# 受说话人变化检测启发的口语语言变更检测

Jagabandhu Mishra[1], [a] 和 S. R. Mahadeva Prasanna[1], [b]

*Department of Electrical Engineering, IIT Dharwad, Dharwad 580011, Karnataka, India.*

（日期：2023年2月13日）

口语语言变更检测（LCD）是指识别语码转换话语中的语言转换。类似地，识别多说话人话语中的说话人转换称为说话人变更检测（SCD）。由于这两项任务在任务层面上相似，为SCD任务开发的架构/框架可能适用于LCD任务。因此，本研究旨在开发受SCD启发的LCD系统。最初，LCD和SCD都由人类执行。研究表明，与SCD相比，人类执行LCD需要（a）在变化点附近更长的时长，以及（b）特定语言的先验暴露。通过增加无监督距离法的分析窗口长度，满足对更长时长的要求。这使合成和实际语码转换数据集上的性能分别相对提升29.1%和2.4%；而先验语言知识则分别带来31.63%和14.27%的相对提升。实际数据集与合成数据集之间的性能差异主要源于单语片段时长分布的差异。

©2023 美国声学学会。[https://doi.org(DOInumber)]

[XYZ]

页码：1–12

arXiv:2302.05265v1 [eess.AS] 10 Feb 2023

## I. 引言

口语语言分段标注（LD）是一项任务，旨在自动切分并标注给定多语言语音信号中的单语片段。现有关于LD 的研究非常少（Sitaram等，2019）。其中大多数采用基于音系搭配（即语音单位的分布）的方法（Chan 等，2004；Lyu 等，2013；Spoorthy 等，2018）。采用基于音系搭配的方法开发 LD，需要转写语音话语。但这类数据难以获取，因为语码转换的多语言话语中包含的大多数语言都是低资源语言（Sitaram等，2019；Spoorthy 等，2018）。虽然已有一些迁移学习方法，通过自适应高资源语言的音系搭配模型来获得低资源语言的模型，但如果两种语言不属于同一语系，最终可能导致性能下降（Sitaram等，2019）。此外，人类进行 LD 毫不费力，尤其是对于熟悉的语言而言，而机器则面临挑战。因此，有必要探索其他 LD 方法。

说话人分段标注（SD）是一项针对给定多说话人话语，自动分割并标注单说话人片段的任务，相关研究在文献中已较为充分。尽管执行语言检测（LD）和说话人分段标注（SD）任务时需要捕获的信息存在差异，但二者也有许多相似之处，例如用于近似声道共振峰的特征已成功用于说话人特定音素和语言特定音素的建模（Carrasquillo 等，2021年；Li等，2013；Liu 等，2021年）。此外，大多数用于口语语言识别（LID）的方法都受说话人识别/验证（SID/SV）任务所采用方法的启发（Richardson 等，2015；Snyder 等，2018）。此外，从 SID/SV 文献中借鉴的大多数成功 LID 系统都不需要转写语音数据（Li等，2013；Snyder 等，2018）。另一方面，采用音系搭配方法开发的 LID 系统则需要转写语音数据。这促使人们对 LD 和 SD 任务之间的紧密关联展开研究，并可能据此提出用于 LD 的方法。

说话人分段（SD）领域的发展主要沿着两条路径展开：（1）先进行变点检测，再进行聚类和边界细化；（2）先对固定时长片段进行分段，再提取 i-vector／嵌入向量，随后进行聚类和边界细化（Moattar and Homayounpour, 2012; Park et al., 2022; Tranter and Reynolds, 2006）。( Bredin et al., 2017; Dawalatabad et al., 2020; Hogg et al., 2019; Park et al., 2022) 报告称，初始变点检测提升了整体 SD 性能。因此，本研究通过比较口语语言变更检测（LCD）与说话人变更检测（SCD），着重研究口语语言变更检测。现有的 SCD 方法大致可分为两类：（1）基于距离的无监督方法；（2）基于模型的有监督方法（Moattar and Homayounpour, 2012; Park et al., 2022）。基于距离的方法对使用连续滑动窗口从语音信号中提取的说话人特定特征进行假设检验，以预测说话人是否发生变化（Moattar and Homayounpour, 2012; Park et al., 2022）。

[a]jagabandhu.mishra.18@iitdh.ac.in[b]
prasanna@iitdh.ac.in

（判定其是否来自同一说话人），以预测说话人变化（Moattar and Homayounpour, 2012; Park et al., 2022）。沿用这一方法，文献中提出了多种特征提取技术，例如激励源（Dhananjaya and Yegnanarayana, 2008; Sarma et al., 2015）、基频轮廓（Hogg et al., 2019）等；同时也提出了多种距离度量方法，例如Kullback–Leibler（KL）散度（Siegler et al., 1997）、贝叶斯信息准则（BIC）（Chen et al., 1998）、KL2（Siegler et al., 1997）、广义似然比（GLR）（Gish et al., 1991）和信息瓶颈（IB）（Dawalatabad et al., 2020）。通常，基于距离的无监督方法在环境变化和背景噪声影响下性能会下降（可能预测出错误的变化），因此，文献提出了基于有监督模型的方法来解决这一问题（Moattar and Homayounpour, 2012; Park et al., 2022）。早期提出的方法使用高斯混合模型—通用背景模型（GMM-UBM）（Barras et al., 2006; Moattar and Homayounpour, 2012）、隐马尔可夫模型（HMM）（Meignier et al., 2006）等对各个说话人进行建模；如今，借助深度学习框架，方法通过区分说话人变更片段（说话人变化点附近的片段）与未发生变化的片段，来预测说话人是否发生变化（Moattar and Homayounpour, 2012; Park et al., 2022）。然而，基于模型的方法会平滑输出证据，可能导致漏检变化点（Moattar and Homayounpour, 2012）。此外，训练有监督模型需要来自相似环境／录音条件、说话风格、语言等的带标签语音数据，这使系统开发变得复杂。因此，基于距离的无监督方法在 SCD 任务中更受欢迎且应用更广泛（Dawalatabad et al., 2020; Moattar and Homayounpour, 2012; Park et al., 2022）。

尽管现有的SCD框架看似易于采用，但实际应用仍面临诸多挑战。图1的(a)和(b)分别展示了对应于发生说话人变化和语言变化的话语的时域语音信号。通过聆听并观察这两段话语的时域表示，

话语中，识别出的说话人/语言变化点均由人工标注。从时域信号中，很难同时定位说话人变化点和语言变化点。图1 （c）和（d）展示了两段话语的频谱图。从频谱图可以观察到，说话人变化附近的共振峰结构有显著变化，而语言变化附近的结构则保持完整。说话人发生变化时，声道系统信息也随之变化，因此共振峰结构会发生变化。然而，在语言变化期间，共振峰频率结构保持完整，因为同一位说话人在说两种语言。有趣的是，人类能够区分不同的口语语言，却不必了解

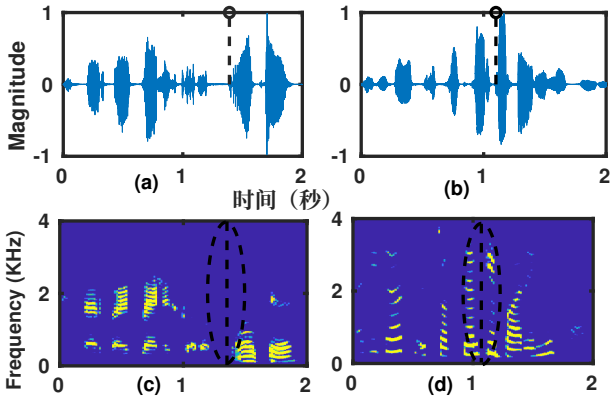


图1。(a)和(c)分别为两位说话人的时域语音信号及其频谱图。(b)和(d)分别为两种语言（双语）的时域语音信号及其频谱图。

各自语言的详细词汇规则和音位分布。当然，人类需要先前接触过这些语言（Li 等，2013）。人类可能会利用长期音位动态特征来区分不同语言。因此，可以通过捕捉长期的语言特定的频谱—时间动态特征来检测语言变化。这可能体现为有效的音位序列及其组合，从而构成某种语言的音节和子词。

基于需要利用长期频谱-时间证据这一点，可以假设，人类/机器进行LCD可能需要比SCD更长的变化点周围邻域时长。此外，LCD还可能受益于先前对相应语言的接触。为验证这一点，开展了一项聚焦于语言/说话人变化检测的主观研究。

为了自动检测语言变化，初步研究采用了现有的无监督距离方法和有监督的基于模型的SCD方法。基于模型的方法包括GMM-UBM、i-vector和x-vector。根据LCD和SCD的实验结果，将对各个框架进行适当的调整，以提升LCD任务的性能。

本研究的主要贡献概括如下：(a) 通过观察说话人变化和语言变化附近的时频表示，我们提出假设：与检测说话人变化相比，检测语言变化需要变化点附近更长时长的语音，以及关于语言的先验知识。人类主观研究也证实了这一假设；(b) 将说话人变化检测框架用作语言变化检测的初始基线，并分析其性能；(c) 进一步改进这些框架，以提升语言变化检测的性能。

## II. 数据库设置

本节简要介绍本研究使用的数据库。为在人类受试者中开展LCD/SCD任务，我们分别选取了 32 段和 15 段话语，用于语言变化和说话人变化研究。所有话语都只有一个语言/说话人变化点，且变化点两侧各有约$6-8$ 个音节。语言变化研究所用的 32 段话语选自公开来源（主要来自YouTube），而说话人变化研究所用的 15 段话语则选自 IITG-MV第三阶段和 DIHARD 数据集（Haris 等，2012；Ryant等，2018）。语言变化研究所用的 32 段话语来自 10 个语言对，数量分别为 4、4、4、4、4、2、2、4、2、2 段，依次对应：(1) 印地语-英语（HIE），(2) 孟加拉语-英语（BEE），(3) 泰卢固语-英语（TEE），(4) 泰米尔语-英语（TAE），(5) 孟加拉语-阿萨姆语（BEA），(6) 孟加拉语-孟加拉语（BEB），(7) 阿萨姆语-阿萨姆语（ASA），(8) 泰米尔语-马拉雅拉姆语（TAM），(9) 泰米尔语-泰米尔语（TAT），以及 (10) 马拉雅拉姆语-马拉雅拉姆语（MAM）。很难找到由同一位说话人说出的孟加拉语-阿萨姆语和泰米尔语-马拉雅拉姆语语言对话语。因此，这些语言对中有语言变化和无语言变化的话语均与说话人变化一并纳入考虑。LCD 和 SCD任务所选话语及其变化点标注可在 https://github.com/jagabandhumishra/HUMAN-SUBJECTIVE-STUDY-FOR-LCD-and-SCD 获取。

最初的研究使用合成生成的语码转换和多说话人话语开展。为生成这些话语，我们使用了印度理工学院马德拉斯分校的文本转语音（IITM-TTS）语料库（Baby等，2016）。IITM-TTS语料库包含来自13种印度语言母语说话人的语音数据录音。每种母语均由两位说话人（一男一女）分别用其母语和英语录制话语。本研究合成语码转换话语时，选用一位以印地语为母语、英语为第二语言的女性说话人。每种语言的前5小时数据用于训练。其余单语话语则随机拼接，用于生成语码转换话语。总计生成了4000条话语，每条包含1至5个语言变化点。生成的印地语和英语语码转换话语中，单语片段的平均时长分别约为6。5秒和5。2秒。生成的数据集称为TTS女性语言变化（TTSF-LC）语料库。同样，为了在语言保持一致的情况下生成说话人变化话语，我们使用了以印地语和阿萨姆语为母语的女性说话人录制的英语语音话语。生成的话语中，单说话人片段的平均时长分别为5。19秒和4。86秒。生成的数据集称为TTS女性说话人切换语料库（TTSF-SC）。

最后，为了验证所得观察结果的普适性，我们在标准语言变化检测语料库上进行了实验。实验使用了微软代码切换挑战任务B（MSC-STB）数据集。该数据集包含开发分区和训练分区，其中包括三种语言对的语码转换话语及语言标签（每段200毫秒）：古吉拉特语-英语（GUE）、泰米尔语-英语（TAE）和泰卢固语-英语（TEE）。训练集和开发集中的每种语言时长分别约为16小时和2小时。数据库详情见（Diwan等，2021年）。

## III. 语言变化检测与说话人变化检测的人类主观研究

我们设计了一项实验流程，让每位人类受试者听取一组话语，其中可能存在语言/说话人变化，也可能不存在。受试者需要标记是否存在语言/说话人变化。话语分为五组。每组以真实/虚假变化点附近选取的有声帧数量（NVF）表示其近似时长。真实变化点指所选话语中实际发生变化的位置。我们在变化点附近切分所选话语，以生成单语言/单说话人语句。虚假变化点表示给定单语言/单说话人语句中居中有声帧的起始位置。有声帧通过以下方式确定：将给定话语的短时帧能量平均值（帧长为20毫秒，帧移为10毫秒）的6%设为阈值（Rabiner，1978年）。我们通过在真实变化点附近切分所选的15条话语，生成了30条单说话人话语。其中，依据时长选取最长的15条用于本研究。我们采用相同流程，使用属于HIE、BEE、TAE和TEE语言对的所选语码转换话语生成单语言语句。不过，属于BEA和TAM的语句是例外，因为这些话语除了语言变化外还存在说话人变化。因此，为了进行公平比较，我们合成了这些情况下的单语言语句，使其也包含说话人变化，即分别为BEB、ASA、MAM和TAT。随后，通过对真实/虚假变化点左右两侧的 $S(n$个有声帧进行处理，来掩蔽每条话语 $x$ （NVF-$x$）。根据 $x$的取值，掩蔽后的话语分为五组，分别称为NVF-10、NVF-20、NVF-30、NVF-50和NVF-75。为避免掩蔽突变，将参数适当的高斯掩码 $G(n)$与话语相乘，以获得掩蔽话语 $S_m(n) = S(n) \times G(n)$。掩蔽信号经过基于能量的端点检测算法处理，以获得最终的掩蔽话语（Rabiner，1978年）。掩蔽话语生成的详细流程见补充材料[1]，，生成的话语也

可在
https://github.com/jagabandhumishra/HUMAN-SUBJECTIVE-STUDY-FOR-LCD-and-SCD获取。

听音实验共有18名受试者参加，其中男性13名，女性5名。所选受试者的年龄为20 – 30 岁。受试者此前均未接触过本研究中所用说话人的语音样本。不过，受试者都熟悉英语，而对其他语言的熟悉程度则各不相同。为了解受试者对语言的熟悉程度，我们请每位受试者针对每一对语言，提供一个从零到三的语言熟悉度评分（LCS）。

听音研究使用了390段话语（即240段用于语言变化检测，150段用于说话人变化检测）。语言变化检测任务与说话人变化检测分开进行，因此分两个不同场次开展；同时，受试者在状态良好、充分休息的情况下参与，以避免听者疲劳。研究设计了一个图形用户界面（GUI），用于开展听音研究。在特定的语言变化检测/说话人变化检测研究中，所有掩蔽话语均以随机顺序呈现给听者，不受片段时长影响。如果听者仅听一遍后无法作答，则允许其重复播放话语多次。我们的目标是观察人们通过聆听来自五个不同组别的话语，能够多准确地识别说话人和语言的变化。因此，此处采用了（Sharma 等，2019）中记录的回答，以分析人们检测说话者变化的能力。记录了三类回答，分别是：（1）是否检测到语言/说话人变化；（2）重放次数（NR）；（3）反应时间（RT）。反应时间是指受试者听完整段话语后作答所用的时长。反应时间的计算方法是从总时长（TD）中减去相应的话语时长（UD）（即 $RT = TD - UD$）。总时长是指受试者作答所用的时长（即从按下播放按钮到按下“是/否”按钮）。

$$DER = \frac{(FA + FR)}{N} \tag{1}$$

对于给定的受试者，本研究计算了三种表现指标：(1) 平均检测错误率 ($DER$) (2) 平均重放次数 ($NR$)，以及 (3) 平均反应时间 ($RT$)。$DER$ 的定义见公式 1，其中 $N$ 为试次总数，$FA$ 为被受试者标记为正确的误判为语言/说话人变化的话语数量，以及

$FR$ 分别表示受试者将真实的语言/说话人变化话语误标为未发生变化的话语数量。$DER$指标定义了受试者检测语言/说话人变化的能力不足程度。$NR$ 用于估算受试者能够舒适地标记其反应所需的平均重播次数。同样，$RT$用于估算受试者听取相应话语后感知到语言/说话人变化所需的平均时长。性能指标的值越高，表示能力不足程度越高

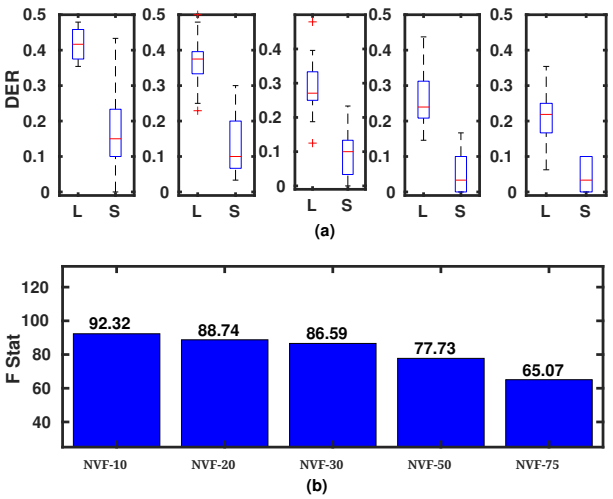


图 2. (a) 受试者的 $DER$ 分布；(b) 语言变化检测（LCD，L）和说话人变化检测（SCD，S）研究中$DER$ 分布的方差分析检验 F 统计量（F Stat）值。

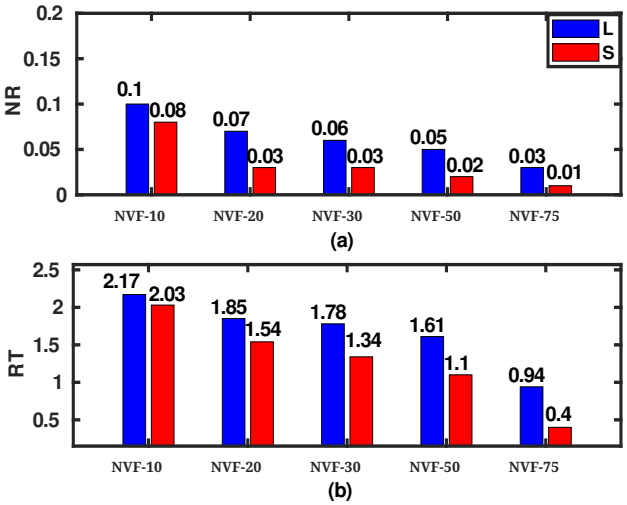


图3. LCD和SCD的(a) $NR$ 和(b) $RT$ 分布的中位数。

人类受试者感知语言/说话人变化，反之亦然。

完成语言变化检测和说话人变化检测实验后，针对 NVF 计算受试者个体的 $DER$、$NR$和 $RT$。图中展示了所得 $DER$ 相对于 NVF 的分布。2(a) 可以看出，$DER$ 值在说话人变化检测中低于语言变化检测，无论 NVF 取何值。这表明，人类受试者检测说话人切换比检测语言变化更轻松。此外，随着 NVF 从 10 增至 75，说话人变化检测和语言变化检测的 DER 均有所下降。

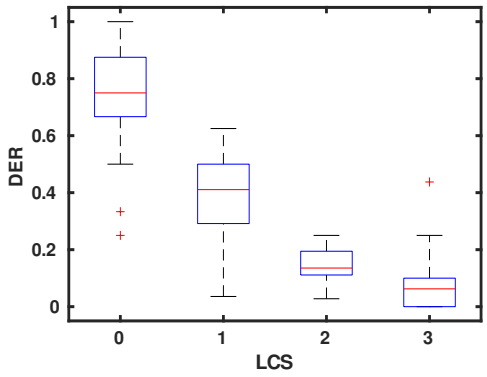


图 4. 使用 NVF-50 和 NVF-75 的语言变化检测中 DER 与语言熟悉度评分（LCS）的关系。

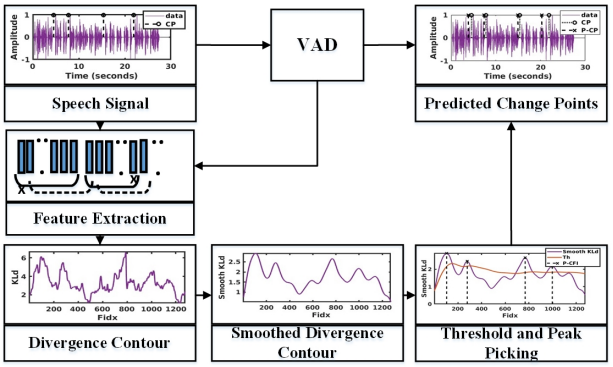


图 5. 无监督距离法变化检测框架的基本框图

随着 NVF 增加，语言变化检测和说话人变化检测的 DER 分布之间的差异逐渐减小。这表明，随着 NVF 增加，人类受试者检测语言变化时的轻松程度有所提高，并逐渐接近检测说话人变化时的水平。为进一步验证这一事实，我们对$DER$ 进行了方差分析（ANOVA）统计检验，比较语言变化检测和说话人变化检测的分布（去除离群值后）。所得 F统计量值如图2(b) 所示。F统计量值越高，表明两个分布的区分能力越强，反之亦然。从图中可以看出，F统计量值随着 NVF 增加而降低。这证实了以下论断：随着 NVF 增加，人类的语言辨别能力有所提高，并逐渐接近说话人区分能力。记录到的 $NR$ 和 $RT$ 值的中位数如图3所示。从图中可以看出，与 $DER$ 类似， $NR$和 $RT$ 的中位数也随着 NVF 增加而降低。 $NR$ 和 $RT$ 的中位数在说话人变化检测中也低于语言变化检测。由此可知，人类受试者需要在变化点附近更长的时长，才能轻松地检测语言变化，而非说话人变化。

为了观察语言熟悉度对检测语言变化的影响，我们考察人类受试者在 NVF-50 和 NVF-75 组中的反应，这两组的中位数 $DER$ 小于 0。25（假设两侧均有足够的时长）。根据 LCS，将受试者的反应分为四组。根据语言熟悉度，分组标准依次为：0：非常低，1：中低，2：中等，3：优秀。所得$DER$ LCS分布如图4所示。从图中可以看出，随着 LCS 增加， $DER$值会降低。这说明，对语言的先验知识有助于人们更好地区分不同语言。

## IV. 使用无监督距离法进行语言变化检测和说话人变化检测

本节旨在借鉴现有的基于无监督距离的 SCD 框架执行语言变化检测任务。通常，SCD任务通过计算滑动分析窗口的特征之间的距离轮廓并对其进行阈值处理来完成，分析窗口长度固定为 $N$。该方法的基本框图如图5所示。首先从语音信号中提取特征向量，然后进行基于能量的语音活动检测（VAD），以获取有声帧索引。有声帧索引会被保存以供后续参考，并将与有声帧对应的特征向量用于后续处理。使用有声特征向量和两个长度固定的连续窗口，对两个不同的高斯分布进行建模（$g_a$ 和 $g_b$）。如式2所述，通过以一帧为步长滑动分析窗口，对给定测试话语进行完整扫描，从而得到散度距离轮廓。随后使用长度为（$h_l$）的汉明窗口对证据轮廓进行平滑。然后，使用峰值选取算法检测平滑后的轮廓中的峰值，该算法设有一个称为 $\gamma$的最小峰值距离参数。 $\gamma$ 的值越大，检测到的峰值数量越少，反之亦然。为减少虚假变化点的数量，这里采用了（Lu和Zhang，2002）提出并在式3中提及的阈值轮廓推导方法。最后，将检测到的峰值强度与阈值轮廓进行比较，以确定变化帧。利用已存储的有声帧位置，获得变化点的实际帧索引和采样位置。

$$D(i) = KL(g_A|g_B) + KL(g_B|g_A), \tag{2}$$

$$Th(i) = \alpha.\frac{1}{N}\sum_{n=0}^{N} D(i-n-1, i-n) \tag{3}$$

最初，我们使用 TTSF-SC 数据集来设计并调整 SCD 系统的超参数。在

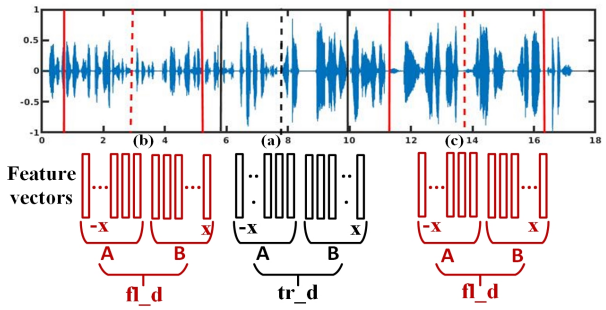


图6. 话语真实变化点和虚假变化点附近的距离计算：(a) 真实变化点，(b)、(c) 虚假变化点；fl d 和 tr d 分别为虚假距离和真实距离。

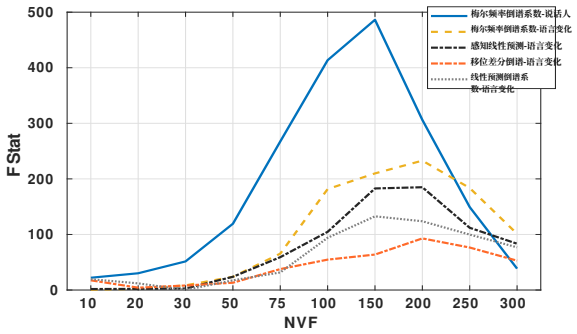


图7. 方差分析检验所得的F统计量（F Stat）值在有声帧数量（NVF）变化时，针对说话人/语言变化研究得到的真实与虚假KL散度距离之间

4000 条测试话语中，前 100 条话语用于调整超参数。观察发现，当 $\alpha = 1$设为 $\gamma$ 0.9 倍分析窗口长度，且分析窗口长度设为 150 时，性能最佳。保持方法和超参数不变，使用 TTSF-LC 和 MSCSTB 数据集执行语言变化检测任务。为评估性能，这里采用事件检测任务中常用的性能指标，即识别率（IDR）、误接受率（FAR）、漏检率（MR）和平均偏差（$D_m$）（Mishra等人，2021年; Murty 和 Yegnanarayana，2008）。两项任务的性能列于表I。

从结果可以看出，SCD 的 IDR 性能为 84.1%，而 LCD 的 IDR 性能为 51.2%。性能下降可能有两个原因：（1）所用的 MFCC 特征可能无法捕捉语言特有的判别信息；（2）超参数（主要是分析窗口长度）是针对 SCD 调整的，可能不适用于 LCD。因此，为了了解这一问题，我们通过改变变化点附近的特征和分析窗口长度开展了一项研究。本文考虑了文献中语言识别（LID）任务最常用的特征，即 MFCC、LPCC、SDC 和 PLP。研究旨在考察这些特征的语言判别能力：选取变化点附近固定数量的有声帧（NVF），$x$并将其与 MFCC 特征的说话人区分能力进行比较。这项研究有助于解释 LCD 相较于 SCD 的性能下降。此外，研究结果也有助于为 LCD 任务选择最优的特征和分析窗口长度。

本研究使用 TTSF-SC 和 TTSF-LC 数据集。从 4000 条测试话语中，选取仅包含一个变化点的话语。说话人变化和语言变化所选取的话语数量分别为 799 和 836。为观察判别能力，这里通过考察真距离与假距离之间的分布差异来进行分析

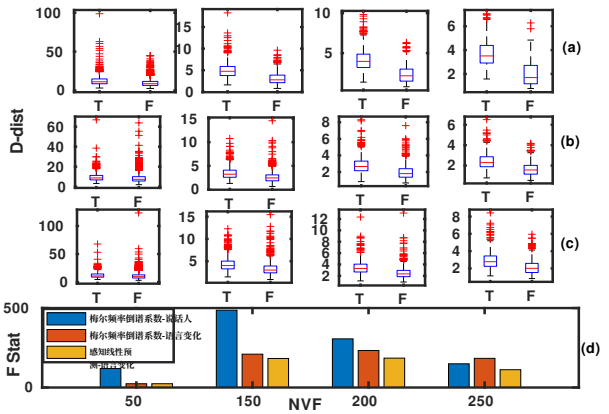


图8. 真实（T）与虚假（F）距离分布（D-dist）：（a）使用MFCC特征检测说话人变化，（b）使用MFCC特征检测语言变化，（c）使用PLP特征检测语言变化；（d）展示相应的F统计量（F Stat）值。

。真距离是变化点两侧的 $x$ 个特征向量之间的 KL 散度。类似地，假距离则通过在单语言／说话人片段中的任意位置随机放置变化点来计算。计算真距离和假距离的过程也如图6所示。为考察时长对判别能力的影响，$x$ 分别取 10、20、30、50、75、100、150、200、250 和 300。对于给定的 $x$ 值，对得到的真距离和假距离进行方差分析检验。方差分析检验所得的 F统计量如图7所示。

从图中可以看出，F统计量值会随着 NVF 的增加而增大，并在有声帧达到一定数量后趋于饱和，随后开始下降。在 LCD 和 SCD 的人类研究中也观察到了类似现象。

表I. 使用无监督距离法进行语言变化检测和说话人变化检测的性能。A：使用 $N = 150$ （针对说话人变化检测进行调优）；B：使用最优 $N$ 值（针对语言变化检测进行调优）。

| | TTSF-SC | TTSF-LC | | | MSCSTB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | GUE | | TAE | | TEE | |
| | MFCC | 梅尔频率倒谱系数 感知线性预测 | | | MFCC | | | | | |
| | A | A | B | B | A | B | A | B | A | B |
| IDR | 84.1 | 51.2 | 66.1 | 64.06 | 42.9 | 44.07 | 47.4 | 48.75 | 44.5 | 45.24 |
| FAR | 10.7 | 41.3 | 20.91 | 22.58 | 8.1 | 9.24 | 8.2 | 8.57 | 7.7 | 7.70 |
| MDR | 5.2 | 7.5 | 12.98 | 13.36 | 49 | 46.69 | 44.4 | 42.68 | 47.8 | 47.06 |
| Dm | 0.19 | 0.45 | 0.51 | 0.57 | 0.5 | 0.49 | 0.5 | 0.56 | 0.5 | 0.56 |

然而，随着 NVF 增加，人类的表现并不会下降。这可能是因为高斯假设（统计独立性）无法对说话人和语言的频谱动态进行建模，从而导致距离分布中的类别特定方差增大。使用 MFCC特征时，无论 NVF 为何值，SCD 的 F统计量值都高于 LCD。此外，还可以观察到，随着 NVF 增加，LCD 的判别能力（以 F统计量衡量）呈现出与 SCD 相同的趋势。另据了解，说话人变化和语言变化研究中，F统计量值分别在 150 和 200 时达到最高。

此外，对于语言变化研究，MFCC特征提供了更好的F统计量值，其次是感知线性预测、线性预测倒谱系数和移位差分倒谱。为便于清晰观察，在NVF为50、150、200和250时，用于执行SCD的MFCC特征，以及用于执行LCD的MFCC和感知线性预测特征的距离分布，通过图中的箱线图进行展示。8。从箱线图中还可以注意到，说话人和语言的判别能力分别在NVF为150和200时趋于饱和。尽管箱线图显示出更好的区分能力，但类间方差的增加会导致F统计量值降低。此外，与感知线性预测相比，MFCC的判别能力更强，因为在NVF等于200时，对于LCD任务，MFCC特征的真实距离分布与虚假距离分布之间的分离程度高于感知线性预测特征。基于此，我们考虑采用分析窗口长度为200的MFCC特征，对TTSF-LC数据集执行LCD。采用修改后的分析窗口长度时LCD任务的性能列于表I中。

该表显示，就IDR、FAR和MDR而言，性能表现与根据F统计量所得的观察结果一致。对于采用MFCC特征的TTSF-LC数据集（分析窗口长度为200），其IDR性能为66.1%，相对提升29.1%；随后，采用PLP特征时的IDR为64.06%。使用MSCSTB数据集也得到了类似的观察结果，其中就性能而言，对于GUE、TAE和TEE语言对，分别采用160、180和170的分析窗口长度时，IDR相对提升了2.72%、2.85%和1.63%。通过在前100个测试试验上评估分析窗口长度从100到250、步长为10时的性能，采用贪心策略确定分析窗口长度160、180和170。因此，这证实了以下假设：与SCD相比，执行LCD需要相对更长时长的信息。

## V. 基于模型的方法进行语言变化检测

人类进行的SCD和LCD表明，事先接触过某种语言有助于人们更高效地检测语言变化。因此，我们尝试从训练好的机器学习（ML）/深度学习（DL）框架中提取统计/嵌入向量，并将其用于执行变化检测任务。具体流程将在以下小节中说明。

### A. 基于模型的变化检测框架

基于模型的变化检测框架框图如图9所示。首先，从训练数据中计算梅尔频率倒谱系数+Δ + ΔΔ，并使用语音活动检测选出有声特征向量，以供后续处理。利用有声特征向量训练通用背景模型（UBM）、自适应模型、总变异矩阵（T矩阵）等统计模型，以及时延神经网络等深度学习模型（如基于 TDNN 的 x 向量模型）。分别使用训练好的 UBM、自适应模型和 T矩阵提取 u/a/i 向量等统计向量。分别通过计算 UBM 和自适应模型的零阶统计量，得到 u 向量和 a向量。根据公式4计算零阶统计量，其中 $i$的取值范围为 $1 \leq i \leq M$， $M$ 表示混合分量的数量， $x_j$ 表示梅尔频率倒谱系数特征， $T$ 表示有声帧的数量。u 向量是使用 UBM 模型提取的 $M$ 维向量，而 a向量则是将从类别特定自适应模型中提取的 $M$ 维向量串联而成

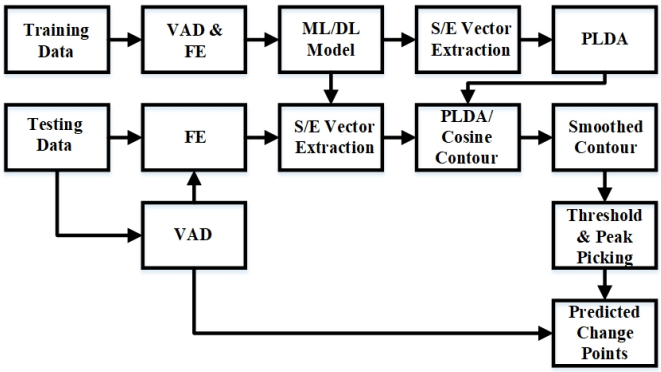


图 9. 基于模型的变化检测研究框图。

。i 向量的提取方法如（Dehak等人，2010）所述。同样，x向量是从训练好的基于 TDNN 的 x 向量模型中提取的。计算统计向量/嵌入向量时，均将 $N$个有声特征向量作为分析窗口长度。随后使用提取出的向量训练线性判别分析（LDA）、类内协方差归一化（WCCN）矩阵和概率线性判别分析（PLDA）模型。

$$N(i) = \frac{1}{T}\sum_{j=1}^{T} P(i|x_j) \tag{4}$$

测试期间，从语码转换话语中提取特征向量。之后，利用语音活动检测标签，使用训练好的模型从固定数量的有声帧中提取统计／嵌入（S/E）向量。每个测试话语的 S/E 向量提取和距离轮廓均使用公式计算：5。其中，$x_i$s' 为有声特征向量，$\psi(.)$ 为距离计算函数，而 $\mathbb{F}(.)$ 是从特征空间映射到统计／嵌入向量空间的映射函数。

$$D(i) = \psi(\mathbb{F}(x_{i-N},\ldots,x_i),\mathbb{F}(x_{i+1},\ldots,x_{i+N})) \tag{5}$$

随后，使用长度为（$h_l$）的汉明窗对距离轮廓进行平滑处理。 $h_l$ 被视为$1/\delta$ 倍的 $N$。计算平滑后轮廓的峰值，并将大于阈值轮廓的峰值幅度视为变化点。

### B. 实验设置

TTSF-SC 数据集用于 SCD，而 TTSF-LC 和 MSCSTB 则用于执行语言变化检测任务。根据语音信号计算 39 维梅尔频率倒谱系数$+\Delta+\Delta\Delta$ 特征向量，窗长和帧移时长分别为 20 毫秒和 10 毫秒。将帧能量高于话语平均帧能量 6% 的帧判定为有声帧。通用背景模型和自适应模型采用 32 的聚类大小进行训练。u/a/i 向量的维度分别为 32、64 和 50。使用 SpeechBrain 的配方训练并提取 512 维 x-vector（Ravanelli 等，2021年）。在说话人特定研究中，训练 x-vector 时不使用丢弃和 L2 归一化；而在语言特定研究中，则在第二、第三、第四和第六层使用 0.2 的丢弃率，并进行 L2 归一化。

训练期间，使用说话人／语言特定有声特征向量，以固定的 $N$不相连地提取 S/E 向量；测试期间则以采样帧移提取 S/E 向量。所有模型均训练了 20 个训练轮次。对于 TTSF-LC，通过实验确定最优的 $N$ 为 200；对于 TTSF-SC， $N$ 设为 50。训练完成后，通过观察验证损失和准确率，分别为语言特定研究和说话人特定研究选取对应于 $15^{th}$ 和 $11^{th}$ 训练轮次的模型。同样，对于 MSCSTB，针对每种语言对训练 x-vector 模型。训练 100 个训练轮次后，通过观察验证损失和准确率，为 GUE、TAE 和 TEE 语言对分别选取对应于（$54^{th}$、 $29^{th}$ 和 $26^{th}$）训练轮次的模型，以及对应于 $N = 200$ 和（$25^{th}$、 $80^{th}$和 $18^{th}$）训练轮次的模型$N = 50$ 。

对于 TTSF-LC 和 TTSF-SC，提取的嵌入向量在未进行 LDA 和类内协方差归一化的情况下进行归一化处理。归一化后的向量用于建立概率线性判别分析模型，并计算语言变化检测和说话人变化检测任务的距离轮廓。在MSCSTB数据集上观察到，与使用概率线性判别分析距离轮廓相比，同时进行 LDA 和类内协方差归一化并使用余弦核距离，可以提升变化检测性能。这可能是由于数据集的性质所致。TTSF-LC 和 TTSF-SC 是朗读语音的录音棚录音，而 MSCSTB 是办公环境中的对话录音。

对于说话人变化检测任务，在提取 i/x向量后，使用超参数 $\alpha$、 $\delta$ 和 $\gamma$，分别为 2.6、1.3 和 0.9，检测每个测试话语中的变化点。通过观察前 100 次测试试验中的变化检测性能，以贪心方式确定这些超参数。对于语言变化检测任务（使用 TTSF-LC），超参数分别确定为 3.2、1.3 和 0.9。类似地，对于 MSCSTB（$N = 200$），GUE、TAE 和 TEE 的最优超参数分别为 (0.3、4.5 和 1.1)、(0.3、4.5 和 1.1) 以及 (0.3、3.9 和 1.1)。对于 $N = 50$，最优超参数分别为 (0.3、0.9 和 1.1)、(0.3、0.9 和 1.3) 以及 (0.3、0.5 和 1.3)。

### C. 基于统计/嵌入向量的语言判别

此处旨在通过合成模拟代码转换场景，观察所提取的 S/E向量在语言判别方面的判别能力。本研究考察的是 TTSF-LC，即同一说话人使用两种语言进行说话的情形。训练分区用于训练通用背景模型、自适应、T矩阵和基于时延神经网络的

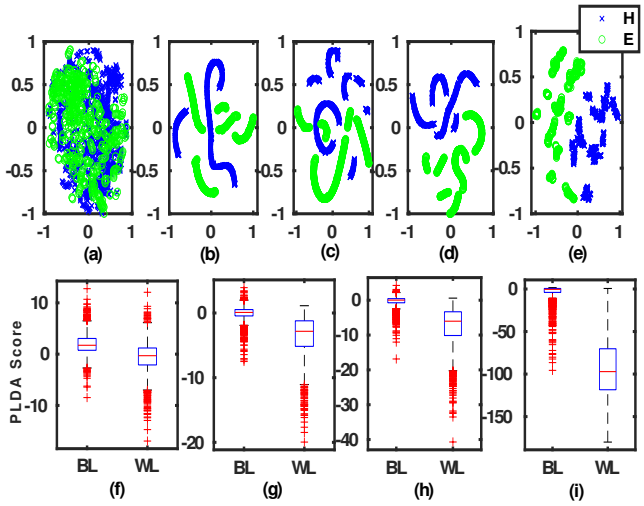


图10. 印地语（H）和英语（E）之间的t-SNE特征分布：（a）MFCC特征、（b）u向量、（c）a向量、（d）i-vector和（e）x-vector。语言内和语言间的PLDA分数分布，以及u、a、i和x-vector分别对应的等错误率：（f）28.5、（g）17.35、（h）12.55和（i）3.6。

x向量模型。从测试分区中选取两段话语，每种语言各一段，均由同一说话人说出。利用选取的话语提取梅尔频率倒谱系数+Δ+ΔΔ 特征和 S/E向量，并使用 t-SNE 将其投影到二维空间（Maaten 和 Hin- ton，2008）。

二维表征如图10（a-e）所示。从图中可以观察到，从特征空间转移到S/E向量空间时，语言之间的重叠减少了。这表明，与人类受试者类似，通过机器学习/深度学习模型先前接触过这些语言，有助于提高语言辨别能力。此外，在S/E向量中，语言之间的重叠在x向量空间中最少，其次依次为i-vector、自适应空间和通用背景模型后验空间。这是由于这些建模技术能够捕捉语言特定的特征动态。

为进一步强化这一观察，我们从测试话语中提取特征，并按给定语言汇集这些特征。将汇集后的特征向量随机切分为上下文长度为 200 的片段，并用于提取 S/E 向量。将提取出的 S/E 向量配对，形成 2000 个语言内（WL）向量对和 2000 个语言间（BL）向量对。使用 PLDA 分数对 WL 和 BL 向量对进行比较。图 10(f-i) 展示了 WL 和 BL 对的 PLDA 分数分布箱线图。从箱线图分布可以看出，随着建模技术从 UBM 改进到 x-vector，WL 与 BL 之间的 PLDA 分数分布重叠程度有所降低。

在变点检测任务中，目标是在语言发生变化时，使距离轮廓出现突变。若语言内的轮廓（PLDA分数的负值）变化较小，

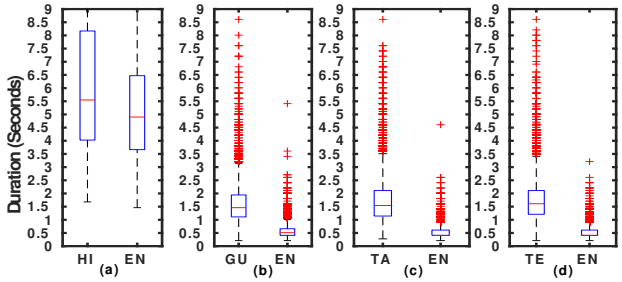


图 11. TTSF-LC 的 HE 以及 MSCSTB 的 (b) GUE、(c) TAE 和 (d) TEE 语言对的单语片段时长分布。

而跨语言对的轮廓出现突变，便可实现这一目标。因此，为确保这一点，应最大化语言内与跨语言对之间的PLDA分数分布差异。基于此，采用等错误率（等错误率）作为客观衡量指标，其中语言内和跨语言对的试验分别称为假分数和真分数。UBM、自适应/ i-vector 和 x-vector 对应的等错误率分别为 28。5、17。35、12。55 和 3。6。因此，根据判别能力，本研究使用 i/x-vector 作为说话人和语言的表征开展变点检测研究。

**D. 实验结果**

首先，使用 i/x 向量作为说话人/语言表示，在 TTSF-SC 和 TTSF-LC 上开展变化检测研究。判别能力以及语言变化检测/说话人变化检测研究表明，与 i-vector 相比，x-vector 是更好的说话人/语言表示。因此，在 MSCSTB数据集上进行语言变化检测任务时，采用 x向量作为语言表示。

实验结果列于表 II中。在说话人变化检测任务中，使用 i-vector 和 x-vector 获得的 IDR 性能分别为 87.75% 和 92.27%。类似地，对于语言变化检测任务，TTSF-LC 上的性能分别为 80.58% 和 87.01%。语言判别研究表明，与基于无监督距离的方法所取得的最佳性能相比，使用 i 向量和 x向量进行语言变化检测的性能分别相对提升 21.9% 和 31.63%。这证实了如下观点：与人类类似，通过计算模型融入特定语言的先验信息，可以提升语言变化检测的性能。

在 MSCSTB 数据集上使用 x向量作为语言表示，并将 N 设为 200（与 TTSF-LC 相同）时，LCD任务在 GUE、TAE 和 TEE 分区上的 IDR 分别为 46.56%、49.91% 和 47.13%。性能分别相对提升了 5.6%、2.3% 和 4.2%。然而，与使用 TTSF-LC 数据取得的提升相比，这一提升幅度较小。这可能

表II. 基于模型的方法在语言变化检测和说话人变化检测中的性能，S：统计 i 向量，E：基于嵌入的 x 向量，N：分析窗口长度。

| | TTSF-SC | | TTSF-LC | | MSCSTB | | | MSCSTB | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | GUE | TAE | TEE | GUE | TAE | TEE |
| | S | E | S | E | E | | | E | | |
| N | 50 | 50 | 200 | 200 | 200 | | | 50 | | |
| IDR | 87.75 | 92.27 | 80.58 | 87.01 | 46.56 | 49.91 | 47.13 | 54.74 | 52.19 | 50.84 |
| FAR | 5.42 | 3.96 | 8.8 | 8.84 | 5.95 | 10.42 | 6.50 | 13.10 | 27.82 | 19.34 |
| 漏检率 | 6.83 | 3.76 | 10.57 | 4.41 | 47.49 | 39.67 | 46.36 | 32.16 | 19.99 | 29.83 |
| Dm | 0.05 | 0.03 | 0.33 | 0.28 | 0.51 | 0.56 | 0.56 | 0.35 | 0.30 | 0.34 |

表 III. 改变分析窗口长度时语言变化检测的性能。

| | MSCSTB | | | | |
|---|---|---|---|---|---|
| | GUE | | | | |
| N | 200 | 150 | 100 | 75 | 50 |
| IDR | 46.56 | 48.12 | 50.12 | 51.78 | 54.74 |
| FAR | 5.95 | 7.38 | 12.54 | 11.01 | 13.10 |
| 漏检率 | 47.49 | 44.50 | 37.35 | 37.21 | 32.16 |
| Dm | 0.51 | 0.48 | 0.37 | 0.33 | 0.35 |

是因为 TTSF-LC 和 MSCSTB 数据集中的单语片段时长分布存在差异。图中以箱线图展示了 TTSF-LC 和 MSCSTB 测试集的单语片段分布，见图 11。从图中可以观察到，TTSF-LC 中主要语言和次要语言的单语片段时长中位数分别为（5.54 秒和 4.9 秒）；MSCSTB 中 GUE、TAE 和 TEE 分区的对应中位数分别为（1.46 秒和 0.51 秒）、（1.54 秒和 0.41 秒）以及（1.61 秒和 0.41 秒）。此外，研究发现，将 $N$ 设为 200（即约 2 秒）时，语言判别效果更好。因此，由于 MSCSTB 数据集的单语片段时长小于所采用的分析窗口时长，导致平滑后的距离轮廓，并使漏检率上升。因此，另一种做法是缩短分析窗口长度，但这可能会影响 x向量的语言辨别能力。

开展了一项研究，用于观察分析窗口长度与语言判别能力之间的权衡关系。利用MSCSTB数据集的GUE划分，通过将分析窗口长度从200缩短至50，进行了语言区分测试和LCD任务。LCD任务的结果列于表III中。图12展示了在改变分析窗口长度的情况下，x向量的语言内和语言间对经过LDA和类内协方差归一化投影后的余弦得分分布。从表中可以

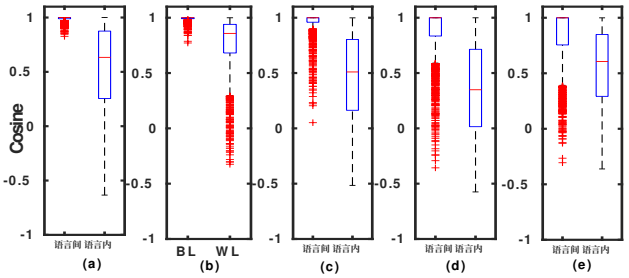


图12. 分析窗口长度分别为 200、150、100、75 和 50 时，语言内与语言间余弦得分分布及等错误率分别为 (a)7.1、(b) 9.8、(c) 12.8、(d)19.8 和 (i) 29.2。

观察到，随着 $N$的减小，LCD任务的性能得到提升，并在。74%处达到54%的最佳性能，此时 $N$ 等于50。因此，对于GUE、TAE和TEE语言对，使用 $N$ 等于50计算变化检测性能，结果列于表II中。然而，通过x向量模型纳入特定语言的先验暴露所带来的相对性能提升，并未像在TTSF-LC数据集中那样明显。这是由于随着 $N$的减小，x向量的语言判别能力也随之降低。从图12可以观察到，随着 $N$值减小，语言内和语言间得分分布之间的重叠增加。作为客观衡量指标，当 $N$ 分别等于200、150、100、75和50时，计算得到的等错误率分别为7。1%、9。8%、12。8%、19。8%和29。2%。

## VI. 讨论

基于人类的LCD和SCD研究表明，相较于说话人，人类需要更多的邻域信息，才能舒适地区分语言。此外，先前接触过这些语言有助于人类更好地区分它们。受此启发，我们提出假设：通过（a）引入

表 IV. 性能比较，RI：相对提升，A：使用 $N = 150$ （针对 SCD 调优），B：使用最优$N$ 值（针对 LCD 调优），C：基于 x-vector 的方法。

| 数据集 | 方法 IDR | | RI |
|---|---|---|---|
| TTSF-SC | A | 84.1 | - |
| | C | 92.27 | 9.71 |
| TTSF-LC | A | 51.2 | - |
| | B | 66.1 | 29.1 |
| | C | 87.01 | 31.63 |
| MSCSTB | A | 44.93 | - |
| | B | 46.02 | 2.4 |
| | C | 52.59 | 14.27 |

更大时长的分析窗口（$N$），以及（b）通过计算模型进行特定语言的接触，可以提升机器执行LCD的性能。

在无监督距离法中，已观察到，提高 $N$的值可以提升语言变化检测的性能。说话人变化检测研究中的最佳 $N$ 值为 150。采用相同的$N$值，对 TTSF-LC 和 MSCSTB 数据集进行语言变化检测任务，结果列于表 IV中。对于 MSCSTB 数据集，表中列出了三个语言对的平均 IDR 值。受人类语言变化检测／说话人变化检测研究的启发，我们增大了 $N$ 值，得到 TTSF-LC 语言变化检测的最佳$N$ 值为 200。类似地，MSCSTB 中 GUE、TAE 和 TEE 的最佳 N 值分别为 160、180 和 170。TTSF-LC 和 MSCSTB 在最佳 $N$ 值下的性能分别为 66。1% 和 46。02%，相对提升分别为 29。1% 和 2。4%。这些观察结果证实，增大分析窗口时长可以提升机器语言变化检测的性能。

此外，正如主观研究所提出的假设，通过计算模型引入特定语言接触能够提升语言变化检测的性能。i/x-vector模型已经过训练，实质上捕捉了特定语言的倒谱动态特征。研究发现，采用x-vector方法时，TTSF-LC上的性能为87.01%，IDR为52.59%；与无监督距离法的性能相比，相对提升分别为31.63%和14.27%。同样，在使用TTSF-SC数据集进行说话人变化检测任务时，性能相对提升了9.71%。比较合成数据上的语言变化检测与说话人变化检测性能可以发现，前者的提升幅度大于后者。由此可见，与人类主观研究的结果一致，在理想条件下（仅存在说话人/语言变化，且其他变化有限），基于需求模型的方法对语言变化检测比对说话人变化检测更为重要。

此外还观察到，在LCD任务中，与合成TTSF-LC数据集上取得的性能提升相比，MSCSTB数据上的性能提升有限。这是由于单语片段时长不同所致。分析窗口时长与语言判别能力之间存在权衡：随着分析窗口时长增加，判别能力会提高。同时，在变化检测过程中，由于单语片段时长可能短于500毫秒（约50个有声帧），采用较大的分析窗口会使证据曲线变得平滑，从而导致性能下降（漏检率上升）。因此，要解决这一问题，（1）需要在 $N$ 值尽可能小的情况下实现显著的语言判别能力；（2）需要开发一种框架，使其性能受分析窗口时长变化的影响尽可能小，或不受其影响。

## 七、结论

在本研究中，我们使用适用于SCD的现有框架进行了LCD。根据主观研究可以观察到，与说话人相比，人类需要在变化点周围获取相对更大的邻域信息。研究还发现，先前的特定语言接触能够提升LCD任务的性能。在无监督距离法中，引入更大的邻域信息后，在合成TTSF-LC数据集和实际MSCSTB数据集上，LCD性能分别相对提升29.1%和2.4%。同样，通过计算模型引入特定语言先验信息后，相较于无监督距离法，性能分别相对提升31.63%和14.27%。

研究还观察到，实际数据集的表现不如预期，无法达到合成数据的水平。这是由于两个数据集中的单语片段时长分布存在差异。MSCSTB数据集包含时长小于0.5秒的单语片段，而要实现更好的语言判别，所需时长约为2秒（约200个有声帧）。因此，分析窗口时长的确定颇具挑战性。较大的时长会平滑证据轮廓并增加MDR，而较小的0.5秒时长则无法提供恰当的语言判别能力。

因此，我们未来将尝试开发更好的框架，使其能够在较短时长内更准确地区分语言；此外，我们还计划提出一种变化检测框架，其性能应不受分析窗口时长变化的影响，或较少受到其影响。

**致谢**

作者谨此感谢印度理工学院达尔瓦德分校的高性能计算（HPC）设施“Anatganak”，为我们开展实验提供了支持。此外，感谢电子和信息技术部（Me-

itY）以及印度政府通过多个项目给予我们的支持。

[1]关于人类语言变化检测和说话人变化检测研究的数据生成流程的补充材料见 [AIP 将插入 URL]

Baby, A., Thomas, A. L., Nishanthi, N., Consortium, T. *et al.* (2016)。“印度语言资源”，载于 *Proceedings of Text, Speech and Dialogue* 。Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006)。“广播新闻的多阶段说话人分离”，IEEE 音频、语音与语言处理汇刊 14(5), 1505–1512。Bredin, H., Barras, C. *et al.*(2017)。“使用双向长短期记忆网络进行广播电视中的说话人变化检测”，载于 *Interspeech 2017*, ISCA。Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., and Deller, J. (2002)。“使用高斯混合模型和移位差分倒谱特征进行语言识别的方法”，ICSLP2002-INTERSPEECH2002 会议录，16–20。Chan, J. Y., Ching, P., Lee, T., and Meng, H. M. (2004)。“通过双音素概率检测语码转换话语中的语言边界”，载于 *2004 International Symposium on Chinese Spoken Language Processing* 293 296 ，IEEE, 第 – 页。Chen, S., Gopalakrishnan, P. *et al.*(1998)。“基于贝叶斯信息准则的说话人、环境和通道变化检测与聚类”，载于 *Proc. DARPA broadcast news transcription and understanding workshop* 8, ，美国弗吉尼亚州，第 卷，第 127–132 页。Dawalatabad, N., Madikeri, S., Sekhar, C. C., and Murthy, H. A. (2020)。“用于会议无监督信息瓶颈说话人分离的新型架构”，IEEE/ACM 音频、语音与语言处理汇刊 29, 14 27–。Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010)。“用于说话人验证的前端因子分析”，IEEE 音频、语音与语言处理汇刊 19(4), 788–798。Dhananjaya, N., and Yegnanarayana, B. (2008)。“利用激励源特征检测非正式对话中的说话人变化”，语音通信 50(2), 153–161。Diwan, A., Vaideeswaran, R., Shah, S., Singh, A., Raghavan, S., Khare, S., Unni, V., Vyas, S., Rajpuria, A., Yarra, C., Mittal, A., Ghosh, P. K., Jyothi, P., Bali, K., Seshadri, V., Sitaram, S., Bharadwaj, S., Nanavati, J., Nanavati, R., Sankaranarayanan, K., Seeram, T., and Abraham, B. (2021)。“面向低资源印度语言的多语言与语码转换自动语音识别挑战”，Interspeech 会议录。Gish, H., Siu, M.-H., and Rohlicek, J. R. (1991)。“用于语音识别和说话人识别的说话人分离”，载于 *icassp*, 第 91 卷，第 873–876 页。Haris, B. C., Pradhan, G., Misra, A., Prasanna, S., Das, R. K., and Sinha, R. (2012)。“印度场景中的多变异性说话人识别数据库”，国际语音技术期刊 15(4), 441–453。Hogg, A. O., Evers, C., and Naylor, P. A. (2019)。“利用基频检测说话人变化及其在多说话人分段中的应用”，载于 *ICASSP 2019-2019 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* 5826 5830 ，IEEE, 第 – 页。Li, H., Ma, B., and Lee, K. A. (2013)。“口语识别：从基础到实践”，IEEE 会议录 101(5), 1136–1159。Liu, H., Perera, L. P. G., Zhang, X., Dauwels, J., Khong, A. W., Khudanpur, S., and Styles, S. J. (2021)。“双语语码转换语音的端到端语言说话人日志分析”，载于 *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, 国际语音通信协会，第 2 卷。Lu, L., and Zhang, H.-J. (2002)。“实时新闻广播分析中的说话人变化检测与跟踪”，载于 *Proceedings of the tenth ACM international conference on Multimedia*，第 602–610 页。Lyu, D. C., Chng, E. S., and Li, H. (2013)。“通过发音词典自适应对会话式语码转换语音进行语言说话人日志分析”，载于 *Signal and Information Processing (ChinaSIP) 2013 IEEE China Summit and International Conference on* 147 150 ，IEEE, 第 – 页。Maaten, L. v. d., and Hinton, G. (2008)。“使用 t-SNE 对数据进行可视化”，机器学习研究期刊 9(Nov), 2579–2605。Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., and Besacier, L. (2006)。“广播新闻说话人日志分析的逐步式与集成式方法”，计算机语音与语言 20(2-3), 303–330。Mishra, J., Agarwal, A., and Prasanna, S. M. (2021)。“使用基于注意力的神经网络进行口语说话人日志分析”，载于 *2021 National Conference on Communications (NCC)*, IEEE, 第 1–6 页。Moattar, M. H., and Homayounpour, M. M. (2012)。“说话人日志分析系统与方法综述”，语音通信 54(10), 1065–1103。Murty, K. S. R., and Yegnanarayana, B. (2008)。“从语音信号中提取声门脉冲”，IEEE 音频、语音与语言处理汇刊 16(8), 1602–1613。Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022)。“说话人分离综述：深度学习的最新进展”，计算机语音与语言 72, 101317。Rabiner, L. R. (1978)。 *Digital processing of speech signals* (培生印度教育出版社)。Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J. *et al.* (2021)。“SpeechBrain：通用语音工具包”，arXiv 预印本 arXiv:2106.04624。Richardson, F., Reynolds, D., and Dehak, N. (2015)。“用于说话人与语言识别的深度神经网络方法”，IEEE 信号处理快报 22(10), 1671–1675。Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2018)。“首届DIHARD挑战赛评估计划”，2018，技术报告。Sarma, M., Gadre, S. N., Sarma, B. D., and Prasanna, S. M. (2015)。“利用激励源和声道系统信息进行说话人变化检测”，载于 *2015 Twenty First National Conference on Communications (NCC)* 1 6 ，IEEE, 第 – 页。Sharma, N. K., Ganesh, S., Ganapathy, S., and Holt, L. L. (2019)。“说话者变更检测：人类与机器表现的比较”，美国声学学会期刊 145(1), 131–142。Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. (1997)。“广播新闻音频的自动分段、分类和聚类”，载于 *Proc. DARPA speech recognition workshop*, 第 1997 卷。Sitaram, S., Chandu, K. R., Rallabandi, S. K., and Black, A. W. (2019)。“语码转换语音与语言处理综述”，arXiv:1904.00784 [cs.CL] 。Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (2018)。“使用 x 向量进行口语识别”，载于 *Odyssey*, 第 105–111 页。Spoorthy, V., Thenkanidiyoor, V., and Dinesh, D. A. (2018)。“基于支持向量机的语码转换双语印度语音语言说话人日志分析：采用瓶颈特征”，载于 *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages* 132 136 ，第 – 页，http://dx.doi.org/10.21437/SLTU.2018-28 10.21437/SLTU.2018-28 ，doi: 。Tranter, S. E., and Reynolds, D. A. (2006)。“自动说话人日志分析系统概述”，IEEE 音频、语音与语言处理汇刊 14(5), 1557–1565。