说话人识别

2019 年 1 月 9 日

1 什么是说话人识别

1.1 基本概念

语音信号作为人类信息交互的一种载体，其在形式简单的一维信号中蕴含着丰富的信息(“形简意丰”)，包括语言信息(如语音内容)、副语言信息(如音高、音量、语调等)以及非语言信息(如健康状况、性别、年龄、环境背景等)[1]。声纹作为语音信号中的一种信息，是对语音信号中所蕴含的能表征说话人身份的语音特征以及基于这些特征所建立的语音模型的总称 [2]。由于不同说话人在讲话时所使用的发声器官(如舌头、口腔、鼻腔、声带、肺等)在尺寸和形态等方面均有所不同，再考虑到不同说话人在年龄、性格、语言习惯等因素上的差异，使得不同说话人的发音容量和发音频率等特性大不相同。可以说，任何两个人的声纹特性都不尽相同 [3]。

说话人识别(SRE)，又称声纹识别(VPR)，就是基于语音信号中的声纹信息，利用计算机以及各种信息识别技术，自动地实现说话人身份识别的一种生物特征识别技术 [4, 5, 6]。说话人识别主要由训练和识别两个阶段组成，下图 1 是一个基本的说话人识别系统框架 [5, 7]：

1. 训练阶段：首先对使用系统的说话人预留充足的语音，并提取该语音中的声纹特征，然后根据说话人的声纹特征训练得到说话人模型，最后将全部说话人模型构成系统的说话人模型库。

2. 识别阶段：说话人在进行识别认证时，系统对待识别语音进行与训练阶段相同的声纹特征提取过程，并将声纹特征与说话人模型库进行比对，得到对应的相似性打分，最后根据相似性打分判决待识别语音的说话人身份。

1.2 任务分类

从不同的分类角度看，说话人识别可大致分为以下几类。

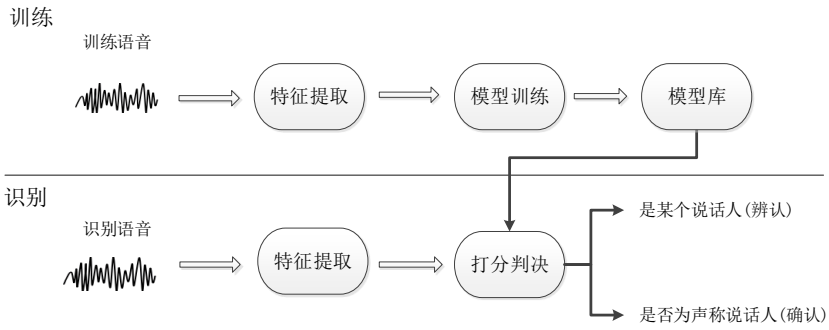

Figure 1: 一个基本的说话人识别系统框架

1. 说话人辨认和说话人确认

从实际应用的范畴，说话人识别可分为说话人辨认和说话人确认。说话人确认是确定待识别语音是否来自其所声称的目标说话人，是一个“一对一”的判决问题；说话人辨认是判定待识别语音属于目标说话人集合中哪一个说话人，是一个“多选一”的选择问题。此外，根据测试范围的不同，说话人辨认又可划分为闭集辨认和开集辨认。闭集辨认是指待识别语音必定属于目标说话人集合中的某一个说话人；而开集辨认是指待识别语音不受限于目标说话人集合，其可属于该集合外的某一位说话人。

除此之外，在实际应用中，说话人识别还涵盖了说话人检测(即检测目标说话人是否在某段语音中出现) 和说话人追踪(即以时间为索引，实时检测每段语音所对应的说话人) [8]等。

2. 文本相关、文本无关和文本提示

从发音文本的范畴，说话人识别可分为文本无关、文本相关和文本提示三类 [5, 6, 7]。文本无关是指说话人识别系统对于语音文本内容无任何要求，说话人的发音内容将不受任何限制，只要语音达到一定时长即可；而文本相关则要求用户需按照预先指定的固定文本内容进行发音。对比这两类说话人识别，文本相关的说话人识别的文本内容匹配性明显优于文本无关的说话人识别，所以一般来说其系统性能也会相对好很多。但是，文本相关对说话人预留和识别时的语音录制有着更为严格的限制，并且相对单一的识别文本更容易被误闯。相比于文本相关，文本无关的说话人识别使用起来更加方便灵活，具有更好的体验性和推广性。为此，综合二者的优点，文本提示型的说话人识别应运而生。对文本提示而言，系统从说话人的训练文本库中随机地抽取组合若干词汇，作为用户的发音提示。这样不仅降低了文本相关所存在的系统闯入风险，提高了系统的安全性，而且实现起来也相对简单。

1.3 评价指标

根据说话人识别任务的不同，其系统性能的评价指标也略有不同。

1. 说话人确认系统的性能指标

说话人确认系统的性能评价主要依据两个参量，分别是错误接受率(FAR)和错误拒绝率(FRR)。FAR 是指将非目标说话人误判为目标说话人而产生的错误。FRR 是指将目标说话人误识成非目标说话人而产生的错误。在说话人确认系统中，可通过设定不同的阈值对FAR 和FRR 进行权衡，并采用检测错误权衡(DET) 曲线 [9]来反映两个错误率之间的关系。在DET 曲线上，第一象限的角平分线与其交点之处的FAR 和FRR 值相等，通常称该交点所对应的错误率称为等错误率(EER)。EER 代表了说话人确认系统的一个整体性能，其越小系统性能相对越好，是衡量系统性能的一个重要参数。

2. 说话人辨认系统性能指标

通常情况下，在开集说话人辨认系统中仍可采用等错误率(EER) 和检测代价函数(DCF) 作为系统性能的评价指标。在闭集说话人辨认系统中通常采用前N 辨认正确率(Top-N IDR) 作为评价系统性能的指标。识别率是指待识别语音从目标说话人集合中正确地找出所对应真实说话人的比率。通常将待识别语音与目标说话人集合中相似度最大的说话人作为辨认说话人，其辨认正确的比率称为Top-1 辨认正确率(Top-1 IDR)。同理，若在目标说话人集合上相似度最大的前n 个说话人中包含真实说话人即认为辨认正确，则由此统计出来的辨认正确率称为Top-n 辨认正确率(Top-n IDR)。

2 发展历史

“闻其声而知其人”，通过人耳听觉感知来辨别声音中的说话人身份，古已有之。下图 2 总结了说话人识别技术的发展历史 [3, 6]。

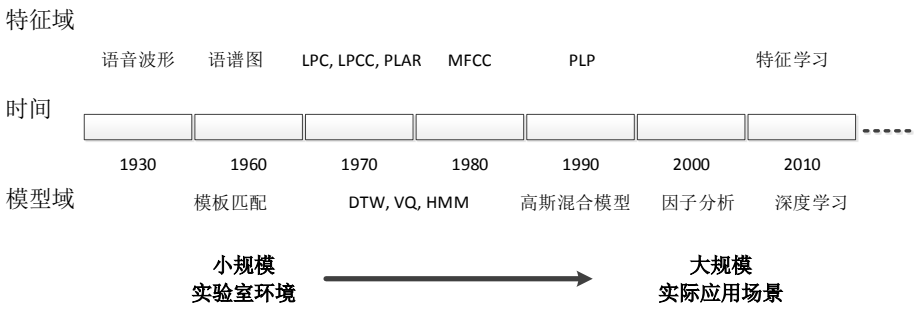

Figure 2: 说话人识别技术的发展历史

以语音作为身份认证的手段，最早可追溯到17世纪60年代英国查尔斯一世之死的案件审判中。1945年，Bell实验室的L. G. Kesta等人借助肉眼观察，完成语谱图匹配，并首次提出了“声纹”的概念；并随后在1962年第一次介绍了采用此方法进行说话人识别的可能性。随着研究手段和计算机技术的不断进步，说话人识别逐步由单纯的人耳听辨转向基于计算机的自动识别。从语音信号处理的角度，研究者们提出了倒谱 [10]、共振峰 [11]、基频轮廓 [12]等特征，并将其应用于说话人识别中，取得了不错的效果。

从20世纪70年代至80年代，有效的声学特征参数和模式匹配方法成为说话人识别的研究重点。研究者们相继提出了线性预测编码(LPC) [13]、线谱对(LSP) [14, 15]、线性预测倒谱系数(LPCC) [16]、感知线性预测(PLP) [17]、梅尔频率倒谱系数(MFCC) [18]等一系列声学特征参数。与此同时，动态时间规整(DTW) [19]、矢量量化(VQ) [20]、隐马尔科夫模型(HMM) [21]等已在语音识别领域得到广泛运用的技术，也逐渐成为说话人识别的重要技术。

20世纪90年代以来，尤其是D. Reynolds对高斯混合模型(GMM) [22]做了详细介绍后，基于最大似然的概率统计模型GMM迅速成为了文本无关说话人识别中的主流技术，将说话人识别研究带入了一个新的阶段。2000年，D. Reynolds在说话人确认任务中提出了高斯混合模型-通用背景模型(GMM-UBM)结构 [23]，为说话人识别技术从实验室走向实用作出了重要贡献。

进入21世纪，在传统GMM-UBM方法上，P. Kenny、N. Dehak等人先后提出了联合因子分析(JFA) [24] 和i-vector模型 [25]，将说话人模型映射到低维子空间中，得到了一个低维的说话人向量表示。在i-vector模型后端还可以通过类内协方差归一化(WCCN) [26]、扰动属性投影(NAP) [27]、线性判别分析(LDA) [25, 28] 、概率线性判别分析(PLDA) [29, 30]等方法，进一步去除与说话人无关的会话信息，从而提高了i-vector对说话人的区分能力。近年来，随着深度学习在语音识别等语音信号处理领域的快速发展和成功应用，基于深度学习的相关方法 [31, 32, 33, 34, 35, 36]也逐渐应用到说话人识别中，并取得了不俗的效果。

3 研究领域

4 小结

References

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[2] 中华人民共和国电子行业标准, "自动声纹识别(说话人识别)技术规范," Tech. Rep. SJ/T 11380-2008, 2008.

[3] 吴朝晖, *说话人识别模型与方法*. 清华大学出版社, 2009.

[4] N. Lass, *Contemporary issues in experimental phonetics*. Elsevier, 2012.

[5] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[6] T. F. Zheng and L. Li, *Robustness-Related Issues in Speaker Recognition*. Springer, 2017.

[7] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.

[10] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[11] G. R. Doddington, J. L. Flanagan, and R. C. Lummis, "Automatic speaker verification by non-linear time alignment of acoustic parameters," 1972, uS Patent 3,700,815.

[12] B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America*, vol. 52, no. 6B, pp. 1687–1697, 1972.

[13] J. Makhoul and L. Cosell, "Lpcw: An lpc vocoder with linear predictive spectral warping," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.*, vol. 1. IEEE, 1976, pp. 466–469.

[14] F. Zheng, Z. Song, L. Li, W. Yu, F. Zheng, and W. Wu, "The distance measure for line spectrum pairs applied to speech recognition," in *Fifth International Conference on Spoken Language Processing*, 1998.

[15] M. Sahidullah, S. Chakroborty, and G. Saha, "On the use of perceptual line spectral pairs frequencies and higher-order residual moments for speaker identification," *International Journal of Biometrics*, vol. 2, no. 4, pp. 358–378, 2010.

[16] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460–475, 1976.

[17] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[18] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[20] D. Burton, J. Shore, and J. Buck, "A generalization of isolated word recognition using vector quantization," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 1021–1024.

[21] B. Schuster-Böckler and A. Bateman, "An introduction to hidden markov models," *Current protocols in bioinformatics*, vol. 18, no. 1, pp. A–3A, 2007.

[22] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[24] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[25] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[26] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Ninth international conference on spoken language processing*, 2006.

[27] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for svm speaker recognition." in *Odyssey*, vol. 4. Citeseer, 2004, pp. 219–226.

[28] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted lda for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[29] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[30] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[31] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[32] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014, pp. 293–298.

[33] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[34] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 171–178.

[35] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[36] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.