基于多传感器多源的鲁棒语音处理信息融合——最新技术概述

Parham Aarabi a，\*，Belur V。

2003年12月3日，加拿大安大略省多伦多市国王学院路10号多伦多大学电气与计算机工程学士学位，

地址：M5S3G4 b Huntsville，AL，USA；2004年2月3日收到订正表格；2004年2月3日接受

摘要

本文概述了鲁棒语音处理的最新进展，并描述了信息融合在进一步实现其目标中的作用。此外，它还具有传统特刊客座社论的功能，对特刊内容进行简要介绍。

@2004爱思唯尔公司版权所有。

关键词：稳健语音处理；多传感器信息融合；语音识别

1.简介

几十年来，人们对使计算机能够处理（即定位、增强和/或识别）语音进行了广泛的研究[1–4,16–20,33]。在此期间，在自动语音识别和处理方面取得了许多进展，产生了各种商业语音识别系统。这些先进的语音识别系统通常通过大量的训练和计算获得较高的识别准确率。然而，高准确率在存在噪声（例如二次对话）的情况下显著降低。因此，噪声环境下的鲁棒语音识别一直是一个热门的研究课题，也是一个尚未解决的挑战[16,18,20,21]。单麦克风语音识别的最新进展（包括概率语音增强和噪声建模技术，如[16,21]）在解决当前语音识别系统的一些鲁棒性问题方面取得了很大进展。但是，使用一个麦克风可以实现的功能是有限的。采用多传感器（包括摄像机和麦克风）的系统具有更强的鲁棒性和更高的识别准确率。融合多个传感器或多个信息源进行鲁棒语音处理是本专题研究的重点。

信息融合方法在语音处理中的应用已经在过去进行了探索[4,6,8]。例如，在文献[4]中，我们使用解释性融合范式探索和分析了一个视听元音识别系统。这项工作不仅说明了融合多模态系统的优点，而且还说明了各个模态对融合结果的相对影响。例如，当信噪比为24分贝时，视觉模态对声学模态的相对影响要大得多，以便做出正确的判断。这清楚地说明了信息融合的好处：在不同的情况下，某些信息源可能是不可靠的，只要存在一种适当的方法对这些信息源进行融合，不可靠的信息源的相对影响将小于那些更可靠的信息源。类似的视听作品支持这项权利要求[4-8,23,34-39]。

2.文献综述

多年来，视听信息融合一直是一个广泛研究的领域。在[25]中，采用了一种混合的视听特征融合和视听决策融合方法来实现鲁棒的语音识别。在[26]中，一个耦合的隐马尔可夫模型（这是动态信念网络的特例）被用于语音识别的视听信息融合。在[27]中，一种加权判决融合方法被用于视听语音识别。人们还提出了多种其他视听语音识别系统，将麦克风的信号与摄像机记录的图像融合在一起[28–32]。在所有这些系统中，与视觉或声学模态本身相比，融合的视听系统导致字错误率的显著降低。

尽管进行了所有这些研究，但融合多个传感器获得的信息（其中一些传感器可能具有不同的模式）仍然是一个开放的问题，其本质是一个定义模糊的问题[2,8,22]。首先，有一个问题是究竟应该使用哪种类型的传感器。例如，将相机和麦克风结合起来可以产生更大的鲁棒性，因为相机不受噪声的影响，而麦克风不受影响相机的事件（如物体、障碍物、照明等）的影响。然而，融合不同的模式往往更加困难和复杂，这可能导致更大的计算需求。然而，即使存在这种复杂性-有效性的权衡，许多系统都试图融合多模态信息以用于各种应用。示例包括采用单个摄像机和麦克风的视听语音识别系统，从而获得更高的语音识别准确率和更强的噪声鲁棒性[5–15]。其他应用包括视听声音定位[2,3]，其中使用多个摄像头和多个麦克风对扬声器进行视觉定位。正如预期的那样，这些系统说明了与仅音频或仅视频定位系统相比的改进的性能。出现的下一个问题是信息融合的点。例如，由一组麦克风获得的信号值可以在信号电平上被集成（使用诸如波束形成之类的技术），或者，它们可以在每个信道上执行语音识别之后被集成。事实上，这种集成或融合可以发生在语音识别或定位过程中的任何地方。该领域的大部分前期工作都集中在识别过程之前的数据融合上（即，将信号进行整合，得到更清晰的信号，然后用于识别）。这种预识别融合具有计算简单的优点，尽管其最优性不是很清楚。

当使用麦克风阵列时，预识别数据融合是有效的。当相机和麦克风联合使用时，在信号水平上组合它们的数据要困难得多。任何这样的融合都必须涉及到视觉结果到声学结果空间的投影，反之亦然。显然，这个投影可能是一个非常复杂和计算要求很高的任务。然而，早期的视听信息融合[1,11,12]和后期的视听信息融合[14,15]都有尝试。这些例子探讨了复杂性增加和语音识别准确率提高之间的相对折衷。

因此，我们手头的问题是（1）我们融合什么，（2）什么时候（即在识别过程的哪个阶段）我们进行融合，以及（3）我们实际如何进行融合。这是三个相互关联的问题，它们的最终答案可能还需要几年的时间。尽管如此，本期特刊的六篇论文为这些问题提供了一些独特而创新的解决方案。

3.特刊概述

第一篇论文介绍了用于视听语音识别的异步隐马尔可夫模型（HMMs），提出了一种异步HMM结构，它分别应用于从麦克风和摄像机获取的声音和视觉数据。结果表明，该融合模型无论是对视听语音识别还是对纯声语音识别都优于标准的HMM结构，特别是在较低的信噪比下。

第二篇论文还考虑了音视频信息融合问题，对连续音视频数字识别进行决策融合。他们的方法涉及从视觉数据获得的识别结果和从声学数据获得的结果的识别后融合（决策融合）。结果表明，在很低的信噪比（15db）下，它们的系统性能与纯音频系统相似，远优于纯视频系统。在中等信噪比范围（0–15 dB）内，他们的系统性能优于纯音频和纯视频系统。

本文的第三篇论文试图将两个麦克风和两个摄像机融合在信号级，以实现语音增强。这里，与语音相关的视觉数据（即唇部运动）最初被转换为两个相应的声学信号。然后，这些视觉衍生或虚拟声学信号被用作使用从两个麦克风获得的信号进行语音分离的基础.

第四篇论文讨论了一个非常有趣的话题，即在多传感器环境中量化传感器的价值。如果存在多个传感器（如麦克风），则可能不需要将所有传感器用于语音处理应用。但是，为了确定应该使用哪些传感器，必须测量该传感器提供的唯一信息量。为此，本研究提出了一种信息论的方法。

本期特刊的第五篇论文使用了24个麦克风阵列来控制和导航机器人导游。这里，对从24个麦克风收集的数据进行处理，以估计语音源的位置，假设语音源是机器人。这种多传感器的声学定位是用来纠正机器人的位置，并引导它走向下一个目的地。有趣的是，将第五篇论文与Enzo等人[24]提出的系统进行比较。与[24]中的技术不同，机器人上的麦克风阵列用于人-操作员定位和改进人-机器人交互，本研究使用放置在环境中的固定阵列。值得注意的是，所提出的声学机器人定位系统实际上是互补的，使得两者的融合成为未来工作中一个有趣且可能富有成效的方向。

第六篇论文针对鲁棒语音识别中的特征融合问题，提出了一种提高语音识别准确率的新方法。它们不依赖于多个麦克风或视听信息融合，而是依赖于由声学特征（Mel频率倒谱系数）和发音特征组成的两个信息源。

4.结论意见

实现一个健壮而准确的语音识别系统的最终目标需要对定位、增强和识别语音的有效计算方法进行广泛的研究。就像我们人类用两只耳朵和两只眼睛进行语音定位和识别一样，人工语音识别系统也需要多模态和多传感器的信息融合，才能在实际情况下获得准确的结果。因此，实现这样的系统需要对融合从多个麦克风和照相机以及其他传感器/源获得的信息进行广泛的研究。本期特刊的六篇论文为实现这一最终目标提供了一些相当独特和积极的步骤；最终的目标是有一天使强大的语音识别能够应用于汽车、电器、平板电脑、便携式电脑、个人数字助理和残疾人。