SIDEKIT是一个新的开源Python工具包，它包含一个由最新组件组成的大型面板，允许对端到端的说话人识别系统进行快速原型设计。对于前端特征提取、规范化、语音活动检测、建模、评分和可视化的每个步骤，SIDEKIT提供了广泛的标准算法和灵活的界面。使用单一高效的编程和脚本语言（本例中为Python）以及有限的依赖性，有助于部署工业应用程序和扩展，以将新算法作为SIDEKIT提供的整个工具链的一部分。SIDEKIT的性能在两个标准评估任务中得到了验证，即RSR2015和NIST-SRE 2010。

1. INTRODUCTION  
   一。导言  
   Speaker verification is the task of comparing audio recordings to answer the question ” Is the same speaker speaking in all the recordings?” [1]. The domain is still an active research area as many problems are not solved; the performance of systems in adverse conditions such as noisy environment, de- graded communication channels or short duration of speech samples [2] is still limiting an extensive use of the technology. At the same time, performance in more controlled conditions have reached a point that allows a number of commercial ap- plications.  
   说话人确认是比较两个录音的任务，以回答“在所有录音中，同一个说话人讲话吗？”？“[1]。由于许多问题尚未得到解决，该领域仍然是一个活跃的研究领域；系统在噪声环境、降级通信信道或短时间语音样本等不利条件下的性能[2]仍然限制了该技术的广泛应用。同时，在更可控的条件下，性能已经达到允许许多商业应用程序使用的程度。

Over the years, a number of toolkits have been developed that fulfil different purposes; some are dedicated to research and focus on flexibility while others target on efficiency to be compatible with industrial requests. As researchers, we aim at developing new algorithms while keeping close to industrial standards in order to enable quick technology transfer. To achieve these two goals, a speaker recognition toolkit should fulfil a number of requirements:  
多年来，研究者已经开发了一些满足不同目的的工具包；一些工具包致力于研究和注重灵活性，而另一些工具包则着眼于提高效率，以符合工业要求。作为研究人员，我们的目标是开发新的算法，同时保持接近工业标准，以便实现快速的技术转让。为了实现这两个目标，说话人识别工具包应满足以下几个要求：  
• easy to understand and modify;  
• easy to install and start with;  
• allow the development of an end-to-end speaker recog- nition system;  
•易于理解和修改；•易于安装和启动；•允许开发端到端说话人识别系统；  
• minimum dependencies on other tools;  
•对其他工具的依赖性较小；  
• implement a wide range of standard algorithms;  
•实施各种标准算法；  
• enable the use of large data sets an fast computation to obtain state-of-the-art performance;  
•能够对大数据集进行快速计算，以获得最先进的性能；  
• manage standard data formats to allow compatibility with existing tools.  
•管理标准数据格式，以便与现有工具兼容。  
Considering the advantages and drawbacks of existing tools, we developed a new toolkit for speaker recognition, SIDEKIT, that aims at fulfilling the above-mentioned require- ments while providing an end-to-end solution to integrate a wide choice of state-of-the-art algorithms. Focusing on the easiness of use, we included a complete documentation, ex- amples and tutorials on standard tasks for an easy first-use of the toolkit.  
考虑到现有工具的优缺点，我们开发了一个新的说话人识别工具SIDEKIT，它旨在满足上述要求，同时提供端到端的解决方案，以集成多种最新算法。为了便于使用，我们提供了一个完整的文档、示例和关于标准任务的教程，以便于首次使用工具包。  
Additionally, the use of an open-source licence would en- able a wide diffusion, a quick development and facilitate the technology transfer, if the licence is permissive enough.  
此外，如果许可证允许的话，使用开源许可证将能够广泛传播、快速发展和促进技术转让。  
This article details our motivations, describes the main functionalities of the Speaker IDEentification toolKIT, SIDEKIT, explains how to start with this new tool and demonstrates the performance of SIDEKIT on two standard tasks.  
本文详细介绍了我们的动机，描述了Speaker identification工具包SIDEKIT的主要功能，解释了如何从这个新工具开始，并演示了SIDEKIT在两个标准任务上的性能。

1. MOTIVATIONS

SIDEKIT aims at providing an end-to-end tool-chain encom- passing various state-of-the-art methods, easy to start with and to modify. The content of SIDEKIT has been thought to address the lacks of existing toolkits. Our intention is to keep the architecture simple so as to facilitate the use and the development of new approaches.  
SIDEKIT旨在提供端到端的工具链，包括各种最先进的方法，易于开始和修改。SIDEKIT的内容被认为解决了现有工具包的不足。我们的目的是保持体系结构的简单，以便于新方法的使用和开发。

2.1. Comparison with other tools  
2.1条。与其他工具的比较

Several good tools are available but don’t serve the purpose for one or multiple reasons. ALIZE [3] is an open-source C++ toolkit widely used. It includes recent developments in speaker recognition and its efficient implementation in C++ provides fast integration for commercial applications. Modi- fying the C++ code efficiently requires a deep knowledge of the software architecture and is usually time consuming. Fur- thermore, ALIZE does not provide feature extraction or visu- alization tool.  
有几种好的工具可供使用，但由于一个或多个原因不能达到目的。 ALIZE（3）是一个广泛使用的开源C++工具包。它包括最近的发展，在说话人识别和有效的实现在C++提供快速集成的商业应用。有效地修改C++代码需要对软件体系结构有深入的了解，而且通常是耗时的。而且，ALIZE不提供特征提取或可视化工具。

[3] A.Larcher,J.-F.Bonastre,B.Fauve,K.A.Lee,C.Le ́vy, H. Li, J. S. Mason, and J.-Y. Parfait, “ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition,” in Annual Conference of the International Speech Communication Association (Interspeech), 2013, pp. 2768–2773.

Kaldi [4] is an open-source C++ toolkit dedicated to speech recognition. Due to the recent use of i-vectors for session adaptation [5], an i-vector module has been added into Kaldi that can be used for speaker recognition. Kaldi is evolving quickly thanks to a very dynamic community but the toolkit, for instance the front-end processing, is highly motivated for speech recognition task.  
Kaldi[4 ]是一个专用于语音识别的开源C++工具包。**由于最近在会话自适应中使用了i-向量[5]，因此在Kaldi中添加了i-向量模块，可用于说话人识别**。Kaldi由于一个非常活跃的社区而发展迅速，但是工具箱，例如前端处理，对于语音识别任务来说是非常有动力的。

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.  
[5] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, “I- vector-based Speaker Adaptation of Deep Neural Net- works for French Broadcast Audio Transcription,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2014.

MSR [6] is a Matlab toolbox that includes the entire tool- chain to develop an i-vector PLDA system. It includes basic feature extraction and visualization tools but is limited to the i-vector approach. The cost of the Matlab environment limits the use of this tool and integration in a commercial applica- tion imposes to rewrite the code in a more standard computer language.  
MSR[6]是一个Matlab工具箱，它包含开发i矢量PLDA系统的整个工具链。它包括基本的特征提取和可视化工具，但仅限于i矢量方法。Matlab环境的成本（收费）限制了此工具的使用，并且在商业应用中集成需要用更标准的计算机语言重写代码。

[6] S. O. Sadjadi, M. Slaney, and L. Heck, “MSR Iden- tity Toolbox v1.0: A MATLAB Toolbox for Speaker- Recognition Research,” Speech and Language Process- ing Technical Committee Newsletter, vol. 1, no. 4, November 2013.

Spear-BOB [7] is one of the most recent toolbox for speaker recognition. The whole chain of recognition is ef- ficiently implemented in C++ and Python including basic feature extraction, GMM modelling, joint factor analysis (JFA), i-vector, back-end and visualization tools. The Python higher layer of Spear makes it easy to set-up a state-of-the- art system but modification of the lower C++ layer could be complex and time consuming.  
Spear BOB[7]是最新的说话人识别工具箱之一。在C++和Python中实现了完整的识别链，包括基本特征提取、GMM建模、联合因子分析（JFA）、I向量、后端和可视化工具。Python的更高层次的矛使它易于设置最先进的系统，但较低的C++层的修改可能是复杂和耗时的。

[7] E.Khoury,L.E.Shafey,andS.Marcel,“Spear: An open source toolbox for speaker recognition based on Bob,” in International Conference on Audio, Speech and Sig- nal Processing (ICASSP), 2014.

2.2. Compatibilities with existing tools  
2.2条。与现有工具的兼容性  
In order to benefit from the best of all available tools and to facilitate smooth transitions between them, SIDEKIT is com- patible with some of the most popular formats for speaker recognition. SIDEKIT is able to read and write features in both SPRO4,1, and HTK [8] formats, and GMMs in ALIZE [3] and HTK formats. Most of the objects in SIDEKIT can also be saved in the open and portable HDF5 format used in BOSARIS,2 .  
为了从所有可用的工具中受益并促进它们之间的平滑过渡，SIDEKIT可以与一些最流行的说话人识别格式兼容。SIDEKIT能够读写SPRO4、1,和HTK[8]格式的特性，以及ALIZE[3]和HTK格式的GMMs。SIDEKIT中的大多数对象也可以以BOSARIS中使用的开放和便携式HDF5格式保存。

1 http://www.irisa.fr/metiss/guig/spro/  
2 https://sites.google.com/site/bosaristoolkit/

[8] S. Young and S.J.Young, “The HTK Hidden Markov Model Toolkit: Design and Philosophy,” Entropic Cam- bridge Research Laboratory, Ltd, vol. 2, pp. 2–44, 1994.

2.3. Structure of SIDEKIT  
2.3条。SIDEKIT结构  
SIDEKIT is 100% Python and has been tested on several plat- forms under Python 2.7 and > 3.4. The toolkit has been de- veloped with minimum dependencies to external modules and to make full use of the most standard Python modules for lin- ear algebra, matrix manipulation, etc… To maximize readabil- ity and flexibility, SIDEKIT is built on a limited number of classes that are listed below.  
SIDEKIT是100%的Python，已经在Python 2.7和>3.4下的几个平台上进行了测试。开发工具包时，对外部模块的依赖性最小，并充分利用最标准的Python模块进行线性代数、矩阵操作等。为了最大限度地提高可读性和灵活性，SIDEKIT构建在下面列出的有限数量的类上。

FeaturesServer offers a simple interface to load and save acoustic features read in SPRO4, HTK format or ex- tracted from audio files (RAW, WAV, SPHERE)  
**FeaturesServer**提供了一个简单的接口，用于加载和保存以SPRO4、HTK格式读取或从音频文件（RAW, WAV, SPHERE）提取的声学特性

StatServer class used to store and process zero and first or- der statistics considering different types of observations (acoustic features, i-vectors or super-vectors)  
**StatServer类**，该类用于存储和处理零和一阶或二阶统计数据，同时考虑到不同类型的观测数据（声学特性、i矢量或超级矢量）

**Mixture** stores and process GaussianMixtureModels(GMM)  
**Mixture**：存储和处理高斯混合模型（GMM）

Bosaris classes SIDEKIT makes use of the main classes of the BOSARIS toolkit to manage files and trial lists, scores matrices and DET plots.  
**Bosaris类**：SIDEKIT使用Bosaris工具包的主要类来管理文件和trial lists、分数矩阵和DET图。

1. WHAT IS IN SIDEKIT?

This section describes the main features included in the toolkit by the time we wrote this article. On-going devel- opment that will be included in the toolkit will be discussed in the last section of this article.  
本节描述了我们撰写本文时工具包中包含的主要功能。本文的最后一节将讨论工具包中包含的正在进行的开发。

3.1. Front-End  
3.1条。前端  
SIDEKIT offers a simple interface to extract, extend and nor- malize filter banks and cepstral coefficients with linear- or Mel-scale filter bank (LFCC and MFCC). Two voice-activity detection algorithms based on energy are available. Addition- ally, SIDEKIT supports selection of feature frames based on external labels and exports labels in ALIZE format.  
SIDEKIT提供了一个简单的接口，可以使用线性或Mel尺度滤波器组（LFCC和MFCC）提取、扩展和非恶意化滤波器组和倒谱系数。提出了两种基于能量的语音活动检测算法。另外，SIDEKIT支持基于外部标签选择特征框，并以ALIZE格式导出标签。

Several options are offered for contextualization of acous- tic features. In particular, ∆ and ∆∆ can be computed with a simple two points difference or by using a window filtering as described in [9]. Alternatively, a recently proposed method based on a 2D-DCT followed by Principal Component Anal- ysis dimension reduction is also provided [10].  
为声学特性的上下文化提供了几种选择。特别是，∏和∏∏可以通过简单的两点差分或使用如[9]所述的窗口滤波来计算。或者，还提供了最近提出的基于2D-DCT和主成分分析降维的方法[10]。

[9] M. McLaren, N. Scheffer, L. Ferrer, and Y. Lei, “Ef- fective use of DCTs for contextualizing features for speaker recognition,” in International Conference on Audio, Speech and Signal Processing (ICASSP), 2014, pp. 4027–4031.  
[10] M. McLaren and Y. Lei, “Improved speaker recogni- tion using DCT coefficients as features.” in Interna- tional Conference on Audio, Speech and Signal Process- ing (ICASSP), IEEE, Ed., 2015, pp. 4430–4434.

Standard normalizations are implemented: cepstral mean subtraction (CMS), cepstral mean variance normalization (CMVN) and short term Gaussianization (STG) [11]. RASTA filtering is also included in the toolkit.  
实现了标准规范化：倒谱均值减（CMS）、倒谱均值方差规范化（CMVN）和短期高斯化（STG）[11]。RASTA过滤也包含在工具包中。

The FeaturesServer includes standard front-end al- gorithms organized in sequential function calls that enable easy integration of new methods for the different steps of the process.  
特性服务器包括标准的前端算法，这些前端算法是按序列函数调用组织的，可以方便地为流程的不同步骤集成新方法。

3.2. Modelling and classifiers  
3.2条。模型和分类器  
The core of SIDEKIT is based on GMM-based approaches. The Mixture class includes two versions of the Expecta- tion Maximization (EM) algorithm with Maximum Likelihood criteria to train a Universal Background Model (UBM). One that initializes a single Gaussian and perform iterative split- ting based on variance gradient, and a second that randomly initializes a GMM and performs EM algorithms with a con- stant number of distributions. The mixtures variance can be constrained between a flooring and a ceiling value. Target model can be enrolled using Maximum a Posteriori (MAP) adaptation [12]  
SIDEKIT的核心是基于GMM的方法。The Mixture class包含两个版本的期望最大化（EM）算法，该算法具有训练通用背景模型（UBM）的最大似然准则。一种是初始化单个高斯函数并基于方差梯度执行迭代分割，另一种是随机初始化GMM并执行具有恒定分布数的EM算法。混合料的差异可以限制在地板和天花板值之间。目标模型可以使用最大后验（MAP）自适应来注册[12]。

On top of the simple GMM modelling, SIDEKIT inte- grates Factor Analysis based approaches in a single frame- work. Indeed, both Joint-Factor Analysis (JFA)[13, 14] and Probabilistic Linear Discriminant Analysis (PLDA)[15] were derived from a basic Factor Analyser [16] and therefore share a common decoupled implementation despite exhibiting two major differences. Firstly, JFA considers acoustic frames as observations while PLDA models a single distribution of i- vectors [17] or super-vectors [18]. Secondly, JFA ties latent factors across a temporal sequence of observations and across mixtures while PLDA ties the latent factors across speakers [19]. The Factor Analysis implementation follows [20] with minimum divergence step as described in[21].  
在简单的GMM建模的基础上，SIDEKIT将基于因子分析的方法集成到单个框架中。事实上，联合因子分析（JFA）[13，14]和概率线性判别分析（PLDA）[15]都是从基本因子分析器[16]中导出的，因此，尽管显示出两个主要差异，但它们共享一个共同的解耦实现。首先，JFA将声学帧视为观测值，而PLDA将i矢量或超级矢量的单个分布建模[17]或[18]。其次，JFA通过观察的时间序列和混合体将潜在因素联系起来，而PLDA通过说话人将潜在因素联系起来[19]。因子分析的实现遵循[20]，最小发散步如[21]所述。

SIDEKIT includes a standard i-vector extractor as well as two fast implementations based on the work of [22]. Several normalization algorithms are included: Eigen Factor Radial [23], Spherical Nuisance Normalization [23], LDA, WCCN [17] and various scoring methods: Cosine [24], Mahalanobis, Two-Covariance model [25], as well as partially closed-set PLDA likelihood ratio scoring [26, 27, 28].  
SIDEKIT包括一个标准的i向量提取器以及两个基于[22]工作的快速实现。包括几个标准化算法：特征因子径向[23]、球面干扰标准化[23]、LDA、WCCN[17]和各种评分方法：余弦[24]、马氏体、两个协方差模型[25]以及部分闭集PLDA似然比评分[26、27、28]。

[22] O. Glembeck, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, “Simplification and optimization of I- Vector extraction,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2011, pp. 4516–4519.  
[17] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine similarity scoring without score nor- malization techniques,” in Odyssey Speaker and Lan- guage Recognition Workshop. Odyssey, 2010, pp. 1–5.  
[24] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification,” in Annual Conference of the International Speech Communication Association (Interspeech), 2009, pp. 1559–1562.

We also made binding of Support Vector Machines [29, 1] available by simply compiling the LibSVM toolkit [30] and placing a copy of the library in SIDEKIT’s directory. Nuisance Attribute Projection (NAP)[31] commonly used for speaker verification was implemented as well.  
我们还通过编译LibSVM工具包[30]并在SIDEKIT的目录中放置库的副本，使支持向量机的绑定[29，1]变得可用。实现了常用于说话人验证的干扰属性投影（NAP）[31]。

Models and classifiers available in SIDEKIT cover the standard development for speaker recognition.  
SIDEKIT中提供的模型和分类器涵盖了说话人识别的标准开发。

3.3. Evaluation and visualization  
3.3条。评估和可视化

Based on the BOSARIS toolkit, SIDEKIT includes tools to compute Equal Error Rate (EER), Decision Cost Function (DCF) and minimum-DCF and plot two types of Detection Error Trade-off (DET) curves: steppy (from the ROC) and ROC Convex-Hull. It is also possible to indicate points on the curve beyond which miss rate and false alarm rates are not reliable. Figures 1 and 2 were generated using SIDEKIT.

SIDEKIT基于BOSARIS工具箱，包括计算等错误率（EER）、决策成本函数（DCF）和最小DCF的工具，并绘制两种类型的检测错误权衡（DET）曲线：steppy（来自ROC）和ROC凸包。也可以在曲线上指出漏报率和误报率不可靠的点。