论文的规定。特授权 天津大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名：刘畅 导师签名：

签字日期：2020 年 6 月 21 日 签字日期：2020 年 6 月 21 日

摘　要

自动说话人确认系统（ASV）作为一种常用身份确认系统，目前被广泛地应用于银行身份验证、手机解锁登录等场景。近年来，出现了一些不法分子利用语音合成和语音转换技术，攻击ASV系统，以盗取他人的信息和钱财的情况。这无疑给ASV系统带来了严重的安全隐患。

目前，常用的语音合成和语音转换技术中，大多数都是通过基于音素的拼接和调整方法来生成语音的。这种方法会导致生成的语音与自然语音相比，不同的音素中，存在很多明显的差异。如果能够有效地利用这些差异信息，可以很好地提升合成语音检测任务的性能。此外，语音合成技术通常是基于文本信息进行合成的，在合成的过程中往往没有考虑到语音中的情感因素。因此，通过分析语音中的情感，也是一种有效地区分合成语音的重要方法。针对合成语音中存在的上述问题，本文提出了两种合成语音检测的算法。首先，是使用音素级的F-Ratio分析方法，来寻找频域中合成语音与真实语音之间差异信息分布较为集中的频段，再根据分析的结果，修改语音特征滤波器设计，从而得到更适合于合成语音检测的语音特征。其次，是针对合成语音缺乏情感的问题，基于迁移学习的思路，提出了一种利用预训练的情感识别网络对合成语音进行情感特征提取的方法。

本文通过实验，验证了提出的两种方法在合成语音检测任务中的识别能力。其中，基于音素分析的方法，在ASVspoof 2019 LA数据集中的EER和t-DCF两项评价指标均优于目前最佳的单系统结果；而基于情感特征的识别方法，则表现出了良好的泛化能力。

关键词：合成语音检测，自动说话人确认系统，反欺诈检测，音素分析，情感特征

ABSTRACT

As a commonly used verification system, the automatic speaker verification (ASV) system is currently widely used in scenarios such as bank identity verification and mobile phone unlock login. Recently, some criminals begin to use the speech synthesis and voice conversion technologies to attack the ASV system in order to steal information and money. This undoubtedly brings serious security risks to the ASV system.

At present, most of the speech synthesis and voice conversion technologies generate speech through phoneme-based stitching and adjustment methods. The speeches generated by such methods will have many obvious differences in different phonemes compared to natural speeches. If these difference information can be effectively used, the performance of the synthesized speech detection task can be well improved. In addition, speech synthesis technology is usually synthesized based on text, which does not consider the emotion in speech. Therefore, by analyzing the emotion in speech, it is also an effective method for distinguishing synthesized speech. In view of the above problems, this paper proposes two algorithms for the detection of synthesized speech. First, the phoneme-level F-Ratio analysis method is used to find the frequency band where the difference information distribution between the synthesized speech and the natural speech in the frequency domain is concentrated, and then the feature’ s filter is modified according to the analysis result. Secondly, for the lack of emotion in synthesized speech, a method for extracting emotion features of synthesized speech using a pre-trained emotion recognition network is proposed.

The experimental results show that both methods have certain synthetic speech detection capabilities. Among them, based on the phoneme analysis method, the EER and t-DCF in the ASVspoof 2019 LA dataset are better than the best single system results. The method based on emotional features has shown generalization ability.

KEY WORDS: Synthetic Speech Detection，Automatic Speaker Verification System，Antispoofing，Phoneme analysis，Emotion Feature

目 录

第 1 章 绪论

1.1 研究背景及意义

近年来，随着人工智能技术的飞速发展，利用人的指纹、虹膜以及声音等个性化的生物特征进行个人身份鉴别的生物识别技术，成为了许多科研人员当前的研究热点。生物识别技术，顾名思义就是利用人体上某些在一定时间内具有持续稳定性的、且能够反映出个体与个体之间差异的生物特性，对这些特性进行采样和转化，进而形成具有唯一性数字编码，再对这些编码进行比较，可以达到身份鉴定的目的。很多的生物识别技术，如指纹识别（Fingerprint Recognition）技术、虹膜识别（Iris Recognition）技术以及声纹识别（Voiceprint Recognition）技术等目前已经发展的比较成熟，如今正被广泛地应用于人们日常的生产和生活当中。

声纹识别技术，又称说话人识别（Speaker Recognition）技术，实现方法是对说话人的声音进行采集，并使用语音信号处理等方法，提取音频中的说话人信息，再根据提取到的信息进行说话人身份的鉴定。这一技术与指纹识别、虹膜识别等技术相比，具有实现成本低且易于操作等特点，它既不像指纹识别技术那样需要用到专门的设备，也不像虹膜识别技术那样需要进行特定的动作，只需简单的发出声音就可以进行准确的身份鉴定。因此自动说话人确认（Automatic Speaker Verification，ASV）系统得到了用户广泛的认可，在智能手机登录以及电子商务验证等情景中，都可以看到这一系统的身影。然而，最近一段时间，随着深度神经网络（Deep Neural Networks, DNN）的不断发展，尤其是生成对抗网络（Generative Adversarial Networks，GAN）模型的出现，语音合成（Speech Synthesis）技术和语音转换（Voice Conversion）技术突破了波形拼接（Waveform Concatenation）、声码器（Vocoder）等传统方法的束缚，出现了很多基于神经网络的方法，使用这些方法得到的合成语音以及转换语音在质量上都有了很大的提升。利用这些方法，许多不法分子可以轻易地对目标说话人的声音进行模仿，生成欺诈语音，进而攻击ASV系统，以达到盗取他人的隐私和钱财的目的。

上个世纪90年代末，T.Masuko等人[1]对基于隐马尔科夫模型的ASV系统在面对合成语音时的安全性进行了测试，测试结果显示一个在处理真实人声时错

误接受率（False Acceptance Rate，FAR）可以达到0%的ASV系统，在遭到合成语音的攻击时，系统的FAR猛增至超过了70%，而且实验中使用到的语料库仅包含了20个说话人的信息。2005年，P.Perrot等人[2]使用语音转换技术对说话人自动确认系统进行了类似的实验，实验将所有非目标说话人的音频都替换成了由语音转换技术得到的音频，最终系统的等错误率（Equal Error Rate，EER）由原来的16%增涨至超过60%。2012年，P.L.De Leon等人[3]使用《华尔街日报》语料库进行了一个包含300个说话人信息的合成语音检测试验，实验中测试了两个ASV系统，它们分别基于GMM-UBM模型和SVM模型，实验结果显示两个系统在面对合成语音攻击时的FAR分别达到了86%和81%。同一年，T. Kinnunen等人[4]对五个不同的ASV系统进行了面对语音转换攻击时的鲁棒性测试，测试发现即使是人耳听起来存在很明显差异的转换语音，系统也不能很好地将他们分辨出来，五个系统中表现最好的FAR也从3%增加到了17%。上述研究表明，ASV系统在没有保护措施的情况下，很容易受到合成语音和转换语音的影响，安全性方面存在较大的隐患。

为了应对欺诈语音给ASV系统带来的挑战，从2015年起，语音与信息处理技术领域顶级的国际会议Interspeech开始举办针对自动说话人系统面对的欺诈攻击的对策研究挑战赛（Automatic Speaker Verification Spoofing and Countermeasure Challenge，ASVspoof）。赛事旨在使用统一的标准评估现有技术对欺诈语音的识别能力并推动相关学者在这一领域上的研究。

1.2 国内外研究现状

在2015年ASVspoof挑战赛举办前，关于如何区分真实语音和欺诈语音的研究相对较少，且研究的方式也多是将合成语音和转换语音分开单独进行研究。

在合成语音检测方面，2001年，T.Satoh等人[5]设计了一个对于合成语音具有鲁棒性的说话人识别系统，系统通过计算输入语音帧与帧之间的对数似然差分的平均值来判断语音是否为合成语音，当平均值低于设定的阈值时，语音就会被判定为合成语音。实验选取了16名说话人作为说话人确认系统已知人员，选取了另外20名说话人作为系统的未知人员，另外使用了一个基于隐马尔科夫模型的语音合成系统，对已知人员的声音进行了合成，生成了欺诈语音。实验结果显示，没有合成语音检测的说话人确认系统，对于真实语音的FAR仅为0.005%，而对于合成语音的FAR则高达86.3%，在加入合成语音检测算法后，对于合成语音的FAR降至了0.69%，大大地提高了系统在面对合成语音攻击时的鲁棒性。然而这一方法只适用于基于隐马尔科夫模型的语音合成方法，对于其他类型的合成语音不具有很好的泛化性。在此之后，陆续出现了一些从不同的

角度对合成语音进行分析检测的研究。2005年，A. Ogihara等人[6]通过提取真实语音和合成语音的基频，发现二者存在较大的差异，并基于此提出了一种通过语音的基频信息检测合成语音的方法。2012年，P.L.De Leon等人[7]在Ogihara的基础上进行了进一步的研究，研究发现合成语音的基频在音素之间过渡时的变化要比真实语音的基频变化的更加迅速，这一差异可以通过捕捉基频的抖动来获取，于是他们提出了一种基于语音基频抖动特征的合成语音检测算法，实验结果显示该方法对于合成语音的识别率可以达到96%。除了基频方面，P.L.De Leon等人在2011年的还从相位角度对合成语音进行了研究。由于人的听觉系统对语音信号的相位信息较为不敏感[8]，所以使用声码器的语音合成方法，为了简便，通常会使用最小相位的声道模型，这种简化设计会导致合成语音与真实语音在相位上存在差异，P.L.De Leon等人[9,10]便根据这一发现提出了一种基于相对相位变化（Relative Phase Shift，RPS）的新特征用于检测使用声码器合成得到的语音，实验结果显示系统的对于合成语音的识别准确率为88%。

早期的转换语音检测方法，很多都借鉴了P.L.De Leon[9]在语音合成中的思路。Z. Wu等人[11]2012年在RPS的基础上提出了两种与相位相关的新特征，分别为相位谱的余弦归一化特征和改进的群延迟特征（Modified Group Delay，MGD），实验表明两种特征对转换语音均有较好的识别效果，其中MGD特征至今仍被许多欺诈语音检测算法所使用。此外，F. Alegre等人[12,13]研究发现，基于超向量的支持向量机（Support Vector Machine ，SVM）对欺诈攻击的检测具有较高的鲁棒性，这一研究结果也促使后续许多的相关研究采用了以SVM为后端分类器的实验方案。

自2015年起，语音与信息处理技术领域顶级的国际会议Interspeech开始举办ASVspoof挑战赛，比赛每两年举办一届，目前已经举办了3届。其中，第一届比赛[14]主要是检测使用语音合成技术和语音转换技术生成的欺诈语音，这类攻击只要针对ASV系统的逻辑访问（Logical Access，LA）部分；而第二届比赛[15]的研究重点则是录音重放得到的欺诈语音，这类攻击主要发生在ASV系统物理访问（Physical Access，PA）部分；在2019年刚刚举办的第三届比赛[16]则是同时涵盖了LA和PA两种类型的攻击，赛事的主办方将它们设置成了两个子任务进行考察。在每次比赛过程中，赛事的主办方都会发布统一的数据集，数据集分为三部分，分别为：训练集、开发集和测试集。数据集中的语音数据均由真实语音和欺诈语音组成，其中训练集和开发集中包含了每条语音的真假标签，而测试集中则没有这些标签。此外，为了验证识别方法的泛化能力，测试集中包含了训练集和开发集中没有使用到的欺诈语音生成方法。最终，各支参赛队只需将测试集中的语音在其设计的识别系统中的打分结果提交给赛方，赛方再根据提交的结果对各支队伍的系统使用EER作为标准进行评估和排名。此项赛事

的举办，解决了以往缺少统一的数据集和统一的评价标准，难以对不同方法的效果进行比较和评估的问题。随着赛事的举办，许多新的方法和思路开始涌现出来，极大地促进了欺诈语音检测技术的进步和发展。

因为ASVspoof 2015的数据集[14]中同时包含了使用语音合成方法和使用语音转换方法得到的语音数据，所以在此后的研究中，大多将这两种欺诈语音统称为逻辑访问欺诈语音，并一同进行检测。在比赛中，T.B. Patel等人[17]将Q. Li等人[18]提出的耳蜗滤波器倒谱系数与T. F. Quatieri[8]提出的瞬时频率相结合，提出了一种基于瞬时频率的耳蜗滤波器倒谱系数特征CFCCIF（Cochlear Filter Cepstral Coefficients with Instantaneous Frequency），并将它与梅尔频率倒谱系数（Mel Frequency Cestrum Coefficient, MFCC）进行了特征融合，使用高斯混合模型（Gaussian Mixture Model，GMM）作为后端分类器，最终在测试集中的EER仅为2.013%，在各支参赛队中排名第一。

2015年，M. Sahidullah等人[19]在ASVspoof 2015数据集上使用GMM和SVM两种后端分类器分别对包括线性频率倒谱系数（Linear Frequency Cepstral Coefficient，LFCC）、矩形滤波器倒谱系数（Rectangular Filter Cepstral Coefficient，RFCC）和逆梅尔频率倒谱系数（Inverted Mel Frequency Cepstral Coefficient, IMFCC）等在内的19种语音特征在欺诈攻击检测任务中的效果进行了分析比较，同时还研究了这些特征在加入一阶和二阶动态信息之后的识别效果。研究结果显示，在这19种前端特征中，识别效果最好的为LFCC特征，它在与GMM分类器组合后可以得到1.67%的EER。此外，研究中还发现，语音信号中的动态信息以及高频部分的信息对于检测欺诈语音具有较好的增益效果。

2016年，M. Todisco等人[20]首次提出将原本用于音乐信号分析的常数Q倒谱系数（Constant Q Cepstral Coefficient，CQCC）用于欺诈语音检测。与传统的MFCC等特征使用短时傅里叶变换（Short-Time Fourier Transform，STFT）对音频信号进行时域到频域的转换不同，CQCC特征是语音信号经过常数Q变换（Constant Q Transform，CQT）得到的。二者的区别在于使用STFT得到的频谱是线性分布的，而CQT得到频谱则是非线性的，其频谱在低频部分具有更高的频率分辨率，而在高频部分则具有更高的时间分辨率，这一特点使得CQCC特征可以捕捉到传统特征难以发现的细节特征，因此可以很好地进行欺诈语音的检测。实验中，研究者把CQCC特征与GMM分类器进行组合后，在ASVspoof 2015的测试集中进行实验，得到了EER仅为0.26%的出色结果。

由于上述的研究中验证了LFCC特征和CQCC特征在在欺诈语音检测任务中具有非常良好的识别效果，ASVspoof挑战赛的组织者将使用这两种特征与GMM分类器组合后的系统设计成了基线系统，发布在了2019年的挑战赛中。

与此同时，在ASVspoof 2019的比赛中，除了使用传统的EER作为评价指标之外，还引入了Tomi Kinnunen等人[21]在2018你提出的串联成本检测函数（Tandem Detection Cost Function，t-DCF）作为另外一个评价指标，这一函数不仅可以评价合成语音检测系统的性能，还可以同时考察合成语音检测系统与ASV系统组合后的实验效果。图 1-1 所示为ASVspoof 2019官方公布的比赛结果。其中队伍编号为加粗字体的表示所提交的系统中使用了多分类器融合的方法，队伍编号背景加深的表示所提交的系统中的特征提取部分和分类器设计部分中至少有一部分使用了神经网络的方法。

ASVspoof 2019 LA scenario

#	ID	t-DCF	EER	#	ID	t-DCF	EER
1	**T05**	0.0069	0.22	26	T57	0.2059	10.65
2	**T45**	0.0510	1.86	27	**T42**	0.2080	8.01
3	**T60**	0.0755	2.64	28	*B02*	0.2116	8.09
4	**T24**	0.0953	3.45	29	**T17**	0.2129	7.63
5	**T50**	0.1118	3.56	30	**T23**	0.2180	8.27
6	**T41**	0.1131	4.50	31	**T53**	0.2252	8.20
7	**T39**	0.1203	7.42	32	**T59**	0.2298	7.95
8	**T32**	0.1239	4.92	33	*B01*	0.2366	9.57
9	**T58**	0.1333	6.14	34	T52	0.2366	9.25
10	T04	0.1404	5.74	35	**T40**	0.2417	8.82
11	**T01**	0.1409	6.01	36	T55	0.2681	10.88
12	**T22**	0.1545	6.20	37	**T43**	0.2720	13.35
13	T02	0.1552	6.34	38	T31	0.2788	15.11
14	**T44**	0.1554	6.70	39	**T25**	0.3025	23.21
15	**T16**	0.1569	6.02	40	**T26**	0.3036	15.09
16	T08	0.1583	6.38	41	T47	0.3049	18.34
17	**T62**	0.1628	6.74	42	T46	0.3214	12.59
18	**T27**	0.1648	6.84	43	T21	0.3393	19.01
19	**T29**	0.1677	6.76	44	T61	0.3437	15.66
20	**T13**	0.1778	6.57	45	**T11**	0.3742	18.15
21	**T48**	0.1791	9.08	46	**T56**	0.3856	15.32
22	**T10**	0.1829	6.81	47	T12	0.4088	18.27
23	T54	0.1852	7.71	48	T14	0.4143	20.60
24	T38	0.1940	7.51	49	T20	1.0000	92.36
25	T33	0.1960	8.93	50	T30	1.0000	49.60

ASVspoof 2019 PA scenario

#	ID	t-DCF	EER	#	ID	t-DCF	EER
1	**T28**	0.0096	0.39	27	**T29**	0.2129	8.48
2	**T45**	0.0122	0.54	28	**T01**	0.2129	9.07
3	**T44**	0.0161	0.59	29	T54	0.2130	11.93
4	**T10**	0.0168	0.66	30	T35	0.2286	7.77
5	**T24**	0.0215	0.77	31	T46	0.2372	8.82
6	**T53**	0.0219	0.88	32	**T34**	0.2402	10.35
7	**T17**	0.0266	0.96	33	*B01*	0.2454	11.04
8	**T50**	0.0350	1.16	34	**T38**	0.2460	9.12
9	**T42**	0.0372	1.51	35	**T59**	0.2490	10.53
10	**T07**	0.0570	2.45	36	T03	0.2593	11.26
11	T02	0.0614	2.23	37	**T51**	0.2617	11.92
12	**T05**	0.0672	2.66	38	T08	0.2635	10.97
13	**T25**	0.0749	3.01	39	**T58**	0.2767	11.28
14	**T48**	0.1133	4.48	40	T47	0.2785	10.60
15	T57	0.1297	4.57	41	**T09**	0.2793	12.09
16	**T31**	0.1299	5.20	42	T32	0.2810	12.20
17	**T56**	0.1309	4.87	43	T61	0.2958	12.53
18	T49	0.1351	5.74	44	*B02*	0.3017	13.54
19	**T40**	0.1381	5.95	45	**T62**	0.3641	13.85
20	**T60**	0.1492	6.11	46	T19	0.4269	21.25
21	T14	0.1712	6.50	47	T36	0.4537	18.99
22	**T23**	0.1728	7.19	48	**T41**	0.5452	28.98
23	**T13**	0.1765	7.61	49	T21	0.6368	27.50
24	**T27**	0.1819	7.98	50	T15	0.9948	42.28
25	**T22**	0.1859	7.44	51	T30	0.9998	50.19
26	**T55**	0.1979	8.19	52	T20	1.0000	92.64

图 1-1　ASVspoof 2019比赛结果[16]

从官方公布的比赛结果中，我们可以发现大多数队伍的系统，在检测合成语音和转换语音时的识别效果要低于在检测重放语音时的效果。在包含合成语音和转换语音的LA数据集上，只有很少的参赛系统可以能够得到低于5%的EER，大多数系统的EER都集中在5%到10%之间。分析这些系统使用的方法可以发现，排名前二十的系统中，有九成都使用了多分类器打分融合的设计，尤其是排名第一的系统，在进行多分类器融合后，效果有了明显的提高，其提交的单系统模型，在测试集中的EER高达11.40%，而融合后的系统EER降低到了仅为0.22%。除此之外，神经网络的方法在这次比赛中也被大家所广泛使用，有超过六成的队伍在系统中使用了神经网络的方法。例如，Alejandro Gomez-Alanis等人[22]设计的系统中使用了一种名为LC-GRNN（Light Convolutional Gated Recurrent Neural Network）的循环神经网络结构作为特征提取器，

使用这一模型可以提取出语音信号的语句级特征，用来进行欺诈语音的检测；Cheng-I Lai等人[23]设计的系统中则是使用了残差神经网络（Residual Neural Network，ResNet）作为系统的后端分类器模型，这种网络结构可以有效地解决神经网络的深度增加后出现的退化问题。目前，ResNet已经被广泛地应用于图像识别领域，在语音识别领域中，也显示出了非常出色的效果。

除了上述研究外，2019年Gajan Suthokumar等人[24]创新地从音素的层面对ASVspoof 2017中的数据进行了实验分析。实验中，研究人员对数据集中的音频根据音素进行了分类，所有的音频会通过一个音素的识别器，这个识别器可以通过计算音频中一帧对应于某个音素的后验概率来判断这一帧属于哪一个音素。当某一帧在某个音素上的后验概率高于75%时，这一帧就会被认定为是这个音素。通过对音频中不同的音素进行分类，可以得到不同音素在真实语音和欺诈语音中的详细信息，之后利用这些信息，可以分别训练每个音素的GMM分类器，最后通过分析不同音素的分类器的KL散度（Kullback–Leibler Divergence）以及EER，就可以比较出不同音素在欺诈语音检测任务中的性能差异。图 1-2 所示为实验的结果，其中KL散度表示真实语音的模型与重放语音的模型之间的差异，越大说明两个模型的差异越多，越易于进行区分，而EER则是越低表示识别结果越好。

从图 1-2 中可以发现，不同的音素在重放语音检测任务中的性能存在着很大的差异，其中轻音音素的效果要普遍好于浊音音素的效果。

1.3 本文的主要工作

音素是语音合成和语音转换技术中的重要基础，大多数的合成语音和转换语音（后文中统称为合成语音）都是基于不同音素的拼接和调整得到的。因此，合成语音与真实语音中，不同的音素中都会存在着一定的差异性信息。本文以此为出发点，同时受到Gajan Suthokumar等人[24]从音素角度对语音重放攻击进行研究的启发，提出了一种基于音素分析的合成语音检测方法。方法中借鉴了F-Ratio分析方法的思路，从音素的层面挖掘合成语音与真实语音之间的差异，进而找出语音中更有利于鉴别合成语音频段，之后再根据这些频段的分布，调整语音特征的提取方法，最终设计出更适合于合成语音检测任务的语音特征，达到提高识别准确率的目的。此方法是首次将音素级的研究应用于合成语音检测任务中。

除此之外，情感信息一直都是语音中包含的一项重要信息，而这一信息在合成语音中却很难被模拟。因此，提取语音中的情感信息作为特征进行合成语音的检测，是一种非常具有潜力的方法。本文基于这一观点，提出了一种基于

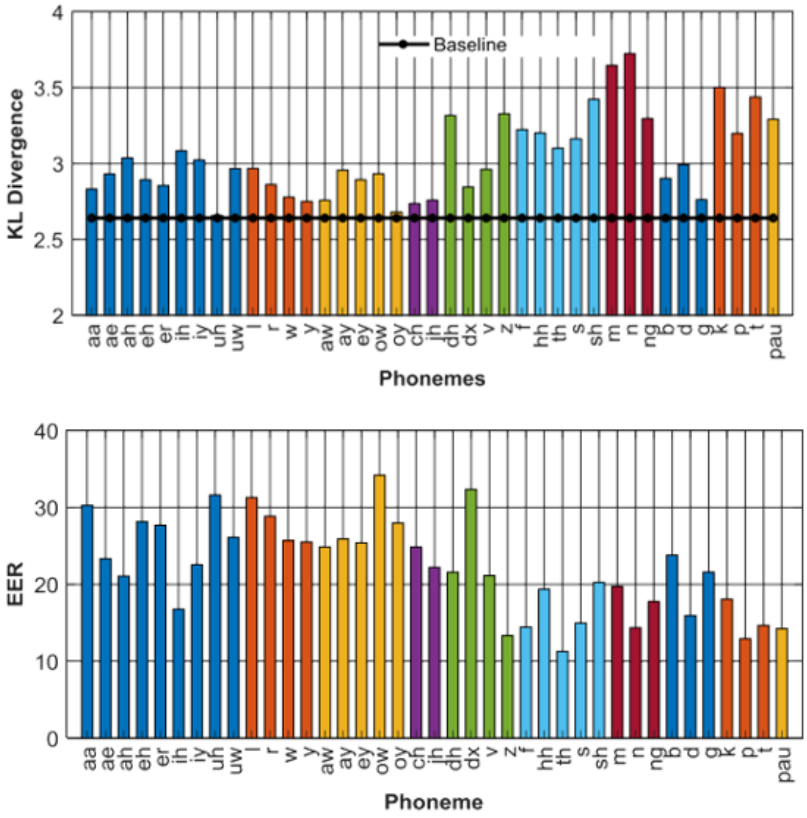


图 1-2 不同音素在重放攻击检测任务中的效果对比[24]

情感的合成语音检测方法，这也是合成语音检测领域中首次从情感角度进行尝试。

1.4 本文的文章结构

本文一共分为六章，每章中的主要内容如下：

第一章，为绪论部分，首先详细地介绍了合成语音检测技术的研究背景和实际意义，然后系统地阐述了这一技术的发展历史以及研究现状，分析了目前主流的研究方向以及发展趋势，最后提出了本文的创新工作。

第二章，相关理论及技术，首先简述了合成语音检测系统的设计方法以及它与ASV系统的融合方法。之后分别对系统的前端特征提取部分以及后端分类器部分进行了详细地介绍，介绍的内容包括目前研究中主流的算法。最后对合成语音检测系统的两种评价指标进行了说明。

第三章，数据集介绍及数据预处理，首先对本文中实验使用的ASVspoof 2019 LA数据集进行了详细的介绍，包括数据集中的数据来源以及各类数据等

分布情况等。之后对实验中涉及到的数据预处理环节进行了说明，介绍了预处理过程中每一个步骤的目的、所需要使用到的工具以及具体的参数配置。

第四章，提出了一种基于音素级的F-Ratio分析方法得到的音素适应的语音特征。详细地介绍了这一特征的设计原理以及提取过程，并通过实验，验证了新特征在合成语音检测任务中的有效性。

第五章，提出了一种基于情感特征的合成语音检测算法。详细地阐述了算法的设计思路，并设计实验对算法的有效性进行了验证，最后对实验的结果进行了分析。

第六章，为总结与展望，对本文中进行的工作以及得到的结果进行了总结和分析，并在此基础上，对本文提出的两种方法今后的研究方向进行了展望。

第 2 章　相关理论及技术

2.1　合成语音检测系统

2.1.1　合成语音检测系统设计方法

合成语音检测系统的基本功能为判断输入的语音信号是否为合成语音，图2-1 所示为合成语音检测系统的工作流程图。

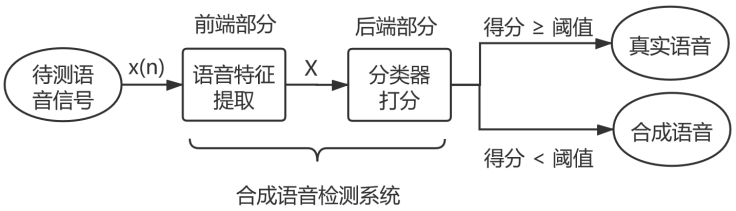


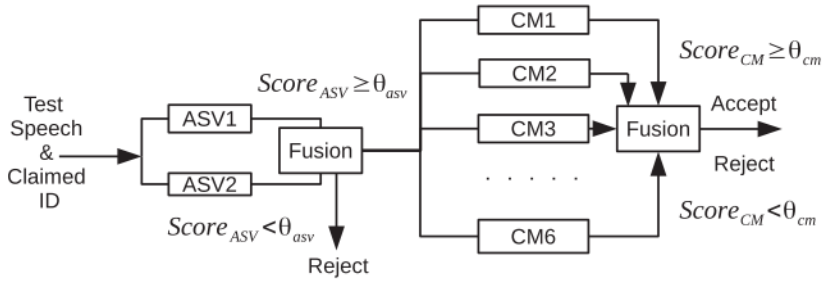
图 2-1　　合成语音检测系统工作流程图

系统主要分为两部分，分别为前端特征提取部分和后端分类器部分。待测试的语音信号输入系统后，会根据系统设定的语音特征提取方法进行对应语音特征的提取。之后，提取得到的语音特征会被送入训练好的后端分类器中，分类器对特征进行打分，最终系统将打分的结果与设定好的阈值进行比较，做出对应的决策。当得小于阈值时，待测语音将会被判定为合成语音，反之，当得分大于或等于阈值时，待测语音将被认定为真实语音。在上述的系统中前端语音特征的选择以及后端分类器的训练效果是直接影响系统判断结果准确性的两个关键因素。目前，在合成语音检测任务中，常用的语音前端特征主要有LFCC特征和CQCC特征等，常用的后端分类器包括传统的GMM分类器以及在图像识别领域广泛应用的ResNet神经网络。本章的第二、第三两节会对上述特征提取方法以及分类器设计进行详细地介绍。

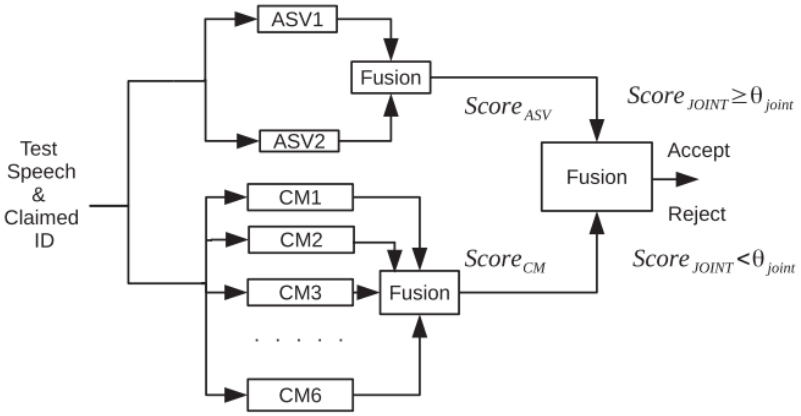
2.1.2　合成语音检测系统与ASV系统的融合方法

合成语音检测系统要实现的最终效果是与ASV系统相结合，在ASV系统遭

受到合成语音攻击时，做出正确的判断，保护ASV系统的信息安全。因此，如何将合成语音检测系统与ASV系统有效地结合起来，是在系统设计时必须要考虑的一个问题。M.Sahidullah等人[25]在2016年的研究中，提出了两种合成语音检测系统与ASV系统的融合策略，分别为串联策略和并联策略。图 2-2 所示为两种融合策略的系统流程图。



(a) 串联策略系统流程图



(b) 并联策略系统流程图

图 2-2　　两种合成语音检测系统与ASV系统的融合策略[25]

所谓串联策略，就是指先进行合成语音的检测，如果检测结果为自然语音，再执行说话人身份的判断和确认。而并联策略则是指同时进行合成语音检测和说话确认两个任务，然后综合两个任务得到的结果做出最终的决策。研究结果表明，上述两种融合方法，在真实场景中都可以有效地检测出合成语音的攻击，提高ASV系统的安全性。

2.2 合成语音检测特征提取方法

2.2.1 引言

语音是指由人的声道等器官发出的一段承载着特定信息的模拟信号，通过采样、量化和编码，连续的模拟信号可以被转换成离散的数字信号，这里用采样率来表示转换过程中每秒从模拟信号中提取的采样点个数。要直接从这些离散的数字中获取信息通常是比较困难的，这就需要使用语音信号处理的方法，对这些数字进行表征，提取出相关的语音特征，进而分析出语音中所包含的信息。因此，选择合适的特征表征方法，提取出适当的特征对于后续的相关研究就显得尤为重要。下面将介绍合成语音检测中常用的两种特征提取方法。

2.2.2 LFCC特征

LFCC特征是语音信号处理中一种常用的振幅特征。与传统的MFCC特征相比，它的区别在于滤波器组中滤波器的分布方式。MFCC特征提取时使用的滤波器分布是为了模仿人类的听觉系统而设计的，在设计时考虑到了人耳对不同频段声音的敏感度不同，因此提高了滤波器在低频部分的密度，降低了在高频部分的密度。这样的设计会导致MFCC特征在高频部分的分辨率较低，不利于发现合成语音中高频部分的信息。而LFCC特征的滤波器是沿频率轴均匀分布的，这样能够保证提取到的特征对于各个频段的关注度是相同的，而不会忽略某些频段的信息。在合成语音检测的任务中，高频部分往往被认为存在较多的有效信息，因此LFCC特征在这一任务中与MFCC特征相比具有一定的优势。LFCC特征的提取过程如图 2-3 所示。

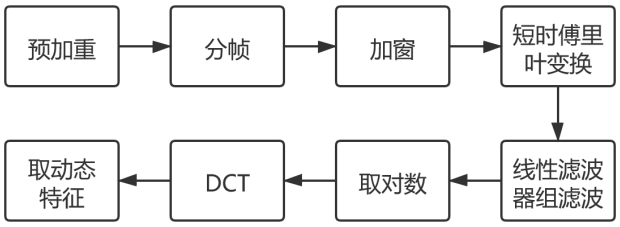


图 2-3 LFCC特征提取过程

首先，对于要对接收到的语音信号进行预加重处理。这样做的目的是为了对语音的高频部分进行增强。通常情况下处理的声音都存在高频部分与低频部分相比比较微弱的情况，机器在处理这些高频部分时就会存在一定的问题，因

此需要对高频部分做加重处理。预加重操作的计算方法如公式（2-1）所示：

$$x(n) = x'(n) - ax'(n-1) \tag{2-1}$$

其中$x'(n)$表示第n个采样点真实读取到的信号，$x(n)$表示处理后的信号，a为一个略小于1的系数，通常取0.97左右。语音信号预加重之后需要进行分帧和加窗处理，分帧就是将较长的信号分割成一段一段等长的信号样本，这样做的原因是语音信号是随时间不停变化的，只有在非常短的一段时间内，可以被认为是近似稳定的，因此需要将语音信号划分为很短的帧来研究。实验中通常使用的帧长为20ms到30ms之间，以20ms的帧长为例，当采样率为16kHz时，每一帧中包含的采样点个数为320个。为了避免帧与帧之间的变化过于突兀，通常两个相邻的帧之间会保留一段重叠区域，区域的长度一般为帧长的1/2或1/3。此外，为了进一步的平滑相邻帧之间的变化，还会对信号进行加窗处理，最常用的窗函数为汉明窗，它的特点是窗体的中间较高而两边较低。汉明窗的计算公式和加窗操作的计算公式分别如公式（2-2）和公式（2-3）所示：

$$w(m) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi m}{L}), & 0 \le m \le L-1 \\ 0, & \text{else} \end{cases} \tag{2-2}$$

$$x(n) = w(n) \cdot x'(n) \tag{2-3}$$

接下来由于语音信号在时域中很难观察出规律和特性，就需要借助STFT对每帧中的数据进行从时域到频域的转换，这样可以得到每一帧信号中的频域能量分布，进而更好地分析信号中蕴含的特性。STFT的公式如公式（2-4）所示：

$$X(k) = \sum_{n=1}^{N} x(n) e^{-\frac{j2\pi kn}{N}}, 1 \le k \le K \tag{2-4}$$

其中N为时域上每帧中的采样点个数，K表示经过傅里叶变换之后频域范围内的采样点个数。由于在听觉中，人耳对单独的频率是不敏感的，因此一些相近的频率可以被放在一起看作一个频段进行分析。具体方法就是使用滤波器组对每一帧中的各个频段进行滤波，用每个滤波器所得到的值来表征对应频段上的能量信息，这样既有利于对频谱中的能量信息进行分析，又起到了降低数据维度的效果。这里不同的特征使用到的滤波器组设计方法大多是不同的，图 2-4所示为语音信号处理中几种常用的振幅特征的滤波器设计，它们依次为LFCC特征、RFCC特征、MFCC特征以及IMFCC特征。其中LFCC特征使用的是沿频率轴均匀分布的线性三角形滤波器组，RFCC特征使用的是均匀分布的线性矩形滤波器组，而MFCC特征和IMFCC特征使用的则是非均匀分布的三角形滤波器组。

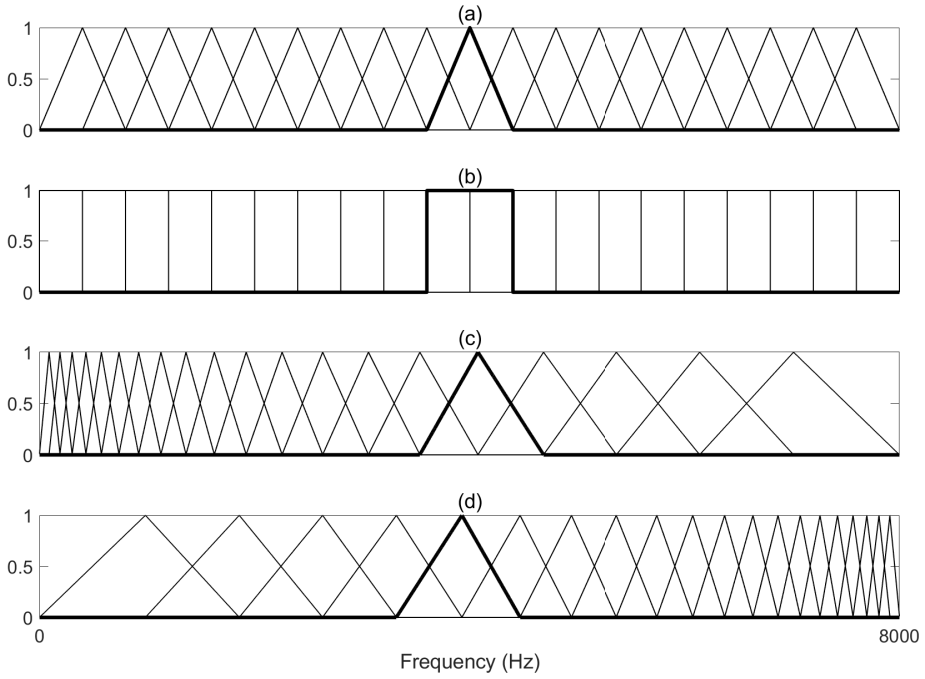


图 2-4　　几种常见的振幅特征滤波器分布

接下来，求得的数值会进行取对数运算，这样做的目的有两个，第一是因为人耳对于声音的感知并不是线性的，而是一种类似对数尺度的，它不会忽略细小的差异，也不会对非常大的差异过于敏感，所以取对数后的数值更加符合人的感知；第二是因为在时域中，表示说话人声道形态的信息与表示说话人声带振动产生气流的信息是以卷积的形式结合在一起的，经过STFT后，在频域中，这两种信息被解卷积成了乘的关系，再继续进行取对数操作，可以将它们的关系进一步转换成相加的关系，后续通过逆傅里叶变换操作，可以比较容易的将这两种信息分开。所以下一步需要进行逆傅里叶变换操作，除了可以有效地提取声道信息和声带信息外，逆傅里叶变换还可以起到压缩降维的效果，经过逆傅里叶变换的后的特征数据之间会具有高度不相关性。由于这里处理的数据已经经过了对数运算，所以数据只包含实数，因此逆傅里叶变换可以简化为离散余弦变换（Discrete Cosine Transform， DCT），取对数操作以及DCT运算的计算公式如公式（2-5）所示：

$$y(j)=\sum_{m=0}^{M-1}\log[Y(m)]\cos[\frac{j\pi}{M}(m-\frac{1}{2})],\ j=0,1,\ldots,J-1<M \tag{2-5}$$

其中，$Y(m)$为经过滤波器组后得到的能量信息，M为滤波器个数，即经过滤波器组后的数据维度，j为经过DCT变换后保留的数据维数。到目前为止，可以求得每一帧中LFCC特征的静态信息，考虑到语音是一种不断变化的信息，所以还可以通过差分的方法，提取相邻帧之间的一阶差分和二阶差分，作为LFCC特征的动态信息。一阶差分和二阶差分的计算公式分别为公式（2-6）和公式（2-7）：

$$\Delta y_t(j) = \frac{\sum_{m=-p}^{p} m y_{t-m}(j)}{\sum_{m=-p}^{p} m^2} \tag{2-6}$$

$$\Delta^2 y_t(j) = \frac{\sum_{m=-p}^{p} m \Delta y_{t-m}(j)}{\sum_{m=-p}^{p} m^2} \tag{2-7}$$

其中，$\Delta y_t(j)$和$\Delta^2 y_t(j)$分别表第t维数据的一阶差分和二阶差分，p表示计算差分时考虑的帧数半径。

2.2.3 CQCC特征

CQCC特征原本是一种用于音乐信号处理分析的振幅特征，在2016年M.Todisco等人[20]的研究中发现，CQCC特征在合成语音识别任务中有良好的效果。近年来，许多的合成语音检测系统在设计时都选择CQCC特征作为系统的前端特征。本文中使用的CQCC特征提取方法为M. Todisco等人[26]提出的改进的提取方法，图 2-5 所示为详细的特征提取过程。

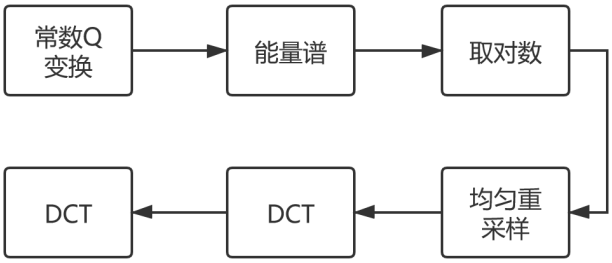


图 2-5　　CQCC特征提取过程

与传统的MFCC特征和LFCC特征不同，CQCC特征在进行时频域变换时，采用的方法不是STFT，而是常数Q变换（Constant Q Transform，CQT）。STFT和CQT本质上都可以看作是一个滤波器组，其中STFT的滤波器是沿频率轴线性分布的，而CQT的滤波器则是以指数分布的，另外两种变换中滤波器的

频带带宽也不相同。这里可以用因子Q来衡量其中每个滤波器的选择性，它和滤波器的中心频率以及频带带宽的关系如公式（2-8）所示：

$$Q = \frac{f_k}{\delta f} \tag{2-8}$$

其中，f_k表示中心频率，δf表示频带宽度。在STFT中，每个滤波器的带宽只与窗函数有关，且是一个定值，所以由公式（2-8）可以看出，Q因子会随着频率增高而变大。而在CQT中，Q因子始终为一个常数，所以当中心频率越高时，带宽也会随之增大。这样做的好处在于变换后得到的频谱在低频部分具有较高的频率分辨率，而在高频部分具有较高的时间分辨率。图 2-6所示为STFT和CQT所得的频谱图对比。其中，上图为经过STFT变换后得到的频谱图，下图为经过CQT变换后得到的频谱图。对比两张图片可以发现，经过STFT变换后得到的频谱图频率轴上的刻度是均匀分布的，而经过CQT变换后得到的频谱图频率轴上的刻度是沿指数分布的，且由低频向高频变化的过程中，时间轴的分辨率也逐渐提高。

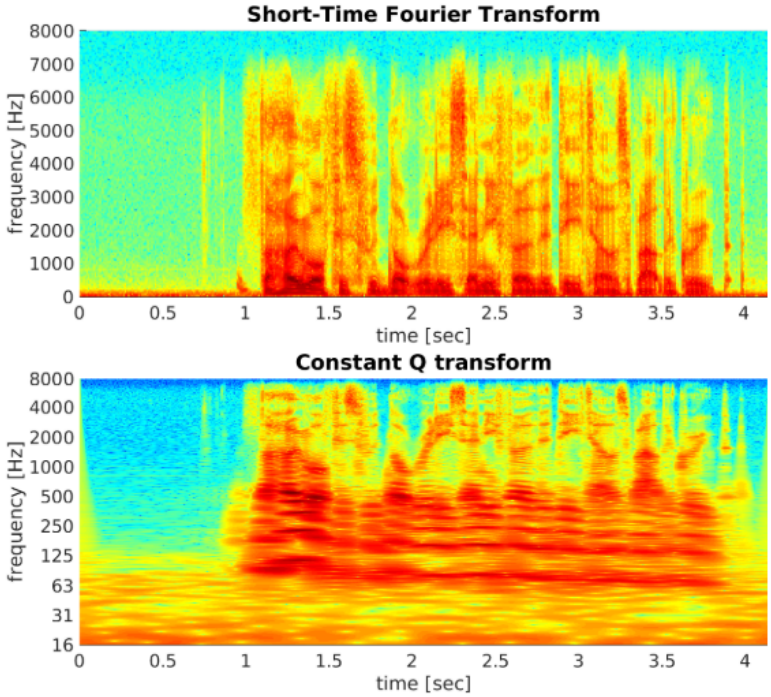


图 2-6　　STFT与CQT频谱图对比[26]

经过CQT变换得到的频率谱数据，通过取绝对值平方，可以得到对应的能量谱数据，再取对数可以得到对数能量谱数据，由于经过CQT变换的对数能量谱中频率轴是沿几何分布的，而后续的DCT变换要求能量谱满足线性分布，为

了解决这一问题，M.Todisco借鉴了Wolberg等人在信号重构领域提出的方法，提出了一种重采样的方法，解决了能量谱由几何空间向线性空间转换的问题。后续的步骤与提取LFCC特征时相同，需要通过DCT变换进行数据压缩，再分别计算静态特征的一阶差分和二阶差分作为CQCC特征的动态信息，便可以得到实验中使用的CQCC特征。

2.3 合成语音检测分类器模型

2.3.1 引言

合成语音检测系统中的后端分类器主要分为生成模型和判别模型两大类。其中，生成模型中的代表方法包括GMM分类器和基于i-vector的SVM分类器，而判别模型则主要是以深度神经网络的方法为主，如卷积神经网络（Convolutional Neural Networks, CNN）中的ResNet网络和循环神经网络（Recurrent Neural Network, RNN）中的长短期记忆网络（Long Short-Term Memory，LSTM）等。本节将以GMM分类器和ResNet网络为例，对合成语音检测中的两类分类器进行介绍。

2.3.2 GMM

GMM，即高斯混合模型，是一种参数估计模型，其本质是用有限个规则的高斯分布来拟合其它复杂的概率分布。其中，每个单独的高斯分布都可以看做是混合模型的隐变量，混合模型的概率密度与各个高斯分量之间的关系如公式（2-9）所示：

$$p(x \mid \lambda) = \sum_{k=1}^{K} \alpha_k N(x \mid \mu_k, \Sigma_k) \tag{2-9}$$

其中，$p(x \mid \lambda)$表示样本x在模型中的条件概率，K表示GMM模型中高斯分量的个数，α_k表示第k个高斯分量的线性组合系数，即组合的权重，这里要满足各分量的权值之和恒为1，$N(x \mid \mu_k, \Sigma_k)$表示第$k$个高斯分量的概率密度函数，$\mu_k$和$\Sigma_k$和分别表示第$k$个高斯分布的均值和方差。

对于多分类问题，可以使用不同分类中的样本数据，训练不同的概率分布模型。在分类时，选择条件概率最大的分类模型作为分类结果。以合成语音检测任务为例，语音样本可以分为真实的自然语音和合成的欺诈语音两类，分别进行特征提取后用来训练得到各自的GMM模型，这里训练的方法多采用最大期望（Expectation-Maximization，EM）算法，它是一种通过不断地迭代进行极大

似然估计的方法，常被用于对含有隐变量的概率模型进行参数估计。算法的计算过程包括两步，分别是根据参数的现有值计算期望的E步骤和根据期望求参数极大值的M步骤，两个步骤不断的迭代，直到参数收敛于某一极值为止。

在得到真实语音与合成语音训练好的GMM模型后，便可以进行对未知语音的真假性判断。分别求出未知语音的特征X在两个模型中的对数似然值，然后如公式（2-10）所示计算两个值的差便可以得到未知语音的得分，当该得分高于设定的阈值时，未知语音便会被认定为真实语音，反之，则会被认定为合成语音。

$$S(X) = \log p(X \mid \lambda_{\text{genunie}}) - \log p(X \mid \lambda_{\text{spoof}}) \tag{2-10}$$

GMM作为一种经典的机器学习方法，在合成语音检测任务中，展现出了非常良好的识别效果，与目前流行的深度神经网络的方法相比，它具有泛化能力强且易于训练的优势。因此，GMM模型成为了合成语音检测领域中被广泛使用的后端分类器之一。

2.3.3 ResNet

ResNet，即残差神经网络，由Kaiming He等人[27]于2015年提出，自提出后，便受到了图像识别领域的广泛关注，目前已经成为应用最为广泛的CNN结构之一。ResNet解决的主要问题是深度神经网络随着网络深度的增加而引发的退化现象。在卷进神经网络中，网络的层数越多，网络中可以提取到的信息也就越丰富，因此在很多的研究中，都会通过增加CNN的层数，来提高网络的识别准确率。然而当网络的层数达到一定数量后，再继续增加网络的深度，反而会导致网络的识别效果变差，这就是所谓的退化问题。图 2-7 所示，为不同层数的网络模型在CIFAR-10数据集上进行测试的实验结果。其中左图为在训练集中的错误率曲线，右图为在测试集中的错误率曲线。观察两张图片中的曲线可以发现，无论是在训练集还是在测试集中，深度为56层的神经网络识别效果均要低于深度为20层的神经网络。

为了解决这一问题，Kaiming He等人参考计算机视觉中常用的残差表示的概念，在神经网络的各层之间引入了一个恒等的快捷连接，设计出来如图 2-8 所示的残差块结构。

残差块的内部可以通过快捷连接跳过一层或多层，将某一层的输入与跳跃后某一层的输出相加作为接下来一层的输入，继续进行后面的训练。具体到CNN中，通常是跳过一个卷积的组合层，将卷积前的输入与卷积后经过激活函数的输出相加。通过这些类似恒等映射的快捷连接，可以将网络中浅层的信

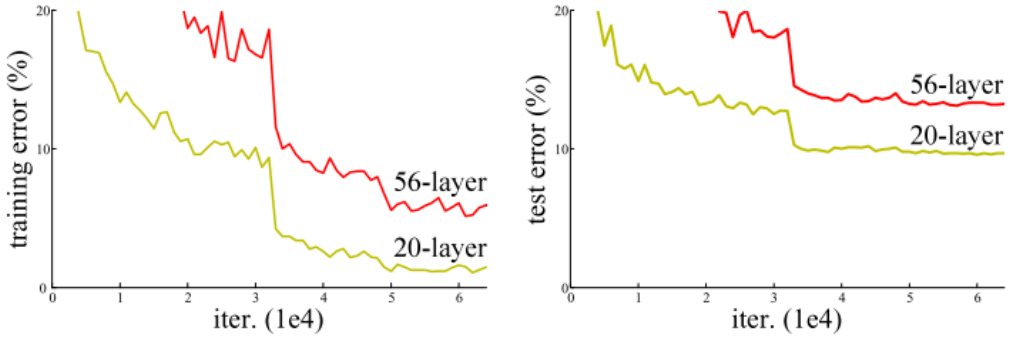


图 2-7　在CIFAR-10数据集上20层网络和56层网络的错误率[27]

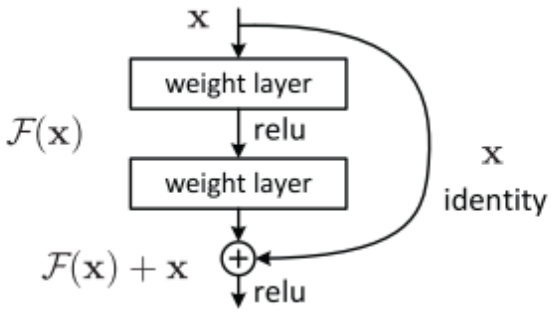


图 2-8　ResNet中的残差块设计[27]

息传递给网络的更深层，从而可以很好地抑制退化问题的影响。实验结果显示，在参数量相近的情况下，ResNet相比于传统的VGG等模型具有更高的准确性。

由于ResNet在图像领域表现出了良好的效果，近年来语音领域中也有越来越多的研究开始使用这一工具。具体的思路是提取语音的语谱图作为输入信息，传入ResNet网络中进行训练和识别。在ASVspoof 2019的挑战赛中，就有相当一部分的队伍使用了这一方法，并且取得了较好的结果。

2.4　合成语音检测系统评价指标

2.4.1　引言

如何能够准确全面地衡量一个合成语音检测系统的准确性，是关系到相关技术进步和发展的重要因素。目前常用的评价指标包括EER、DET（Detection Error Tradeoff ）曲线和t-DCF等。接下来，本节将对ASVspoof 2019挑战赛中所使用的两个评价指标EER和t-DCF进行介绍。

2.4.2 EER

在本章的第一节，介绍了合成语音检测系统的决策方法，即后端分类器对提取到的语音特征进行打分，当分数高于某一阈值时，语音就会被判定为真实语音，反之则会被判定为合成语音。因此，系统中阈值设定的高低将直接影响系统最终的识别效果。在银行的身份校验等具有较高信息安全需求的应用场景中，往往需要采用较高的阈值，以降低系统误将合成语音识别为真实语音，从而用户财产损失的风险。但是，这种设置在保护用户财产安全的同事，可能会导致系统将一部分的真实语音识别为合成语音，进而影响用户在使用时的用户体验。因此，在一些用户使用频率较高，且安全性需求相对较低的使用场景中，则需要适当地降低判定的阈值。为了更好地衡量选择的阈值是否合适，研究人员提出了三个相关的评价指标，分别为错误接受率（FAR）、错误拒绝率（FRR）和等错误率（EER）。其中，FAR和FRR分别表示误将合成语音识别为真实语音的概率和误将真实语音识别为合成语音的概率，它们的计算方法分别如公式（2-11）和公式（2-12）所示。

$$P_{\text{fa}}(\theta) = \frac{\#\{\text{Spoof trials with score} > \theta\}}{\#\{\text{Total spoof trials}\}} \tag{2-11}$$

$$P_{\text{miss}}(\theta) = \frac{\#\{\text{Human trials with score} \leq \theta\}}{\#\{\text{Total spoof trials}\}} \tag{2-12}$$

其中，P_{fa}表示系统阈值为时，系统的FAR，P_{miss}则表示系统阈值为时，系统的FRR。将FAR和FRR之间随着系统阈值变化而变化的关系绘制出曲线图，可以得到系统的DET曲线。其中，曲线的x轴为系统的FAR变化趋势，y轴为系统到的FRR变化趋势。

系统的EER表示系统在取某一阈值θ_{EER}时，系统的FAR和FRR得到了相等的结果，此时FAR或FRR的值即为EER。三者的关系如公式（2-13）所示：

$$\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}}) \tag{2-13}$$

系统的EER在计算时可以从DET曲线中快速获得，DET曲线与直线$y = x$交点的坐标值，即为系统的EER。EER指标可以同时衡量合成语音检测系统出现错误接受和错误拒绝两种错误的概率，EER的值越低，表明系统中发生两种错误的几率越小，即系统的检测性能越好。自2015年第一届ASVspoof挑战赛开始举办以来，EER就被作为评价合成语音检测系统的一项重要评价指标。

2.4.3 t-DCF

在本章的第一节中，曾介绍了合成语音检测系统与ASV系统的融合方法，这两个系统可以理解为是一个具有共同总体目标的主系统的两个子系统。在两个系统协同的工作模式中，一个系统的识别结果势必会对另一个系统产生影响，进而影响系统总体的稳定性。因此，一种可以从系统整体的角度对合成语音检测系统的效果进行评价的指标就显得十分必要。2018年Tomi Kinnunen等人[21]对ASV系统中的常用评价指标——最小检测代价（Minimum Detection Cost Function，minDCF）进行了拓展，提出了一种可以反映合成语音检测系统与ASV系统采用串联的融合方式后，整体识别性能的评价指标——t-DCF，其计算方法如公式（2-14）所示：

$$\begin{aligned}\text{t-DCF} = & C_{\text{miss}}^{\text{asv}} \cdot \pi_{\text{tar}} \cdot P_a(s,t) \\ & + C_{\text{fa}}^{\text{asv}} \cdot \pi_{\text{non}} \cdot P_b(s,t) \\ & + C_{\text{fa}}^{\text{cm}} \cdot \pi_{\text{spoof}} \cdot P_c(s,t) \\ & + C_{\text{miss}}^{\text{cm}} \cdot \pi_{\text{tar}} \cdot P_d(s,t)\end{aligned} \tag{2-14}$$

其中，$C_{\text{miss}}^{\text{asv}}$和$C_{\text{fa}}^{\text{asv}}$分别代表ASV系统出现错误拒绝和错误接受时的惩罚代价，$C_{\text{fa}}^{\text{cm}}$和$C_{\text{miss}}^{\text{cm}}$分别代码合成语音检测系统出现错误拒绝和错误接受时的惩罚代价，π_{tar}、π_{non}和π_{spoof}分别表示目标说话人、非目标说话人以及合成语音出现的先验概率，s和t分别表示合成语音检测系统和ASV系统的阈值，$P_a(s,t)$表示目标说话人的语音正确地通过了合成语音检测系统，而被ASV系统错误拒绝的概率；$P_b(s,t)$表示非目标说话人的语语音正确地通过了合成语音检测系统，而被ASV系统错误接受的概率；$P_c(s,t)$表示合成的目标说话人语音被合成语音检测系统错误接受，导致ASV系统接受了此语音的概率；$P_d(s,t)$表示目标说话人的语音被合成语音检测系统错误拒绝的概率，上述四种概率的计算方法如下：

$$P_a(s,t) \triangleq (1 - P_{\text{miss}}^{\text{cm}}(s)) \times P_{\text{miss}}^{\text{asv}}(t) \tag{2-15}$$

$$P_b(s,t) \triangleq (1 - P_{\text{miss}}^{\text{cm}}(s)) \times P_{\text{fa}}^{\text{asv}}(t) \tag{2-16}$$

$$P_c(s,t) \triangleq P_{\text{fa}}^{\text{cm}}(s) \times (1 - P_{\text{miss,spoof}}^{\text{asv}}(t)) \tag{2-17}$$

$$P_d(s,t) \triangleq P_{\text{miss}}^{\text{cm}}(s) \tag{2-18}$$

在确定了代价函数的计算方法后，需要对函数中四种错误情况的惩罚代价进行讨论设计，这里需要根据系统的实际使用场景来确定不同的代价值，如果

是银行身份确认等需要较高安全性的使用场景，就需要将错误接受的代价设置为一个较高的值，而如果是在比较注重用户体验的使用场景中，则可以选择较高的错误拒绝代价值。

由于采用t-DCF作为评价指标，可以更好地分析合成语音检测系统与ASV系统结合后的综合性能，更加贴近于实际的使用情景，在ASVspoof 2019挑战赛中，举办者将t-DCF作为了EER之外的另一个评价指标。

2.5 本章小结

本章首先对合成语音检测系统的设计思路以及系统与ASV系统的融合策略进行了概述，紧接着分别从系统的前端特征提取方面以及后端分类器方面对常用的几种特征提取方法和分类器进行了详细的阐述，最后介绍了两种合成语音检测系统的评价指标。

第 3 章　数据集介绍及数据预处理

3.1　数据集介绍

3.1.1　引言

本文中实验所使用的数据集为ASVspoof 2019语料库，它是ASVspoof 2019挑战赛中发布的专门用于研究ASV系统所面临的欺诈攻击时的应对策略的语料库。语料库中将欺诈攻击按照不同的情景分为了两类，分别为LA情景和PA情景，并为两类情景分别准备了对应的数据集。其中，LA情景是指对ASV系统的逻辑访问部分进行攻击，主要的攻击方法为语音合成技术和语音转换技术；PA情景则是指对ASV系统的物理访问部分进行攻击，攻击手段为语音的录音重放。本文研究的主要内容为欺诈攻击检测中合成语音和转换语音的检测，因此，研究过程中的实验均是在ASVspoof 2019语料库中的LA数据集上进行的。

3.1.2　ASVspoof 2019 LA数据集

ASVspoof 2019语料库的LA数据集中的语音数据全部是基于VCTK语料库得到的，VCTK语料库采集了包含46名男性和61名女性在内的107名说话人的真实语音，所有的真实语音均采用了相同的录音配置，且没有信道和背景噪声的干扰。LA数据集中的真实语音均直接选取自VCTK语料库，而数据集中欺诈语音则均是由这些真实语音使用不同的语音合成和语音转换技术得到的，所有语音数据的采样率均为16kHz。LA数据集的详细数据分布如表 3-1 所示。

表 3-1　　ASVspoof 2019 LA数据集数据分布

数据集	说话人数量		攻击类型数量	音频数量	
	男性	女性		真实语音	欺诈语音
训练集	8	12	6	2580	22800
开发集	4	6	6	2548	22296
测试集	21	27	13	7355	63822

数据集中的数据被分成了三个子集，分别为训练集（Train）、开发集（Development）和测试集（Evaluation）。其中，训练集和开发集中的欺诈语音

均来源于6种相同的语音合成和语音转换技术，这6种技术作为已知的攻击类型，可以用来对合成语音检测系统进行训练和调整，而测试集中欺诈语音的生成方法包含了2种上述的已知攻击和11种与6种已知攻击不同的语音合成和语音转换技术，这11种技术作为系统面临的未知攻击类型。在进行合成语音检测系统的开发过程中，研究人员只可以使用训练集和开发集中的数据，对系统进行模型设计和参数调整，测试集中的数据只能用来评价系统的识别效果。

在6种已知的攻击类型中，包括了4种语音合成技术和2种语音转换技术，这些技术中包括了传统的波形图拼接技术、基于频谱图过滤的技术[28]以及基于神经网络的[29]和基于WaveNet的声码器技术[30]等。未知的11种攻击类型中，除了包含6种新的语音合成技术和2种新的语音转换技术外，还包括了3种语音合成和语音转换融合的技术，这些技术中涉及了许多已知攻击类型中所没有涉及的新的手段，例如基于Griffin lim算法的语音合成技术[31]以及使用生成对抗网络的语音合成方法[32]等。通过分析合成语音检测系统在面对这些未知攻击类型时的表现，可以推测出系统在真实应用场景中的防护效果，有助于研究人员开发出泛化能力更强、更具有鲁棒性的合成语音检测系统。表 3-2 和表 3-3 所示分布为ASVspoof 2019官方公布的LA数据集中已知攻击类型和未知攻击类型在实现时所使用的技术详情。

表 3-2　ASVspoof 2019 LA数据集中已知攻击类型使用的技术

攻击类型编号	技术种类	技术实现方法
A01	语音合成	neural waveform mode
A02	语音合成	vocoder
A03	语音合成	vocoder
A04	语音合成	waveform concatenation
A05	语音转换	waveform filtering
A06	语音转换	spectral filtering

3.2　数据预处理

3.2.1　引言

本文中提出的两种合成语音检测算法，分别需要基于语音的音素信息进行分析和使用语音的语谱图进行特征提取。因此，需要对数据集中的数据进行相关的数据预处理操作，包括格式转换、提取文本、进行音素对齐以及提取语谱图等几个步骤，本章接下来的内容将对这些预处理操作的方法以及相关参数配置进行介绍和说明。

表 3-3 ASVspoof 2019 LA数据集中未知攻击类型使用的技术

攻击类型编号	技术种类	技术实现方法
A07	语音合成	vocoder+GAN
A08	语音合成	neural waveform
A09	语音合成	vocoder
A10	语音合成	neural waveform
A11	语音合成	griffin lim
A12	语音合成	neural waveform
A13	语音合成+语音转换	waveform concatenation + waveform filtering
A14	语音合成+语音转换	vocoder
A15	语音合成+语音转换	neural waveform
A16	语音合成	waveform concatenation
A17	语音转换	waveform filtering
A18	语音转换	vocoder
A19	语音转换	spectral filtering

3.2.2 格式转换

ASVspoof 2019 LA语料库中提供的所有语音格式均为FLAC格式，而在进行语音识别和语音标注的过程中，使用到的工具大多要求处理的文件格式为WAV格式，因此需要进行语音格式转换。FLAC和WAV格式都是无损格式，其中FLAC格式是WAV格式的一种无损压缩格式，它通过编码压缩的方式，可以达到节省空间，便于储存的目的。因为它是一种无损压缩，所以对音频进行解码后，得到的WAV格式文件与原本的文件相比不会丢失任何信息，此外，FLAC格式具有较好的抗损性，其压缩后每一帧都是不相关的，在压缩或传输过程中出现损坏时，受影响的内容只有特定的损坏帧，而不会影响前后帧的内容。因此，FLAC格式非常适用于音频文件存档等场景的使用。

本文的实验中使用了Matlab中的audioread函数和audiowrite函数，采用编写脚本的方法对ASVspoof 2019 LA数据集训练集中的所有FLAC格式的语音进行了批量地格式转换，后续的语音文本提取以及音素对齐等工作都是在此基础上使用了WAV格式的文件进行的。

3.2.3 提取文本

由于目前常用的进行语音音素对齐的工具中，大多数都需要将语音的音频文件和对应的语音文本同时输入到音素对齐工具里，而ASVspoof 2019 LA数据集中并没有提供相关的文本文件，所以实验需要使用语音识别系统对训练集中的语音音频进行文本信息的提取。本文的实验中使用的语音识别系统为百度公司提供的免费开放接口，该接口可以将语音精准地识别为文字，目前接口支持包括普通话、粤语以及英语在内的多种识别模式。此外，还支持使用用户自己

的语料库进行自训练，提升识别效果。

本文的实验中使用了接口中的短语音场景，通过编写python脚本的方法，提取了ASVspoof 2019 LA数据集训练集中所有语音对应的文本。由于训练集中相同说话人的合成语音是基于相同的文本使用六种已知攻击技术合成的，同时，考虑到训练集中的欺诈语音存在合成的音频清晰度较差的情况，可能会影响语音识别接口的识别效果。因此，实验中对文本相同的6个语音的识别结果进行了比较，当识别结果不相同时，选择6个语音中出现次数最多的结果，作为6个语音的对应文本，这样可以降低语音合成技术对文本的识别结果造成的影响。

3.2.4 音素对齐

音素对齐（Phoneme alignment）是语音标注（Speech annotation）中的一项常见任务。这项任务的主要目的是获得语音中的所有音素信息，包括音素类型、音素边界和持续时间等。这些信息是许多语音相关研究的基础，它们可以为后续分析语音中各个音素的发音特点，总结语音的发音规律提供帮助。例如，在发音可视化教学研究中，可以通过音素对齐操作，提取每个音素的第一和第二共振峰，并以此来分析发音者在发音时的舌位信息，进而对发音过程中存在的问题进行纠正。

然而，手动的音素对齐操作是一项非常耗费人力的工作，即使是经验丰富语音学研究人员可能也需要花费数分钟来完成一条语音的音素对齐工作。因此，从上个世纪末，便开始出现了一批可以自动完成音素对齐任务的软件，如HTK工具包、Julius和SPPAS等。在使用这些软件时，只需要将语音的音频以及对应的文本信息同时输入到软件中，就可以得到语音中音素的标注结果。这些工具的出现很大程度上减轻了研究人员进行语音分析时的负担，促进了相关研究的发展。

本文的实验中使用的自动标注工具是由Brigitte Bigi研发的SPPAS软件[33,34]，这一软件操作简便，既可以使用图形化界面进行快速操作，也可以使用命令行或编写脚本的方式进行批量处理。软件目前支持对包括法语、英语以及中文普通话在内的十余种语言的语音音频进行自动标注。标注的内容包括根据文本进行音素切分、结合音频进行音素对齐等。研究显示，软件得到的音素对齐结果比较准确，可以满足大多数音素相关研究的精度需求。图 3-1 所示为SPPAS软件的操作界面。

本文的实验中使用SPPAS软件对ASVspoof 2019 LA数据集中的语音进行了标注，标注中使用的文本内容为上一节提取的音频文本。标注完成后，为了验证标注结果的准确性，实验采用随机抽样的方法，随机地选取了一定数量的语

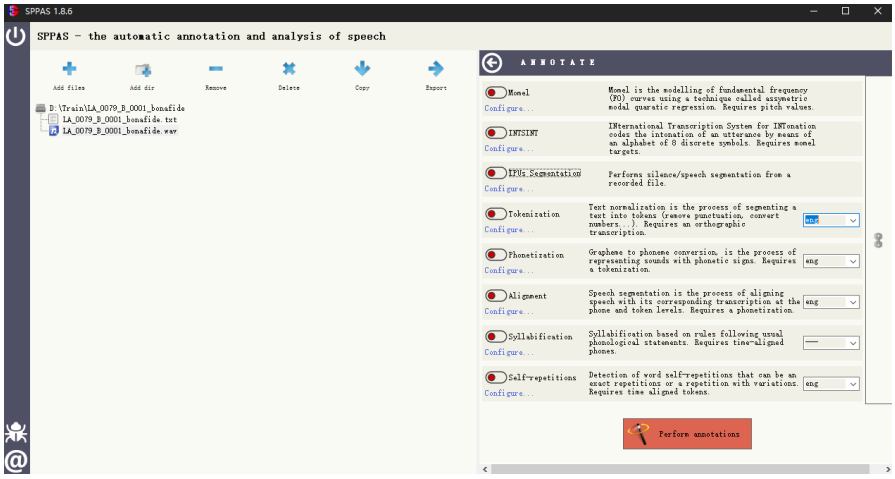


图 3-1 SPPAS软件操作界面

音，进行了人工标注，并将得到的结果与自动标注软件得到的结果进行对比。对比结果显示，使用SPPAS软件得到的标注结果与人工标注的结果相比，音素边界误差均在20ms左右，可以满足本文中实验的精确度需求。图 3-2 所示为使用SPPAS软件得到的自动标注结果在Praat软件中的展示。其中，红色方框中标识出的为音素*d*的结尾处自动标注的结果与人工标注的结果存在19ms的误差。

3.2.5 提取语谱图

本文提出的实验方法中，使用到了基于CNN的特征提取方法。由于CNN通常适用于处理多维的输入数据，而原始的语音信号是一种一维的数据，所以，在语音领域使用CNN方法时，通常是将一维的语音信号转先转换成二维语谱图数据，再输入到CNN中进行训练或特征提取。另外，由于CNN中的输入要求为尺寸固定的数据，而不同的语音信号提取到的语谱图会因为语音的时长不同而导致语谱图的尺寸存在差异。所以，对语音信号提取语谱图的第一步就是对语音进行分段，即将长短不同的语音信号切分成长度相等的语音段，当语音的长度不是段长的整倍数时，就需要对最后一段进行补零或者复制操作，即对长度不足一段的部分用零或者用最后一帧的数据进行填充，补齐一整段的长度。之后在每段内部再进行分帧、加窗和FFT等操作，便可以得到每一段对应的语谱图。

本文的实验中使用了Matlab的spectrogram函数对ASVspoof 2019 LA数据集中包括训练集、开发集和测试集在内的所有语音音频提取了对应的语谱图信息。实验中，提取语谱图时对于不足整段长的部分进行了补零处理，其他涉及到的参数配置如下：每段的段长330ms，每帧的帧长20ms，帧与帧之间的帧移为10ms，即每段内包含32帧，傅里叶变换的点数NFFT为256，最终提取得到的

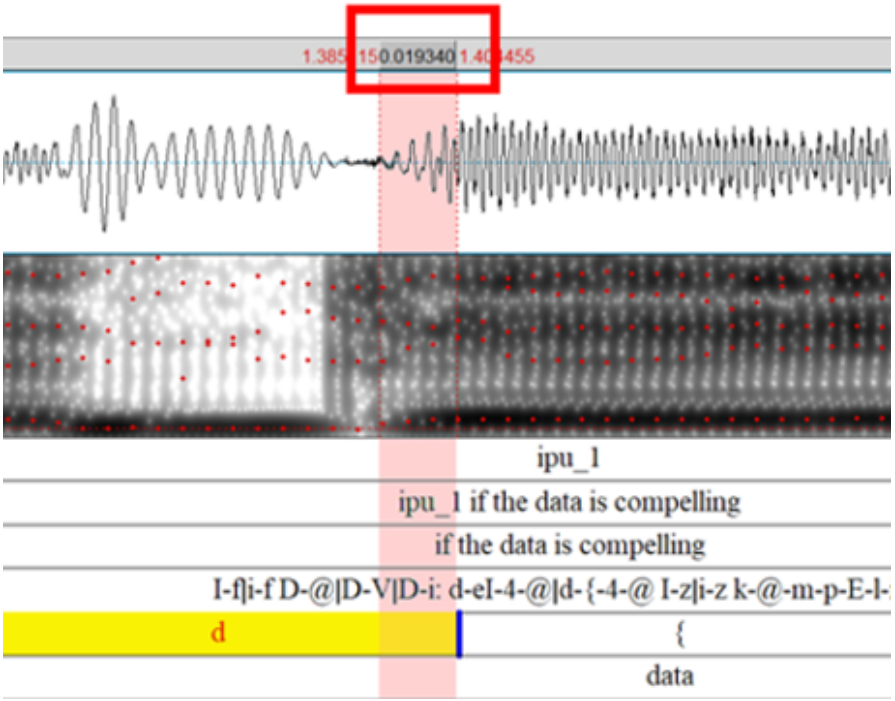


图 3-2　　使用SPPAS得到的自动标注结果

每段语音的语谱图尺寸为32×129。

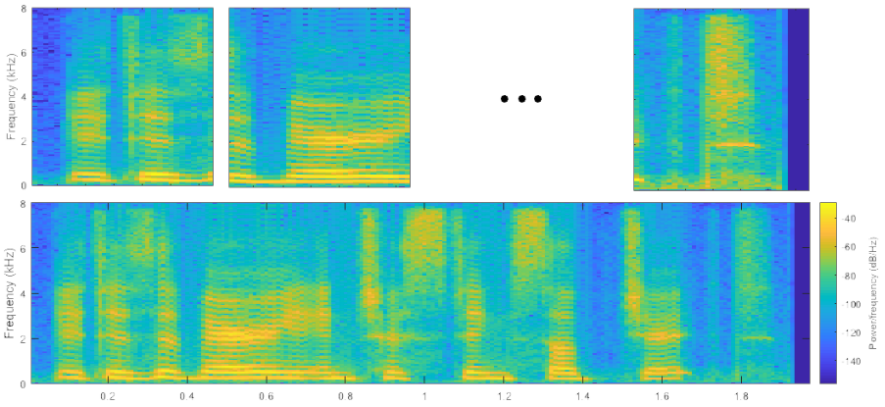


图 3-3　　语音分段前后的语谱图

图 3-2 所示为训练集中一条语音音频分段后各段的语谱图与分段前整体语谱图的对比效果，最后一段语谱图右侧的深蓝色部分，为不足一段长度时补零后的效果。

3.3 本章小结

本章中首先介绍了实验中使用到的ASVspoof 2019 LA数据集的相关信息，包括数据集中各类语音的分布情况以及其中合成语音所使用到的具体实现方法。接下来，本章讲解了实验中涉及的数据预处理操作，包括了语音文件格式转换、利用语音识别系统提取语音文本、借助自动标注工具SPPAS进行音素对齐以及提取语音的语谱图等多个步骤，并详细地说明了每个步骤的实现方法，以及其中的参数配置。

第 4 章　基于F-Ratio分析的音素适应特征提取方法

4.1　引言

音素作为语音发音过程中的最基本单位，是许多语音合成技术的实现基础。如何发现合成语音中的不同音素与真实语音中音素的差别，对于更好地检测合成语音具有非常重要的意义。传统的语音检测技术，往往是简单地尝试不同的特征或不同的分类器在识别任务中的效果，并不能从本质上分析出合成语音与真实语音之间的本质差异，因此很难得到非常好的识别效果。本章采用F-Ratio的分析方法，从音素的层面对合成语音与真实语音进行比较研究，找到了二者在各个频段上的差异分布情况，进而提出了一种更加适合合成语音检测任务的新特征，最后将新的特征提取方法应用于ASVspoof 2019 LA数据集上，验证了其有效性。

4.2　F-Ratio分析方法

语音信号中蕴含着丰富多彩的信息，例如文本信息、说话人信息、语种信息以及情感信息等。语音特征提取就是为了从语音信号中挖掘出这些信息，进行表征，为后续的研究提供基础。使用不同方法提取的到的语音特征，往往可以反映出语音信号中不同方面的信息，如常用的MFCC特征，是参考人的听觉系统进行设计的，常被用于文本相关的语音识别任务中；CQCC特征，在设计时充分考虑了音乐中音阶的原理，因此是音乐识别和分析时最常用的特征之一。由此可以看出，对于不同的任务而言，选择一个合适的特征提取方法是至关重要的，当选择的特征提取方法可以很好地挖掘出任务中所需要的信息时，那么这一特征在任务中的识别效果一定是非常出色的。

F-Ratio分析方法是Xugang Lu和Jianwu Dang[35]在2008年提出的一种频域上的语音信息分析方法，最初被应用于文本无关的说话人识别任务中。这一方法可以通过分析频域中各个频段上的能量信息分布，设计出适合于说话人识别任务的专用特征，之后这一方法又被应用到了许多语音相关的分类识别任务中，均表现出了良好的实验结果。在语音特征提取的过程中，经过FFT处理的语音信号会从时域信号转换为频域信号，使用滤波器组对频域信息进行滤波，本质上

就是用不同的滤波器对不同频段上的信息进行表征，用低维的数据来代替特定频段上的高维频谱数据。因此，如何选择出适合于任务需求的频段并对频段上承载的数据进行表征，将直接影响语音特征在任务中的表现。F-Ratio分析方法的本质就是要通过对一定数量的样本进行分析，找出承载分类任务中所需的个性化信息最多的频段，增大特征在这些频段上提取信息的权重，降低在信息分布较少的频段上的权重，从而设计出一个适合于当前分类任务的语音特征。

在F-Ratio分析方法中，使用F-Ratio值的高低来表示频域中某一频段上用于区分说话人身份信息的多少，某一频段上的F-Ratio值定义如公式(4-1)所示：

$$\text{F-ratio} = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{\sum_{i=1}^{M} N_i}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(x_i^j - u_i)^2} \tag{4-1}$$

其中，M表示用于研究的样本中分类的数量，N_i表示第i个分类中的语音帧数，x_i^j表示第i个分类的第j个帧在这一频段上的频域数据，u_i和u分别表示第i个分类中每一帧数据的平均值以及全体样本中每一帧数据的平均值，它们的定义如公式(4-2)和公式(4-3)所示：

$$u_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_i^j \tag{4-2}$$

$$u = \frac{1}{\frac{1}{\sum_{i=1}^{M} N_i}}\sum_{i=1}^{M}\sum_{j=1}^{N_i} x_i^j \tag{4-3}$$

当某一频段的F-Ratio值越大时，说明该频段上存在的分类信息越多，在后续的特征提取过程中，就应当增加对这一频段的关注度。提高对频域中某一特定频段关注度的方法主要有两种，第一种是增加这一频段上的滤波器数量，使用更加精细、更加密集的滤波器来采集这一频段上的数据特征；另一种方法是对不同频段上的滤波器提取到的数据乘以不同的权重，对需要更高关注的频段提取到的数据乘以一个较大的权重，对于其他不需要过多关注的频段则乘以一个较小的权重。Xugang Lu等人[35]在研究中发现采用第一种方法设计的语音特征的识别效果要好于第二种方法。

本文中所研究的合成语音检测任务，本质上可以看作是一个只包含真假两种类型的二分类问题。因此，可以使用F-Ratio分析方法来发掘真实语音与合成语音之间的差别，设计出更适用于合成语音检测任务的语音特征。

4.3 基于F-Ratio分析的音素适应特征提取方法

上一节中介绍的F-Ratio分析方法是一种文本无关的分析方法，在分析时没有考虑到语音中不同音素的发音对语音的影响。2014年，Songgun Hyon等人[36]在F-Ratio的基础上，提出了一种音素级的F-Ratio分析方法，作者将这一方法应用于说话人识别任务中的，取得了良好的效果。在在语音的发音过程中，不同音素所涉及的发音器官是不同的，由于人与人之间的发音器官存在一定的差异，导致不同人发出的音素也会存在一定的差异。从音素的角度进行F-Ratio分析，可以避免不同音素之间发音时的差异与说话人的个性化信息之间产生混淆，从而能够更好地挖掘出说话人的个性化信息，设计出更为有效的语音特征。同理，在语音合成领域中，绝大多数的合成方法都是以音素为基础，来生成合成语音的，不同的合成方法在生成语音时，对不同音素的处理方式会有很大的不同，这就会导致合成语音中的各个音素与真实语音中的各个音素之间会存在比较大的差异。在Gajan Suthokumar等人[24]对重放语音检测的研究中，可以发现不同的音素在检测任务中的表现是有很大差异的，由此可以推测出不同的音素在合成语音检测任务中的检测效果必定也存在一定的差异。因此，如果只是简单地使用原有的F-Ratio分析方法，把所有的音素混合在一起进行研究的话，不同音素之间本身存在的差异信息将与合成语音中存在的差异信息混淆，进而很难挖掘出音素中所蕴含的对于合成语音检测任务有增益效果差异信息，从而影响最终的识别效果。

基于上述的思想，本文提出从音素角度对合成语音进行F-Ratio分析，找出不同音素在频域中最适合于检测合成语音的频段，之后再根据不同音素在语音可能出现的频率，设置相关的权重，并结合这一频率对之前分析所得的不同音素的有效频段进行整合，得到一个整体的合成语音在频域中的差异信息分布情况。最后，根据分布情况调整语音特征提取过程中的滤波器设计，进而得到一个可以综合利用合成语音中每个音素的有效信息的新特征——音素适应频率倒谱系数（Phoneme Adaptive Frequency Cepstrum Coefficient, PAFCC）。

PAFCC特征的提取过程主要包括以下几个步骤：

首先，对待研究的合成语音和真实语音进行数据预处理，得到语音中的详细音素信息，包括语音中所包含的音素种类以及每个音素在音频中对应的起始时间。

之后，根据上一步提取到的音素信息，对语音中的音素进行分帧、加窗和傅里叶变换，得到不同音素的频谱信息。再使用一组均匀分布的滤波器组对得到的频谱进行处理，以获取音素中各个频段的能量数据。

接下来，使用F-Ratio分析方法对上述步骤中得到的数据进行分析，得到合成语音在不同音素中的差异性信息分布情况，音素k在频段l上的F-Ratio值计算方法如公式（4-4）所示：

$$\text{F-ratio}(k)^{l}=\frac{\frac{1}{T}\sum_{t=1}^{T}(u_k^t-u_k)^2}{\frac{1}{\sum_{t=1}^{T}N_{tk}}\sum_{t}^{T}\sum_{j}^{N_{tk}}[x_k^t(j)-u_k^t]^2} \tag{4-4}$$

其中，T表示分类数量，在合成语音检测任务中，分类只有真实语音和欺诈语音两种，所以T恒等于2; N_{tk}表示音素k在分类t中的帧数；$x_k^t(j)$表示音素k在分类t中的某一帧在频段l上的数据，u_k^t表示音素k在分类t中的各帧在频段l上的数据平均值，u_k表示音素k的所有帧在频段l上的数据平均值。u_k^t和u_k的公式如下：

$$u_k^t=\frac{1}{N_{tk}}\sum_{j=1}^{N_{tk}}x_k^t(j) \tag{4-5}$$

$$u_k=\frac{1}{\sum_{t=1}^{T}N_{tk}}\sum_{t=1}^{T}\sum_{j=1}^{N_{tk}}x_k^t(j) \tag{4-6}$$

接下来需要对刚刚分析得到的不同音素的F-Ratio值进行整合，由于部分音素在某些频段上的F-Ratio值可能会出现过大的情况，如果直接进行整合可能会使其他变化较为平缓的音素的信息被忽略。因此，这里需要先对每个音素在各个频段上的F-Ratio值进行归一化处理，得到每个音素在频域中F-Ratio的分布比例（Phoneme F-Ratio Distribution，PFD），其计算方法如公式（4-7）所示：

$$\text{PFD}_k^l=\frac{\text{F-ratio}(k)^l}{\sum_{i=1}^{L}\text{F-ratio}(k)^i} \tag{4-7}$$

其中，L表示频域中划分的总频段数。

接下来可以对所有音素的F-Ratio分布情况进行整合，这里采用的方法是取加权平均的计算方式，其中权值采用的是不同音素的出现频率，即所有用于分析的合成语音和自然语音中，提取到的不同音素的总帧数。公式（4-8）所示为整合后的整体F-Ratio（Global F-Ratio，GF）在频段l上的计算方法：

$$\text{GF}^l=\sum_{k=1}^{p}\frac{N_k}{N}*\text{PFD}_k^l \tag{4-8}$$

其中N表示语音中所有音素的总帧数，N_k表示语音中音素k的总帧数。

通过上述的分析过程，可以得到用于研究的语音数据在对音素进行整合后的信息分布情况。进而可以推测出合成语音和真实语音在频域中差异性信息分

布情况，并以此为依据，对语音特征提取时所用的滤波器组进行设计。设计的原则为适当增加信息分布较多的频段上的滤波器数量，以提高PAFCC特征在这些频段上的频率分辨率。最后，将重新设计的滤波器组应用到传统的语音特诊提取过程中，便可以得到适合于合成语音检测任务的PAFCC特征。

4.4 实验及结果分析

4.4.1 引言

为了验证本章中提出的PAFCC特征在合成语音检测任务中的有效性，本节在ASVspoof 2019 LA数据集上设计了一系列的实验，来对比PAFCC特征与传统的LFCC特征、CQCC特征的识别效果。接下来将对实验的过程、相关配置以及实验结果进行详细的介绍，并给出相关的实验结果分析。

4.4.2 实验过程及配置

实验中首先对ASVspoof 2019 LA数据集中训练集的语音音频进行数据预处理操作，操作分为音频文件的格式转换、提取语音的对应文本以及语音的音素对齐三个步骤，每个步骤的目的以及详细的操作配置在本文第三章数据预处理部分进行了详细的介绍。

接下来，实验对预处理操作得到的音素数据进行了F-Ratio分析，分析时首先需要对音频进行时频域的转换，转换方法为短时傅里叶变换，变换中的参数设置为：分帧的长度为25ms，帧移为10ms，使用的窗函数为汉明窗，傅里叶变换的点数为512。在得到频域信息后，实验使用了80个均匀分布的三角带通滤波器来提取对应频段的数据，提取到的数据按照上一节中介绍的音素级的F-Ratio分析方法进行了分析，以获取频域中合成语音与真实语音差异信息的分布情况。图 4-1 所示为ASVspoof 2019 LA数据集中训练集语料使用音素级F-Ratio分析方法到的实验结果。

从上图中的结果中可以看出，真实语音和合成语音在0-600Hz、4300-5500Hz、6500-6900Hz以及7600-8000Hz这几个频段上，具有比较明显的差异，说明在这些频段中，存在的可以用于鉴别合成语音的信息比较丰富。基于上述的结果，实验对PAFCC特征的滤波器组进行了设计，主要的设计思路是增加上述频段中，滤波器的密度，即使用一组细小且密集的滤波器组来采集这些频段中的信息。为了更加均匀的提取这些频段上的信息，实验中除了使用传统的三角滤波器外，还尝试使用了矩形滤波器进行实验。图 4-2 所示为实验中设计滤波器分布方案，其中图 4-2 (a) 为采用三角滤波器的方案，图图 4-2 (b) 为采用矩形

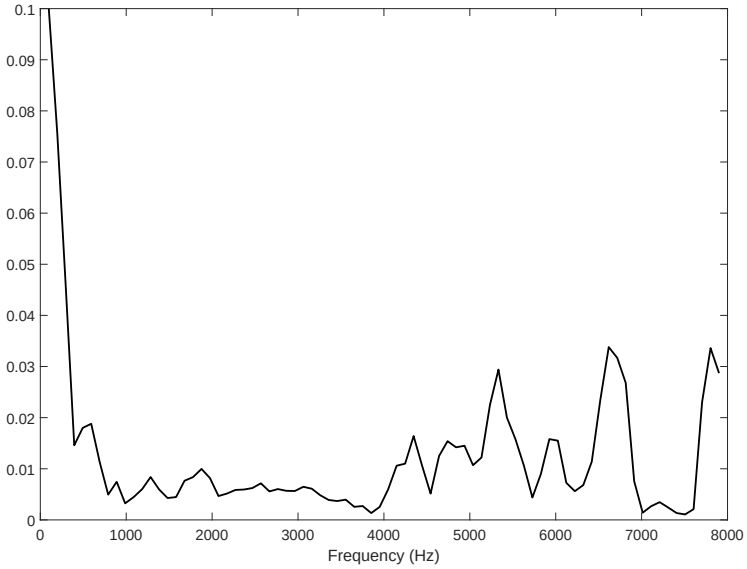


图 4-1　基于音素级的F-Ratio分析方法得到的差异信息分布情况

滤波器的方案，两种方案除了滤波器的形状存在差异外，滤波器的数量、大小以及中心频率的分布均是相同的。

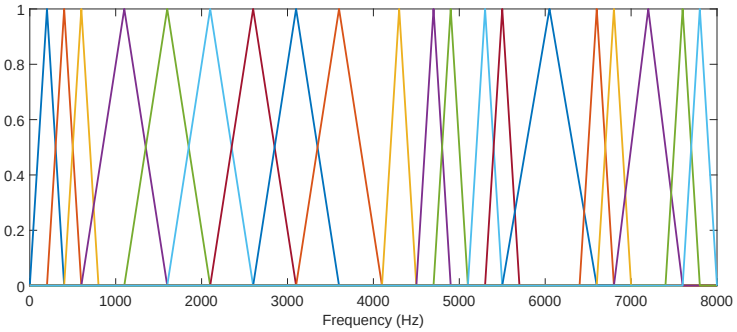
此外，为了验证基于音素的F-Ratio分析方法在合成语音检测任务中相对于传统的F-Ratio方法的优势，实验中又使用传统的F-Ratio分析方法对ASVspoof 2019 LA数据集的训练集进行了分析，分析后得到的信息分布情况如图 4-3 所示。

对比图 4-3 中的结果与图 4-1 中使用音素级F-Ratio分析得到的结果，可以发现使用两种方法得到的信息分布情况存在较大的区别。

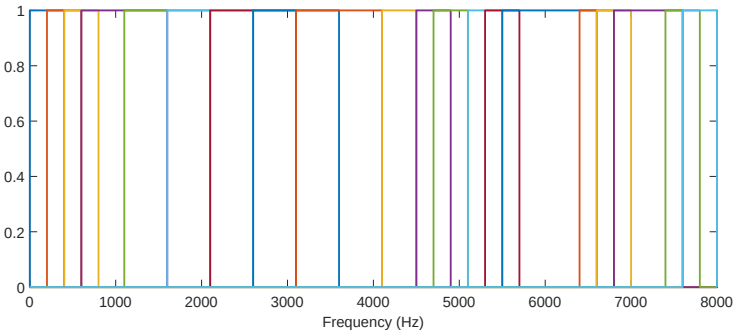
使用上述的几个滤波器设计，实验完成了对应的特征提取，特征提取过程中，涉及到的参数配置为：分帧的大小为20ms，帧长为10ms，采用的窗函数为汉明窗，傅里叶变换的点数为512。提取到的特征共包含60维数据，其中前20维是由滤波器滤波后的频域信息经过DCT变换等操作后得到的静态信息，另外的40维数据是由静态信息的一阶差分和二阶差分组成的动态信息。

此外，实验还选择了目前合成语音检测任务中最常用的LFCC特征和CQCC特征进行对比。其中，LFCC特征提取过程中，除滤波器的分布与PAFCC特征不同外，其余参数均相同。CQCC特征则使用了ASVspoof 2019挑战赛中使用的默认设置，特征的维数为90维。

后端分类器方面，实验中分别使用了GMM和ResNet两种分类器对上述特征的识别效果进行比较。其中，GMM分类器的配置为ASVspoof 2019 挑战赛中



(a) 使用三角滤波器组的设计方案



(b) 使用矩形滤波器组的设计方案

图 4-2 PAFCC特征的两种滤波器分布设计

的默认配置，模型中的高斯分量数为512，最大迭代次数为10次。ResNet采用了Moustafa Alzantot等人[37]在ASVspoof 2019挑战赛中所提出的深度残差神经网络模型，模型的网络各层结构以及残差块的构成如图 4-4 (a) 和图 4-4 (b) 所示。

4.4.3 实验结果及分析

表 4-1 所示为使用GMM模型得到的实验结果，其中使用传统F-Ratio分析方法得到的特征记为FR，使用三角滤波器得到的PAFCC特征记为PAFCC-T，使用矩形得到的PAFCC特征记为PAFCC-R。

表 4-1 五种前端特征基于GMM模型的实验结果

特征名称	开发集		测试集	
	EER(%)	t-DCF	EER(%)	t-DCF
CQCC	0.43	0.0123	9.57	0.2366
LFCC	2.71	0.0663	8.09	0.2116
FR	0.04	0.0007	6.13	0.1671
PAFCC-T	0.04	0.0010	4.16	0.1116
PAFCC-R	0.07	0.0006	3.98	0.1000

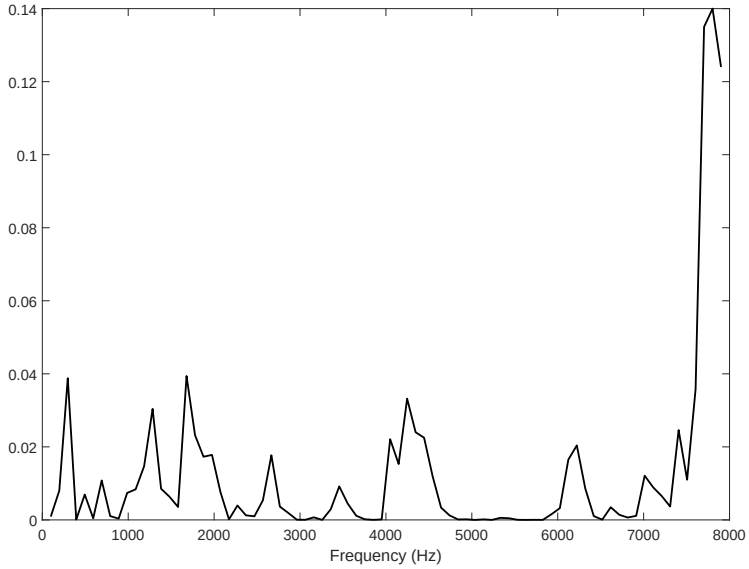
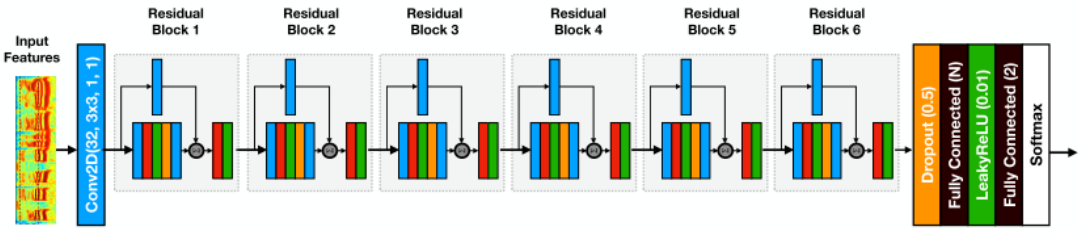


图 4-3　基于传统F-Ratio分析方法得到的差异信息分布情况

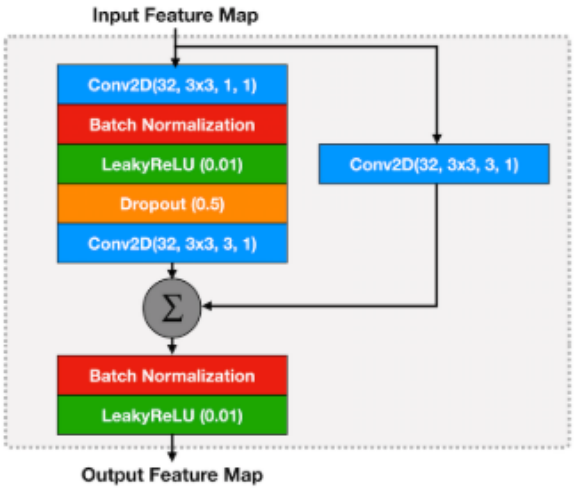
观察GMM模型中的实验结果，可以发现以下结论：

(1) 使用F-Ratio分析方法得到的三种特征，与传统的CQCC特征和LFCC特征相比，无论是以EER作为评价指标，还是以t-DCF作为评价指标，在识别效果上均有了明显的提升。在开发集中，三种特征的EER和t-DCF指标都下降到了几乎为零的水平，说明这三种特征对于已知的攻击类型都具有良好的检测效果。在测试集中，使用传统F-Ratio分析方法得到的FR特征，EER指标相较于CQCC特征下降了35.95%，t-DCF指标则下降了29.37%，相比于LFCC特征，EER和t-DCF两项指标分别下降了24.23%和21.03%。使用音素级的F-Ratio分析方法得到的实验结果中，使用三角滤波器的PAFCC-T特征在EER上相较于CQCC特征和LFCC特征分别下降了56.53%和48.58%，在t-DCF上则分别下降了52.83%和47.26%。把特征提取过程中的滤波器组改为矩形滤波器组后，PAFCC-R特征的EER和t-DCF指标相较于CQCC特征分别下降了58.41%和57.73%，相交于LFCC特征则分别下降了50.80%和52.74%。以上结果都表明，使用F-Ratio分析方法对合成语音和真实语音进行比较分析，有助于找到频域中包含差异信息较多的频段，根据分析的结果有针对性的设计特征提取过程中使用的滤波器组，可以提高特征在合成语音检测任务中的准确率。

(2) 相比于使用传统的F-Ratio分析方法得到的FR特征，使用音素级的F-Ratio方法分析得到的两种PAFCC特征在测试集中所表现出的提升效果更



(a) 实验中使用的ResNet网络各层结构



(b) 实验中使用的ResNet残差块结构

图 4-4 实验中使用的ResNet模型[37]

加明显。其中，使用三角滤波器组的PAFCC-T特征在测试集中的EER指标相比于FR特征下降了32.14%，t-DCF指标则下降了33.21%。使用矩形滤波器组得到的PAFCC-R特征在测试集上的EER指标和t-DCF指标相比于FR特征分别下降了35.07%和40.18%。上述结果可以表明，使用传统的F-Ratio分析方法对合成语音进行研究时，语音中的各个音素会被混合在一起进行分析，这样会导致存在于合成语音音素中的差异信息与不同音素间本身存在的差异信息混合在一起，从而难以很好地发掘出合成语音与真实语音之间的差异。而在使用音素级的F-Ratio分析方法进行分析时，音素间的差异信息所带来的影响得到了很好地抑制，所以合成语音在每个音素中存在的差异性信息都可以被有效地发掘出来，之后再根据这些信息进行系统和综合的分析，便可以将它们有效地利用起来，从而提升特征在合成语音检测任务中的准确性。

(3) 在使用音素级的F-Ratio分析方法得到了两种PAFCC特征中，使用矩形滤波器组的PAFCC-R特征在测试集的识别效果要好于使用三角滤波器组得

到的PAFCC-T特征。其中，在EER指标方面，PAFCC-R特征得到的结果相比于PARCC-T特征降低了4.32%，在t-DCF方面，PAFCC-R特征则下降了10.39%。分析这一结果产生的原因，本文认为在使用三角滤波器对频域中进行信息采集时，对于滤波器滤波频段上的数据的是按照一定的权重进行采集的，不同频率上的数据是按照与滤波器中心频率之间的距离，折合成一定的系数相乘后采集的，这样会导致对每个滤波器中心频率附近的区域关注度更高，而对于距离中心频率较远的区域关注度则相对较低，这种设计不能做到均匀地对频域中的信息进行提取。而使用矩形滤波器则不存在这一问题，它在采集滤波器滤波频段上的信息时，是按照相同的系数进行提取的，对每个频率上的信息都有相同的关注度。因此，当通过F-Ratio分析方法得到合成语音的差异性信息在频域中的分布情况后，使用矩形滤波器组可以更好地从信息分布较为集中的频段中提取相关信息，进而提高语音特征的识别准确性。

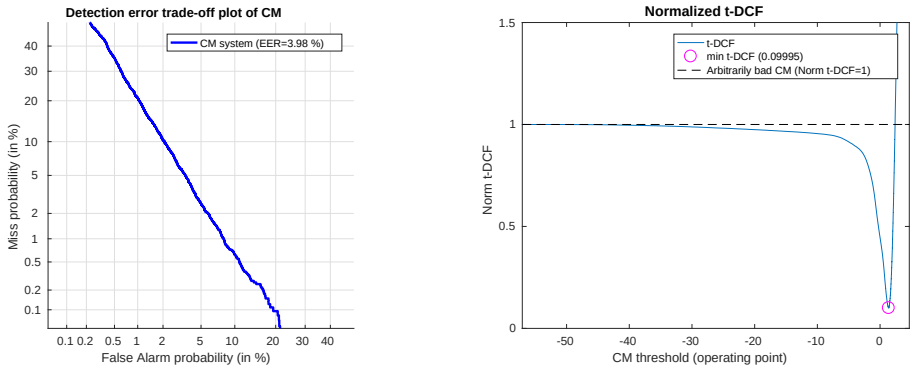
表 4-2　五种前端特征基于ResNet模型的实验结果

特征名称	开发集		测试集	
	EER(%)	t-DCF	EER(%)	t-DCF
CQCC	0.01	0.0002	7.69	0.2166
LFCC	0.71	0.0211	6.85	0.1304
FR	0.08	0.0018	6.78	0.1457
PAFCC-T	0.04	0.0006	6.17	0.1452
PAFCC-R	0.08	0.0020	4.84	0.1299

表 4-2 所示，为使用ResNet模型作为后端分类器，对五种特征进行实验的结果。观察表中的结果，可以发现几种特征在测试集中的识别效果之间的关系与使用GMM模型作为分类器时的关系基本相同，使用基于F-Ratio分析方法得到的三种特征的识别效果均要好于两种传统的语音特征；另外，在这三种特征中，使用基于音素的F-Ratio分析方法得到的两种PAFCC特征，识别效果又均好于使用传统F-Ratio分析方法得到的FR特征；而在两种PAFCC特征中，使用矩形滤波器组的得到的PAFCC-R特征的EER和t-DCF指标均低于使用三角滤波器组的PAFCC-T特征，识别效果更加出色。

观察上述两组实验可以发现，识别结果最好的模型为PAFCC-R特征与GMM模型组合后的系统。该系统在ASVspoof 2019 LA数据集测试集中的EER指标结果为3.98%（如图 4-5 (a) 所示），t-DCF指标结果为0.1000（如图 4-5 (b) 所示）。

为了研究本章中提出的方法与目前研究领域中处于领先水平的方法之间的关系，本文将实验中的最佳系统与ASVspoof 2019挑战赛官方公布的比赛结果进



(a) DET曲线 (b) t-DCF指标

图 4-5 PAFCC-R特征与GMM模型组成的系统在ASVspoof 2019 LA数据集测试集中的实验结果

行了比较。表 4-3 所示为ASVspoof 2019挑战赛中LA部分的两个最佳单系统模型与本章中实验得到的最佳模型的测试结果。其中，ASVspoof 2019 Best-1表示在所有参加ASVspoof 2019挑战赛的62支队伍提交的系统中，测试集中EER指标最优的单系统模型，而ASVspoof 2019 Best-2表示测试集中t-DCF指标最优的单系统模型。PAFCC-R&GMM表示本章中实验的最优模型，系统为单系统模型，系统采用的前端特征为PAFCC-R特征，使用的后端分类器为GMM模型。

表 4-3 本章中实验的最佳结果与ASVspoof 2019挑战赛中LA部分的最佳单系统结果对比

特征名称	开发集		测试集	
	EER(%)	t-DCF	EER(%)	t-DCF
ASVspoof 2019 Best-1	0.63	0.0162	4.04	0.1100
ASVspoof 2019 Best-2	0.16	0.0044	5.06	0.1000
PAFCC-R&GMM	0.07	0.0006	3.98	0.1000

观察表 4-3 中的数据可以发现，本章中提出的PAFCC-R特征与GMM模型组合后得到的单系统模型在ASVspoof 2019 LA数据集的测试集中的表现要优于挑战赛中的两个最佳单系统模型。其中，与ASVspoof 2019 Best-1系统相比，在EER指标方面实验结果降低了1.49%，在t-DCF指标方面实验结果降低了9.09%；与ASVspoof 2019 Best-2系统相比，两者在t-DCF指标方面的实验结果相同，在EER指标方面实验结果则降低了21.34%。上述结果均表明，本文中提出的使用基于音素级的F-Ratio分析方法得到的音素适应特征PAFCC在合成语音检测任务中具有较为出色的识别效果，可以达到甚至超越目前研究的最佳水平。

4.5 本章小结

本章提出了一种基于音素级F-Ratio分析方法得到的音素适应频率倒谱系数。这一方法通过分析合成语音与自然语音中各个音素在频域中的差异性信息分布情况，可以发现更适合于鉴别合成语音的频段，在此基础上改进特征提取时的滤波器分布，进而得到更适合于合成语义检测任务的语音特征。实验表明，该特征在对于合成语音检测任务具有出色的识别效果。

第 5 章　基于情感特征的合成语音检测算法

5.1　引言

语音是人与人之间交流沟通的一种重要方式，它不仅承载着说话人所要表达的文本信息，还蕴含着丰富的情感信息。即使是相同的文本信息，当说话人所要表达的情感不同时，展现出的语音特性也是具有很大的差别的。然而，语音合成技术在生成语音的过程中，往往只是简单地基于文本信息进行合成，却很少加入文本对应的情感信息。在听辨实验中，人们可以非常轻松地分辨出合成语音，一个很重要的原因就是合成语音往往缺乏说话人表达情感时音调的起伏变化。因此，通过情感信息来检测合成语音，是合成语音检测领域中一个重要的研究方向。本章中首次提出了一种基于情感特征的合成语音检测方法，并在ASVspoof 2019 LA数据集上进行了相关的实验。

5.2　基于情感特征的合成语音检测算法

语音中情感信息的差异，一般可以通过分析语音信号中的基频和音高等参数的变化情况来进行研究。然而，传统的帧级别语音特征，如MFCC特征、LFCC特征等，通常难以反映出语音中的这些参数变化，因此往往难以应用于情感识别任务中。

目前，常用的情感识别技术主要分为两种，第一种是使用手工设计的情感特征集，如GeMAPS特征集[38]、eGeMAPS特征集和ComParE特征集等，这些特征集是一些专家手工设计的，通常由一些低水平的描述特征（Low Level Descriptors，LLDs）和高水平的统计特征（High level Statistics Functions，HSFs）组成，使用这些特征集对语音进行表征，再结合一些常用的后端分类模型，如传统的隐马尔科夫模型（Hidden Markov Model, HMM）或一些深度神经网络的模型等，便可以实现识别语音中情感信息的目的。第二种情感识别技术则是直接将语音的语谱图输入神经网络中进行训练和识别，这是一种“端到端”（End-to-End）的神经网络模型。在传统的方法中，神经网络通常是作为后端分类器，用来给提取好的语音特征进行分类的，而在“端到端”的模型中，神经网络需要同时完成前端的特征提取任务以及后端分类器的任务。这种方法不仅可以利用神经

网络的识别能力来进行分类，同时还可以充分地利用神经网络强大的表征能力进行特征提取。通过对训练数据的学习，神经网络可以从语谱图中提取出最适合表现语音中情感信息的特征。使用这种“端到端”的方法进行情感特征提取，得到的特征不会受到人的主观影响，通常可以反映出人工设计的特征中难以发现的信息。

结合上述的语音情感特征提取方法，本文提出了一种基于情感特征的合成语音检测算法。算法中利用了“端到端”的神经网络模型的特征提取的思想，使用经过预训练的情感识别神经网络进行特征的提取。具体方法为将训练集中的合成语音与真实语音输入用情感语音数据库预训练好的神经网络模型中，选取模型中最后一层全连接层的输出作为提取到的情感特征。从全连接层中得到的情感特征，并没有经过神经网络模型中最后的softmax层，所以得到的特征向量并没有被对应到某种具体的情感上，只是对语音中所存在的某些情感特点进行了表征。通过这种方法得到的情感特征，可以将合成语音与真实语音之间在情感上存在的差异很好地反映出来。之后再根据语音的真实性对这些特征使用分类器模型进行训练，模型训练完成后，对待测语音使用相同的方法进行特征提取，便可以判断出语音的真假性。算法的详细实现过程如图 5-1 所示。

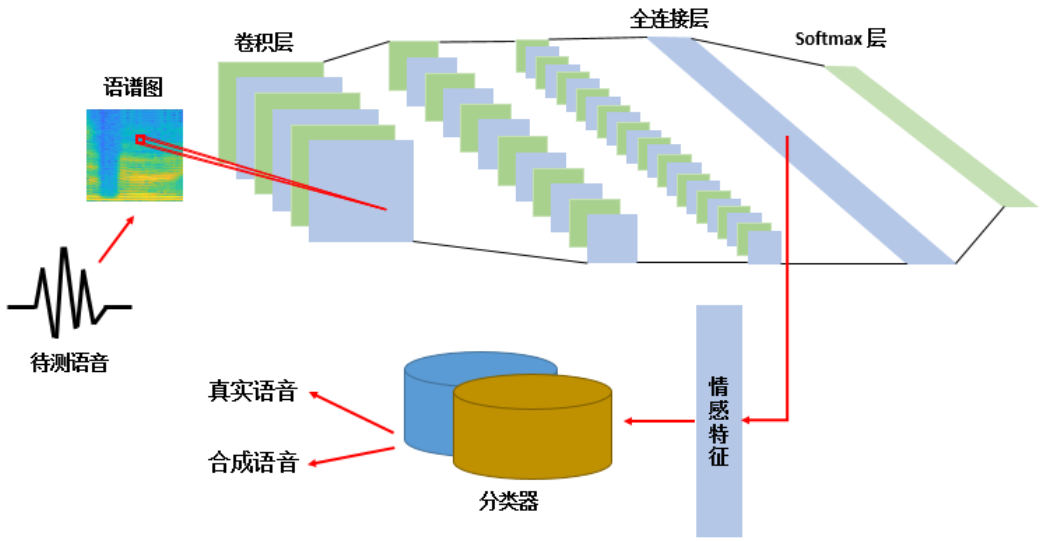


图 5-1　　基于情感特征的合成语音检测算法示意图

5.3 实验

5.3.1 引言

本章中提出的基于情感特征的合成语音检测算法，是合成语音检测研究领域中，首次运用情感信息对合成语音进行检测分析。为了验证算法的可行性，本文使用了Lili Guo等人[39]在2018年提出的情感识别系统中的CNN模型进行了相关实验。实验中，首先使用了IEMOCAP和EmoDB两个情感语音数据库分别对情感模型进行了预训练，然后使用训练好的模型对ASVspoof 2019 LA数据集中的语音数据进行了情感特征的提取，最后实验基于提取得到的情感特征进行了合成语音的检测分析。接下来本节将对实验的详细过程以及得到的实验结果进行介绍。

5.3.2 实验过程及配置

实验中使用的预训练网络为CNN模型，使用CNN模型首先需要对实验中使用到的情感数据库以及合成语音数据库中的语音数据进行数据预处理操作，提取语音所对应的语谱图信息。提取的方法以及涉及的参数配置在第三章中进行了详细的介绍。

在完成语音的语谱图提取后，需要对特征提取时使用的情感识别网络进行预训练。实验中使用的特征提取网络来自于Lili Guo等人在2018年提出的情感识别模型，其模型设计如图 5-2 所示。

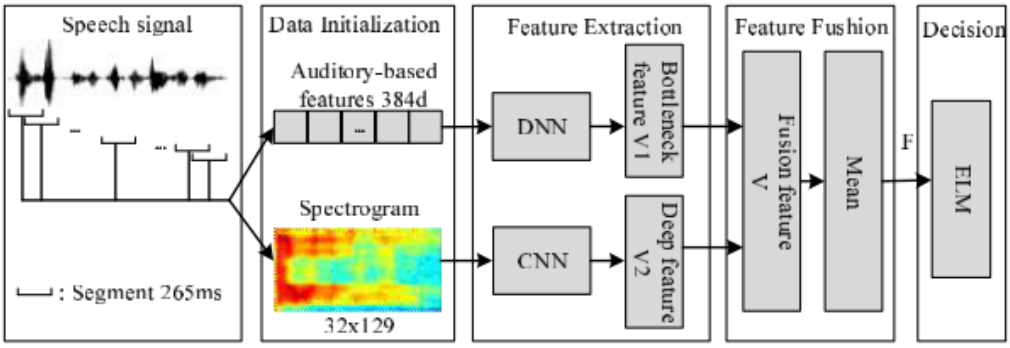


图 5-2 实验中使用的情感识别模型[39]

这一模型中包括DNN和CNN两个特征提取部分，两个部分分别提取对应的情感特征后进行特征融合，再基于融合的特征进行情感识别。由于本文中提

出的方法是基于CNN模型实现的，所以实验中只复现了上述模型中的CNN部分，作为合成语音检测的情感特征提取网络。实验首先使用IEMOCAP情感数据库[40]对模型进行了训练。该数据库是由美国南加州大学采集的一个包含视频、音频及文本信息的多模态英语情感数据库，实验中，只使用了其中的语音信息进行情感网络的训练。训练完成后，实验将ASVspoof 2019 LA数据集中的语音数据依次输入训练得到的网络中，并选取网络中的最后一层全连接层的输出，作为了对应语音的情感特征，这里全连接层的节点数即为特征的维度数。接下来，实验使用了传统的GMM模型作为分类器，对提取到的特征进行了分类测试，GMM模型的配置为ASVspoof 2019挑战赛中的默认配置，模型中的高斯分量数为512，最大迭代次数为10次。

由于，在实验过程中发现使用IEMOCAP数据库训练得到的情感网络在情感识别任务中的准确率要低于使用EmoDB数据库训练的网络，同时考虑到情感特征是一种语言无关的语音信息，于是，实验又尝试了使用EmoDB情感数据库[41]对情感特征提取网络进行预训练。EmoDB数据集是由德国柏林工业大学采集的一个德语情感数据库。实验中，除了将情感网络预训练时使用的数据由IEMOCAP数据库中的英语语音替换为EmoDB数据库中的德语语音外，其余的实验过程和配置均与之前使用IEMOCAP数据库时相同。通过这一对比实验，既可以研究语种对情感特征的影响，还能够分析不同的特征提取网络对合成语音检测算法的影响。

5.3.3 实验结果及分析

实验首先使用IEMOCAP数据集进行了预训练，这里一共进行了10次预训练，得到了10个不同的预训练网络模型。然后，对于每个不同的预训练网络模型，分别提取对应的情感特征。由于GMM模型使用的是EM迭代的方法，在训练时得到的模型是局部最优解，这样会导致使用GMM模型时，几次实验的结果可能不同。为了解决这一问题，本文对每个预训练模型提取到的特征，均进行了三次GMM模型的训练，取三次中最好的结果作为该预训练模型的测试结果。表 5-1 所示为使用10个预训练模型提取的情感特征在ASVspoof 2019 LA数据集中分别得到的测试结果以及10个模型的平均结果，其中模型准确率表示模型在IEMOCAP数据库中情感识别的准确率。

观察表 5-1 中的数据可以发现，10次实验的结果中开发集的EER均集中在37%到42%之间，而测试集中的EER除一次超过了43%以外，其余均集中在41%附近，这些结果并没有随机出现，可以说明本章中提取的情感特征在合成语音检测任务中是具有一定鉴别效果的。

表 5-1 使用IEMOCAP数据库进行预训练的实验结果

实验编号	情感模型准确率 (%)	开发集		测试集	
		EER(%)	t-DCF	EER(%)	t-DCF
1	43.966	39.60	0.9845	41.09	0.9729
2	42.929	40.11	0.9548	40.45	0.9731
3	42.653	40.19	0.9996	43.79	0.9870
4	44.761	39.21	0.9999	40.63	0.9999
5	44.300	40.57	0.9998	41.24	0.9815
6	42.686	39.92	0.9936	41.03	0.9590
7	43.017	38.34	0.9531	40.52	0.9556
8	44.572	42.70	0.9989	41.82	0.9603
9	44.263	37.94	0.9494	40.45	0.9689
10	43.397	42.27	0.9997	41.97	0.9753
平均	-	40.09	0.9833	41.29	0.9733

此外，对比在开发集和测试集中的实验结果可以发现，使用情感特征训练得到的模型在开发集和测试集中表现出的性能差异较小。其中，在第8次和第10次实验中，甚至出现了测试集EER低于开发集EER的情况。综合分析10次实验的结果，开发集的平均EER为40.09%，测试集的平均EER为41.29%，与开发集相比，测试集的平均EER仅提高了2.99%。而传统的LFCC特征测试集EER相比于开发集提高了198.52%，CQCC特征则提高了2125.59%。这些数据均表明基于情感特征的合成语音检测方法具有较好的泛化能力，在面对未知的攻击类型时也可以表现出更高的鲁棒性。

但是，观察实验结果可以发现，使用情感特征训练得到的模型在EER和t-DCF两项指标上均与使用传统语音特征得到的模型存在较大的差距。这里考虑造成这一结果的原因为情感特征提取网络的识别准确率较低，实验中5次预训练得到的模型在情感识别任务中准确率均低于50%，这可能导致提取到的特征不能很好地反映语音中的情感信息。为了验证这一想法，实验又使用EmoDB数据库对模型进行了训练，得到的实验结果如表 5-2 所示。

表 5-2 使用EmoDB数据库进行预训练的实验结果

实验编号	情感模型准确率 (%)	开发集		测试集	
		EER(%)	t-DCF	EER(%)	t-DCF
1	67.684	31.79	0.9401	35.42	0.9601
2	67.467	32.89	0.9584	37.24	0.9828
3	67.024	30.66	0.8011	39.18	0.9570
4	66.822	26.37	0.6897	35.84	0.9421
5	67.474	29.28	0.8539	30.41	0.8832
6	68.667	25.48	0.7395	31.83	0.8883
7	67.437	29.09	0.8278	33.01	0.9042
8	67.857	28.81	0.8647	32.47	0.9006
9	67.692	26.45	0.7269	32.35	0.8815
10	68.029	30.10	0.8734	35.34	0.9426
平均	-	29.09	0.8375	34.31	0.9242

观察表 5-2 中的数据可以发现，使用EmoDB数据库中的德语语料进行训练得到的实验结果在开发集和测试集中的EER同样保持在了一个比较稳定的范围内。这说明使用德语数据进行训练的情感模型提取到的情感特征，对于合成语音检测任务同样具有一定的识别效果，语种的差异并没有对合成语音检测任务产生影响，即本章中所提出的情感特征是一种与语种无关的语音特征。

此外，对比表 5-1 与表 5-2 中的结果可以发现，使用EmoDB数据库训练得到的情感模型，在情感识别任务的准确率方面与使用IEMOCAP数据库训练的模型相比有了一定的提升，识别的准确率均超过了60%。使用经过EmoDB数据库预训练得到的模型提取的情感特征在合成语音检测任务中的检测效果与使用IEMOCAP数据库时相比有了明显的提升，开发集和测试集中的EER都有了明显的下降。因此，可以推测情感模型的准确率对模型提取的情感特征在合成语音检测任务中的识别效果有直接的影响。为了更好地观察这一规律，本文将上述的两组实验按照情感模型的准确率由高到底重新进行了排序，图 5-3 和图 5-4 所示为重新排序后两组实验的EER指标变换规律。

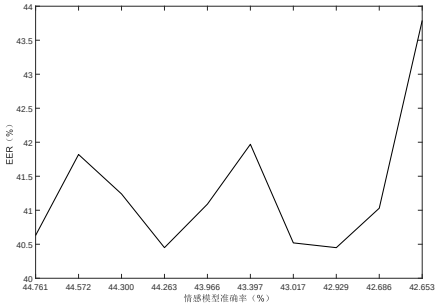


图 5-3　使用IEMOCAP数据库预训练得到的情感特征在ASVspoof 2019 LA 测试集中的EER指标随情感模型准确率的变化曲线

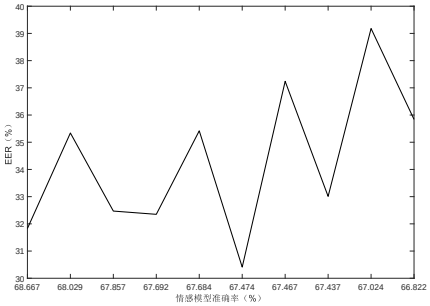


图 5-4　使用Emodb数据库预训练得到的情感特征在ASVspoof 2019 LA 测试集中的EER指标随情感模型准确率的变化曲线

观察图 5-3 和图 5-4 可以发现，使用相同数据库训练的不同模型中，随着模型在情感识别任务中准确率的降低，使用这些模型提取到的特征在合成语音检测任务中EER指标也都呈现出了波动上升的趋势。这表明本章中提出的基于情感特征的合成语音检测算法的检测效果对于使用的情感特征提取模型有较大的依赖性。当情感模型的识别准确率较高时，提取到的情感特征在合成语音检测任务中的表现也较好，反之，表现则较差。目前，在情感识别任务中，模型的准确率仍未达到较高的水平，因此，在一定程度上限制了基于情感特征的合成语音检测算法的表现。

5.4 本章小结

本章中首次提出了一种基于情感特征的合成语音检测算法，该方法借鉴了迁移学习的思想，首先使用预训练的情感识别网络进行特征提取，再利用这些特征进行合成语音检测模型的训练，最后实现判断语音真实性的目的。通过使用IEMOCAP和EmoDB两个情感数据库进行实验，验证了方法的有效性，以及在面对未知攻击方法时的鲁棒性。但是，研究同样发现这一方法的准确性很大程度上受限于情感模型的准确性。今后随着情感识别领域研究的不断进步，情感模型的识别准确性不断提升，基于情感特征的合成语音检测算法将会表现出更好的识别效果。

第 6 章 总结与展望

6.1 总结

随着深度学习领域的不断发展，语音合成技术以及语音转换技术都取得了巨大的进步。许多不法分子开始利用这些技术伪造他人的声音，攻击ASV系统，以盗取他人的信息和钱财。这无疑给ASV系统带来了非常严重的安全隐患。为了更好地应对这些问题，提升ASV系统的安全性，本文对合成语音检测任务进行了深入的研究，并提出了两种相关的解决思路。

首先，本文从音素的角度出发，对合成语音与真实语音在每个音素中的差异进行了分析和研究。目前常用的语音合成技术中，有很多方法是以音素为单位进行拼接和调整，来生成合成语音的。这种做法势必会导致合成语音与真实语音在每个音素中存在一定的差异化信息，而传统的合成语音检测算法，通常难以有效地利用这些隐藏在音素中的有效信息。因此，本文提出使用音素级的F-Ratio分析方法，对合成语音和真实语音进行对比分析，找出每个音素在频域中的差异化信息分布情况。接下来再根据语音中各个音素出现的概率，对分析出的不同音素的信息分布情况进行汇总，进而得到一个综合考虑各个音素的合成语音与真实语音在频域中的差异化信息分布情况。之后，根据信息的分布情况，可以对语音特征提取过程中的滤波器设计进行优化，在信息分布较为集中的频段上，使用更加精细的滤波器进行信息采集，同时加大这些频段中滤波器的密度，进而更好发掘出合成语音中的差异性信息。反之，在信息分布较为稀疏的频段上，则采用较大的滤波器进行滤波，并适当的减小滤波器的个数。除了对滤波器的分布和大小进行改进外，本文还尝试将了常用的三角滤波器更换为矩形滤波器，以达到均匀提取频域信息的目的。在完成滤波器的优化后，便可以使用重新设计的滤波器进行语音特征的提取，从而得到一种适合于合成语音检测的音素适应频率倒谱系数。

根据上述思路，本文在ASVspoof 2019 LA数据集中进行了实验，首先对训练集中的语音数据进行了包括文件格式转换、提取文本信息以及进行音素对齐等在内的数据预处理操作。这些预处理操作可以帮助我们得到训练集中语音的音素信息。之后，实验依照上述思路，使用音素级的F-Ratio分析方法对ASVspoof 2019 LA数据集训练集中的音素信息进行了分析，并根据分析的结

果调整了滤波器的设计。最后，实验分别使用了GMM模型和ResNet模型作为后端，在ASVspoof 2019 LA数据集中的开发集和测试集中对比了新特征与传统特征之间的效果差异。实验结果显示，使用音素级F-Ratio分析方法得到的音素适应频率倒谱系数，在合成语音检测任务中，具有良好的识别效果。无论是使用GMM模型还是ResNet模型作为后端分类器，新特征的EER和t-DCF两项指标与传统的LFCC特征和CQCC特征相比均有了明显的改进，提升比例超过了50%。其中，最好的实验结果甚至优于ASVspoof 2019挑战赛中官方公布的最佳单系统结果。

此外，目前多数的语音合成技术，在生成语音时，通常只是基于文本信息进行合成，并没有考虑到语音中所蕴含的情感信息。因此，合成的语音往往存在缺乏情感的问题。考虑到合成语音的这一特点，本文还提出了一种基于情感特征的合成语音检测算法。算法借鉴了迁移学习的思路，使用预训练的情感识别模型作为特征提取网络，选取模型的最后一层全连接层的输出作为情感特征，并用来训练合成语音检测模型，最后将待测语音用相同的方法提取情感特征，输入到检测模型中，便可以判断语音的真实性。

为了验证上述方法的可行性，本文分别使用IEMOCAP和EmoDB两个情感语音数据库对情感模型进行了预训练，模型训练完成后，使用得到的模型对ASVspoof 2019 LA数据集中的语音数据进行了情感特征的提取，并使用训练集中提取的特征对GMM模型进行了训练。对比多次试验的结果可以发现，使用情感特征训练的合成语音检测模型，对于合成语音的攻击具有一定的鉴别效果，且此方法相比于使用传统语音特征的识别方法，具有较强的泛化能力，在面对测试集中未知类型的攻击时，模型仍可以保持与开发集中相当的检测性能。但是，实验中还发现这一方法的性能对情感模型的识别准确性依赖较高。当情感模型识别准确性较高时，提取到的情感特征在合成语音检测任务中的性能也较好。反之，当情感模型识别准确率较低时，情感特征在合成语音检测任务中的表现也较差。目前，情感识别任务的准确率相对较低，在一定程度上限制了上述方法在合成语音检测任务中的表现。

6.2 展望

针对本文提出的两种方法，后续的研究方向主要有以下两点：

首先，根据本文提出的基于音素级的F-Ratio分析方法得到的PAFCC特征。在特征提取的过程中，需要根据F-Ratio分析得到的差异化信息分布情况，人工地对滤波器进行设计，具有一定的不便性。在后续的研究中，可以对这一部分进行一定的改进，设计出相关的自动生成滤波器的算法，提高本方法的可用性。

其次，是本文提出基于情感特征的合成语音检测算法。实验显示，该方法的性能对使用的情感识别网络的准确性依赖程度较高。目前，实验中使用的情感识别模型所达到的效果，还不足以满足合成语音检测任务的需求。在后续的研究中，可以尝试使用更加复杂的情感识别网络，训练出更加准确的情感识别模型，以提高本文提出的基于情感特征的合成语音检测算法的性能。

参考文献

[1] Masuko T, Hitotsumatsu T, Tokuda K, et al. On the security of HMM-based speaker verification systems against imposture using synthetic speech[C]. In Sixth European Conference on Speech Communication and Technology, 1999.

[2] Patrick P, Aversano G, Blouet R, et al. Voice forgery using ALISP: indexation in a client memory[C]. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., 2005: I–17.

[3] De Leon P L, Pucher M, Yamagishi J, et al. Evaluation of speaker verification security and detection of HMM-based synthetic speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20 (8): 2280–2290.

[4] Kinnunen T, Wu Z-Z, Lee K A, et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech[C]. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012: 4401–4404.

[5] Satoh T, Masuko T, Kobayashi T, et al. A robust speaker verification system against imposture using an HMM-based speech synthesis system[C]. In Seventh European Conference on Speech Communication and Technology, 2001.

[6] Ogihara A, Unno H, Shiozaki A. Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification[J]. IEICE transactions on fundamentals of electronics, communications and computer sciences, 2005, 88 (1): 280–286.

[7] Leon P L D, Stewart B, Yamagishi J. Synthetic speech discrimination using pitch pattern statistics derived from image analysis[C]. In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[8] Quatieri T. Discrete-time speech signal processing principles and practice. Prenticehall. 2002.

[9] De Leon P L, Hernaez I, Saratxaga I, et al. Detection of synthetic speech for the problem of imposture[C]. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011: 4844–4847.

[10] De Leon P L, Pucher M, Yamagishi J, et al. Evaluation of speaker verification security and detection of HMM-based synthetic speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20 (8): 2280–2290.

[11] Wu Z, Chng E S, Li H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition[C]. In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[12] Wu Z, Kinnunen T, Chng E S, et al. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case[C]. In Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 2012: 1–5.

[13] Alegre F, Vipperla R, Evans N. Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals[C]. In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[14] Wu Z, Kinnunen T, Evans N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]. In Sixteenth Annual Conference of the International Speech Communication Association, 2015: 2037–2041.

[15] Kinnunen T, Sahidullah M, Delgado H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[J]. INTERSPEECH, 2017.

[16] Todisco M, Wang X, Vestman V, et al. Asvspoof 2019: Future horizons in spoofed and fake audio detection[J]. arXiv preprint arXiv:1904.05441, 2019.

[17] Patel T B, Patil H A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech[C]. In Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[18] Li Q. An auditory-based transfrom for audio signal processing[C]. In 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009: 181–184.

[19] Sahidullah M, Kinnunen T, Hanilçi C. A comparison of features for synthetic speech detection[J], 2015.

[20] Todisco M, Delgado H, Evans N W. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients.[C]. In Odyssey, 2016: 283–290.

[21] Kinnunen T, Lee K A, Delgado H, et al. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification[J]. arXiv preprint arXiv:1804.09618, 2018.

[22] Gomez-Alanis A, Peinado A M, Gonzalez J A, et al. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection[J]. Proc. Interspeech 2019, 2019: 1068–1072.

[23] Lai C-I, Chen N, Villalba J, et al. ASSERT: Anti-Spoofing with squeeze-excitation and residual networks[J]. arXiv preprint arXiv:1904.01120, 2019.

[24] Suthokumar G, Sriskandaraja K, Sethu V, et al. Phoneme specific modelling and scoring techniques for anti spoofing system[C]. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 6106–6110.

[25] Sahidullah M, Delgado H, Todisco M, et al. Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015[J], 2016.

[26] Todisco M, Delgado H, Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017, 45: 516–535.

[27] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770–778.

[28] Matrouf D, Bonastre J-F, Fredouille C. Effect of speech transformation on impostor acceptance[C]. In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006: I–I.

[29] Morise M, Yokomori F, Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE TRANSACTIONS on Information and Systems, 2016, 99 (7): 1877–1884.

[30] Oord A v d, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[31] Griffin D, Lim J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32 (2): 236–243.

[32] Tanaka K, Kaneko T, Hojo N, et al. Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks[C]. In 2018 IEEE Spoken Language Technology Workshop (SLT), 2018: 632–639.

[33] Bigi B. SPPAS-multi-lingual approaches to the automatic annotation of speech[J]. The Phonetician, 2015, 111 (112): 54–69.

[34] Bigi B, Meunier C. Automatic segmentation of spontaneous speech[J], 2018.

[35] Lu X, Dang J. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification[J]. Speech communication, 2008, 50 (4): 312–322.

[36] Hyon S, Dang J, Feng H, et al. Detection of speaker individual information using a phoneme effect suppression method[J]. Speech Communication, 2014, 57: 87–100.

[37] Alzantot M, Wang Z, Srivastava M B. Deep residual neural networks for audio spoofing detection[J]. arXiv preprint arXiv:1907.00501, 2019.

[38] Eyben F, Scherer K R, Schuller B W, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE transactions on affective computing, 2015, 7 (2): 190–202.

[39] Guo L, Wang L, Dang J, et al. A feature fusion method based on extreme learning machine for speech emotion recognition[C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 2666–2670.

[40] Busso C, Bulut M, Lee C-C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42 (4): 335.

[41] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]. In Ninth European Conference on Speech Communication and Technology, 2005.

关于国际工程师学院人才培养模式情况说明

天津大学国际工程师学院借鉴法国工程师培养理念和模式，结合我国本土教育特色，积极探索工程教育改革路径，以多学科交叉融合为特色，聚焦科学技术前沿领域，服务国家重大战略需求，是天津大学研究生层面的工程教育改革试验区。

2017年，国际工程师学院通过法国工程师职衔委员(CTI)授予的最高等级六年期认证和欧洲工程教育(EUR-ACE)专业认证。这标志着学院的办学模式、人才培养质量得到国际认可，学院毕业生也将获得由法国工程师职衔委员会授权颁发的法国工程师文凭。

国际工程师学院全面推进课程改革，总课时约2000学时，重新配置理论课、练习课和实践课比例，使得实践学时(练习和实践课)占到总学时的 2/3。教学内容与企业工程实际项目深度融合，累计10个月三阶段的渐进式实习，实现学校教育与企业需求的“无缝衔接”。

研究生在学期间着力加强工程实践能力、创新能力和解决实际问题能力的锻炼和提升，不强制要求发表学术论文。该培养体系和学位申请标准已于天津大学学位评定委员会第97次会议审议通过。特此说明。

发表论文和参加科研情况说明

（一）发表的学术论文

[1] 李鹤群，方强，路文焕，魏建国，陈云，苏志华，刘畅等. 基于MRI的发音模型研究[C]. //第十四届全国人机语音通讯学术会议（NCMMSC' 2017）论文集. 2017.

[2] Chang Liu, Jianguo Wei, Hui Feng, Wenhuan Lu, Junhai Xu, Yuqing He, Jiayu Jin, Meng Liu and Zetian Li. A Phoneme Difference Adaptive Feature for ASV Spoofing Logical Access Detection. Interspeech, 2020. (Submitted, Under review).

（二）申请及已获得的专利

[1] 基于F-Ratio分析的抑制音素影响的合成语音检测方法，申请号：202010572748.4，发明人：魏建国，刘畅。

（三）参与的科研项目

[1] 国家重点研发计划课题，No.2018YFC0806802，刑事执行监督多源异构信息自动提取、分析匹配和信息交换关键技术与装备，2018-2020.

致　　谢

在整个毕业设计过程中，困难和迷茫不时困扰着我，是导师和同学们的耐心帮助和指导，让我得以有信心坚持下去，不再迷茫。在此，我要特别感谢那些在我三年的研究生生活中给予我帮助的人。

首先，是我的指导老师——魏建国老师。在我的三年的研究生学习和生活中，魏老师严谨认真的学术态度深深地影响了我，他在学习和生活中的各个方面都给予了我极大的帮助。在学习方面，老师孜孜不倦、严谨治学的态度将成为我毕生奋斗的信条。在生活中，魏老师对我和蔼可亲，像是亲切的家人和朋友，给了我很多的关心和帮助，让我在学校能够感受到家庭般的温暖。在此，我真诚地向魏老师致以深深的敬意。同时，还要感谢路文焕老师，是她不停在督促我的科研进展，又时刻引导着我，给我的科研带来不竭的灵感，其给予的精神财富将使我受益终生。

在这里我还要由衷的感谢天津大学外国语与文学学院的冯卉老师。在进行语音发音的相关研究过程中，冯老师不厌其烦地对我提出的困惑加以讲解，在整个科研过程中，提供了非常多宝贵的意见。同时在书写英文论文的过程中，冯老师在语言的运用上帮我进行了细致的修改，帮助我在科研道路上迅速成长。正是由于有冯老师的指导，我才能顺利完成论文的写作，在开拓思维方式的同时，也提高了我解决问题的能力。

另外也要感谢实验室中为我提供帮助的师兄师姐，尤其是李鹤群师兄、苏志华师兄和白国臣师兄，在我刚刚接触科研工作时，是他们细心地为我提供帮助，将他们科研过程中的经验分享给我，让我少走了很多的弯路。还要感谢与我进行相似课题的李泽田同学、张琳、靳嘉宇同学、刘猛同学和李聿轩同学，与他们在一起时的讨论，往往可以碰撞出非常美妙的火花，在我的科研没有灵感时，帮助我开拓了思路。

此外还要感谢我身边的小伙伴们，比如李聿轩同学、韩智丞同学、魏潇同学和王善宇同学等。在生活中大家互相帮助，一起度过了三年的美好时光。十分感谢大家一直以来对我的鼓励和支持。

最后，也是最重要的，要由衷的感谢我的父母和家人，是他们长期以来的支持和理解，让我可以认真地投入到科研工作中。在今后的工作生活中，我将继续带着他们的期望继续前行，积极向上地投入到未来，不断进步！