噪声环境下稳健的说话人识别特征研究

程小伟，王 健，曾庆宁，谢先明，龙 超
(桂林电子科技大学信息与通信学院，广西桂林 541004)

摘要：针对噪声环境下说话人识别率较低的问题，提出一种基于正规化线性预测功率谱的说话人识别特征。首先对语音信号线性预测分析和正规化处理求出语音频谱包络，然后通过伽马通滤波器组得到对数子带能量，最后对特征参数进行离散余弦变换，得到了一种说话人识别特征正规化线性预测伽马通滤波器倒谱系数(Regularized Linear Prediction Gammatone Filter Cepstral Coefficient, RLP-GFCC)。仿真结果表明，在噪声环境说话人辨认试验中，相比传统特征美尔频率倒谱系数(Mel Frequency Cepstral Coefficient，MFCC)和伽马通滤波器倒谱系数(Gammatone Filter Cepstral Coefficient，GFCC)的系统识别率得到了明显提高，对噪声环境的鲁棒性得到了增强。

关键词：线性预测；正规化；说话人识别；伽马通滤波器组；鲁棒性
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1000-3630(2017)-05-0479-05
DOI 编码：10.16300/j.cnki.1000-3630.2017.05.014

A study of robust speaker recognition feature under noisy environment

CHENG Xiao-wei, WANG Jian, ZENG Qing-ning, XIE Xian-ming, LONG Chao
(*School of Information and Communication*, *Guilin University of Electronic Technology*, *Guilin* 541004, *Guangxi*, *China*)

Abstract: In order to solve the problem that speaker recognition rate is low under noisy environment, a speaker recognition feature based on regularized linear predictive power spectrum is proposed. The method uses linear prediction analysis and regularization of speech signal to get speech spectral envelope and then to get logarithmic sub-band energy through the Gammatone filter group, and finally uses discrete cosine transform to compute feature parameters to get a kind of speaker recognition feature named regularized linear predicted Gammatone filter cepstral coefficients (RLP-GFCC). The simulation results show that the recognition rate of the system is significantly improved in comparison with the systems of traditional feature MFCC and GFCC under noisy environment, and the robustness of the system to noise environment is improved.

Key words: linear prediction; regularization; speaker recognition; Gammatone filter bank; robustness

0 引 言

说话人识别技术是一种重要的生物特征识别技术，应用于身份确认、信息安全、远程控制等领域[1]。如何提取有效的说话人识别特征是识别技术的关键，说话人识别特征要能够描述说话人声道特性，有较高的区分度，对外界环境具有较强的鲁棒性[2]。

线性预测理论应用于语音信号处理，能够提供说话人的声道模型[3]，因此，线性预测系数(Linear Prediction Coefficient，LPC)成为比较普遍的说话人识别特征，基于线性预测理论的特征线性预测倒谱系数(Linear Prediction Cepstral Coefficient，LPCC)[4]能够用于说话人识别特征。这些特征在安静环境下能够取得很高的识别率，但对噪声环境的鲁棒性却很差。梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient，MFCC)[5]是语音识别和说话人识别最有效的特征之一，该特征基于听觉模型，对噪声环境具有一定的鲁棒性，但在低信噪比环境下识别率仍然较低。为应对噪声环境下说话人系统识别率较低的问题，研究人员通过对基于人耳耳蜗听觉模型伽马通滤波器的研究，提出了用于说话人识别的特征——伽马通滤波器倒谱系数(Gammatone Filter Cepstral Coefficient，GFCC)[6]，经实验论证该特征在不同背景噪声环境下可取得比 MFCC 更好的识别率。

为了进一步提高说话人识别系统对噪声环境

收稿日期：2016-12-06；**修回日期**：2017-04-01
基金项目：国家自然科学基金项目(61461011)；教育部重点实验室 2016 年主任基金项目资助(CRKL160107)；广西自然科学基金(2014 GXNSFBA118273)项目。
作者简介：程小伟(1990－)，男，河南漯河人，硕士研究生，研究方向为语音增强和说话人识别。
通讯作者：龙超，E-mail: chengzai05@163.com

的鲁棒性，结合线性预测分析理论和伽马通滤波器组的特殊性质，本文提出一种新的说话人识别特征，先求出语音信号的线性预测功率谱，并对线性预测功率谱进行正规化处理[7]，得到的频谱代替传统傅里叶变换功率谱，最后结合特征 GFCC 的提取方法，得到说话人识别特征正规化线性预测伽马通滤波器倒谱系数(Regularized Linear Prediction Gammatone Filter Cepstral Coefficient, RLP-GFCC)，仿真实验表明，该特征在噪声环境下能够取得比 GFCC 和线性预测伽马通滤波器倒谱系数(Linear Prediction Gammatone Filter Cepstral Coefficient, LP-GFCC)更好的系统识别率。

1 语音信号短时功率谱

1.1 线性预测功率谱

传统语音信号功率谱是通过对语音信号进行加窗分帧，然后对每帧语音信号进行离散傅里叶变换得到其频谱，即通过式(1)实现：

$$S_{\mathrm{FFT}}(f)=\left[\sum_{n=0}^{N-1}w(n)x(n)\mathrm{e}^{-\mathrm{j}2\pi nf/N}\right]^2 \tag{1}$$

其中：f为离散频率；$w(n)$和 $x(n)$分别为加窗函数和语音采样信号；N 为离散傅里叶变换点数。本文实验中加窗函数 $w(n)$采用汉明窗。

线性预测分析是一种有效的语音分析技术，在语音识别和说话人识别中得到了广泛应用[8-10]。线性预测的基本原理是语音信号的一个采样点 $x(n)$可以通过过去 p 个采样点的加权和来估计[11]，即 $\hat{x}(n)=-\sum_{k=1}^{p}a_k x(n-k)$，使预测误差的平方和 $e^2(n)=(x(n)-\hat{x}(n))^2=(x(n)+\sum_{k=1}^{p}a_k x(n-k))^2$ 达到最小，其中 p 为线性预测阶数，α_k 为加权系数。

一般通过自相关方法[12]求取线性预测系数，即通过式(2)求得：

$$\boldsymbol{a}_{\mathrm{opt}}^{\mathrm{LP}}=-\boldsymbol{R}_{\mathrm{LP}}^{-1}\boldsymbol{r}_{\mathrm{LP}} \tag{2}$$

其中：$\boldsymbol{R}_{\mathrm{LP}}$ 为托普利茨自相关矩阵，$\boldsymbol{r}_{\mathrm{LP}}$ 为自相关向量。通过对线性预测系数进行傅里叶变换求得线性预测功率谱，即通过式(3)求得：

$$S_{\mathrm{LP}}(f)=\frac{1}{\left|1+\sum_{k=1}^{p}a_k\mathrm{e}^{-\mathrm{j}2\pi fk}\right|^2} \tag{3}$$

线性预测功率谱比传统离散傅里叶变换频谱更加光滑，能够较好地表示语音信号的频谱包络，同时能提供说话人的声道模型。

1.2 正规化线性预测功率谱

L. Anders Ekman[7]等在 2008 年提出了语音信号的正规化线性预测，正规化线性预测比传统线性预测能更好地描述语音信号的频谱包络。

对于线性预测正规化，在线性预测的基础上，通过添加补偿函数 $\phi(\boldsymbol{b})$，使得 $\sum_n(x(n)+\sum_{k=1}^{p}b_k x(n-k))^2+\lambda\phi(\boldsymbol{b})$ 达到最小值。其中 λ 是大于零的常数，λ 能够控制频谱包络的平滑程度，$\boldsymbol{b}$ 为未知的线性预测向量。$\phi(\boldsymbol{b})$ 可以通过式(4)表示：

$$\phi(\boldsymbol{b})=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{DFDb} \tag{4}$$

其中：$\boldsymbol{D}$ 为对角矩阵，且对角元素为元素所在行数，$\boldsymbol{F}$ 是与自相关序列对应的托普利茨矩阵，$\boldsymbol{F}$ 可由 $f(m)=r(m)v(m)$ 计算得到，其中 $r(m)$ 为自相关序列，$v(m)$ 为加窗函数，文献[13-14]中介绍了加窗函数的种类和使用方法，本文实验中 $v(m)$ 所用的加窗函数为双自相关序列，在加性噪声环境下，该加窗函数对频谱包络进行估计，对噪声具有一定鲁棒性。

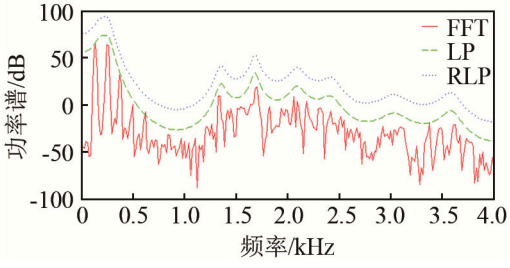
正规化线性预测系数通过式(5)求得：

$$\boldsymbol{b}_{\mathrm{opt}}^{\mathrm{RLP}}=-(\boldsymbol{R}_{\mathrm{LP}}+\lambda\boldsymbol{DFD})^{-1}\boldsymbol{r}_{\mathrm{LP}} \tag{5}$$

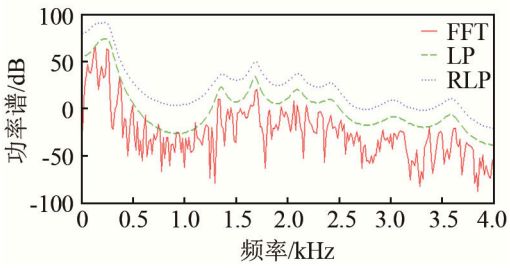
由预测系数 $\boldsymbol{b}_{\mathrm{opt}}^{\mathrm{RLP}}$ 可得正规化线性预测功率谱为

$$S_{\mathrm{RLP}}(f)=\frac{1}{\left|1+\sum_{k=1}^{p}b_k\mathrm{e}^{-\mathrm{j}2\pi fk}\right|^2} \tag{6}$$

图 1 运用 FFT、LP 和 RLP 三种频谱分析方法生成了频谱对比图，使用的语音来自 TIMIT 语音库，图 1(b)为图 1(a)中同一帧语音加 0 dB 信噪比的机枪(machinegun)噪声。LP 和 RLP 所用阶数为 p=20，RLP 中参数 λ=10^{-10}，为了便于观察，RLP 频谱上移 20 dB。从图 1 中可以看出，LP 谱和 RLP



(a) 纯净语言



(b) 带噪语言

图 1 纯净语音与带噪语音频谱对比图

Fig.1 Comparison of spectrum between clean speech and noisy speech

能够体现出短时语音信号的共振峰特性和频谱包络。正规化线性预测通过补偿方法处理非光滑部分，比传统线性预测频谱包络的估计失真低。

2　基于线性预测功率谱的特征提取过程

特征 LP-GFCC 和 RLP-GFCC 提取过程如图 2 所示：

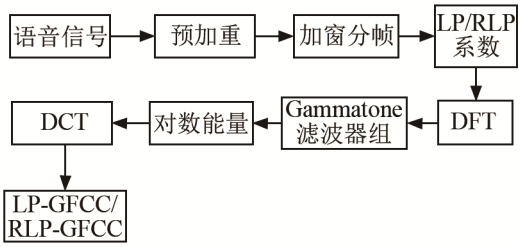


图 2　鲁棒性特征提取流程图

Fig.2　Flow chart of robust feature extraction

首先对语音信号进行预加重处理，通过高通滤波器，提升高频分量；然后利用语音信号的短时平稳性，对语音信号进行加窗分帧，本文采用汉明窗；利用上述计算方法求取 LP 或 RLP 系数，按照式(3)或式(6)对每组预测系数进行离散傅里叶变换(Discrete Fourier Transform，DFT)，得到的能量谱通过 64 通道的伽马通滤波器组[15]，对子带能量取对数，最后对子带对数能量进行离散余弦变换(Discrete Cosine Transform，DCT)，得到特征 LP-GFCC 或 RLP-GFCC。

3　实验分析

本文所用的基线系统是与文本无关的说话人辨认系统，实验使用的语音来自 TIMIT 语音库[16]，采样率是 16 kHz，单通道录音，采样精度为 16 bit，从中选取 85 个说话人(其中男 45 人，女 40 人)，每一个说话人有 10 句语音段，每段语音时长约 3 s。训练模型使用 7 句语音，测试使用 3 句语音，总共测试语音 255 句。说话人识别训练模型采用高斯混合模型(Gaussian Mixture Model，GMM)。实验所用噪声来自 noisex-92 噪声库，语音信号信噪比设为 −5、0、5、10、15、20、25、30 dB。

高斯混合模型阶数由说话人辨认样本数量决定，本次实验样本数量较少，阶数过高会造成过拟合使识别率降低，阶数过低不能充分表达说话人的特征空间。实验使用的参数直接影响系统识别率，文献[2]的实验参数在说话人辨认实验中能够取得较好的识别率，因此本文采用文献[2]的实验参数，实验 1 在基线系统上对 GMM 阶数取值做了对比实验，GMM 阶数取 32 时，基线系统性能达到最好。语音信号预加重系数典型取值在 0.92～0.97 之间，本文取值 0.93，采用汉明窗加窗分帧，帧长为 32 ms，即 512 个采样点，帧移为 8 ms，即 128 个采样点。实验中的端点检测采用基音检测算法。特征 $MFCC_D$ 取 12 阶静态 MFCC 和一阶动态特征，总共 24 维特征参数。在提取说话人识别特征 GFCC 的过程中，采用 64 通道伽马通滤波器组，依照等效矩形带宽(Equivalent Rectangular Bandwidth，ERB)频率分布在 50 Hz 和 8 000 Hz 之间，对对数子带能量进行 DCT 之后，24 维系数作为实验所用的说话人特征。本文实验特征 LP-GFCC 和 RLP-GFCC 线性预测阶数为 20，RLP-GFCC 的参数 λ 取固定值 10^{-10}。

3.1　实验 1　GMM 阶数对基线系统影响

本文实验采用高斯混合模型进行说话人辨认实验，其中 GMM 阶数直接影响说话人识别系统。实验中采用高斯混合模型阶数分别为 4、8、16、32、64，测试语音采用纯净语音，特征使用 24 维 $MFCC_D$ 作为说话人识别系统特征，实验结果如表 1 所示。

表 1　GMM 阶数对基线系统的影响

Table 1　The effect of GMM order on baseline system

GMM 阶数	4	8	16	32	64
识别率/%	89.80	95.69	96.08	98.43	97.25

从表 1 可以看出，随着 GMM 阶数的增加，系统识别性能逐渐变好，当阶数为 32 时，识别性能最好，识别率达到 98.43%，随后开始降低。因此对于实验所用的基线系统，GMM 阶数取值 32 时，系统识别率达到最好，本文实验采用 32 阶 GMM。

3.2　实验 2　平稳噪声环境识别结果

为了验证平稳噪声环境下特征 LP-GFCC 和 RLP-GFCC 识别的鲁棒性，分别用 $MFCC_D$、GFCC、LP-GFCC 和 RLP-GFCC 做仿真实验，四种特征均为 24 维，平稳噪声选用白噪声，信噪比设为 30、25、20、15、10、5、0 dB。系统识别率如表 2 所示。

从表 2 可以看出，特征 LP-GFCC 和 RLP-GFCC 在噪声环境下系统识别率优于 $MFCC_D$，在高噪声环境下系统识别率稍差于 GFCC，在低信噪比时识别率明显优于特征 GFCC，RLP-GFCC 特征对噪声的鲁棒性优于 LP-GFCC。在 0 dB 噪声环境下，四种特征在系统中识别率都很低，在 15 dB 白噪声环境下，特征 RLP-GFCC 的识别率较特征 MFCC、

表 2 白噪声环境下的特征识别率

Table 2 Feature recognition rate in white noise environment

信噪比/dB	识别率/%			
	$MFCC_D$	GFCC	LP-GFCC	RLP-GFCC
30	91.37	98.82	97.65	97.65
25	78.04	96.47	95.29	92.94
20	45.88	80.78	83.53	84.31
15	18.82	52.55	56.86	60.78
10	10.98	24.31	21.18	26.27
5	7.45	3.53	7.06	11.37
0	3.52	2.35	1.57	2.75

GFCC 和 LP-GFCC 分别提高了 41.96%、8.23%和 3.92%。

3.3 实验 3 非平稳噪声环境识别结果

为了验证非平稳噪声环境下特征 LP-GFCC 和 RLP-GFCC 识别的鲁棒性，同实验 2 的实验参数，从 noisex-92 噪声库选取 pink、babble、machinegun 噪声，信噪比设为 30、25、20、15、10、5、0、−5 dB。说话人识别系统仿真结果如图 3～5 所示。

从图 3～5 仿真结果可以看出，在不同信噪比噪声环境下，特征 LP-GFCC 和 RLP-GFCC 系统识别率明显高于传统特征 $MFCC_D$ 和 GFCC，特征 RLP-GFCC 系统识别率稍微高于 LP-GFCC，在 5 dB

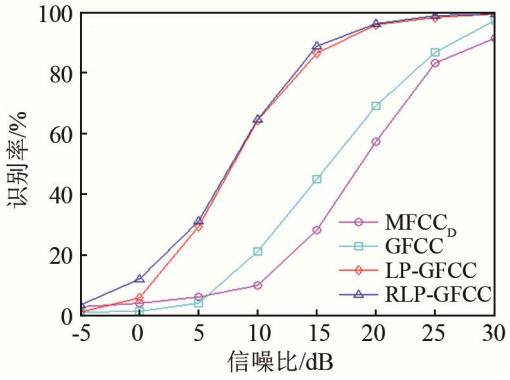


图 3 粉红噪声环境下的特征识别率

Fig.3 Feature recognition rate in pink noise environment

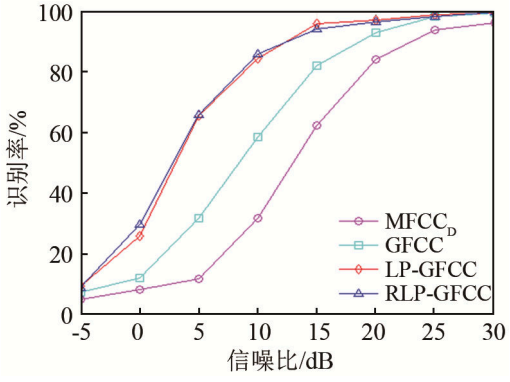


图 4 嘈杂噪声环下的境特征识别率

Fig.4 Feature recognition rate in babble noise environment

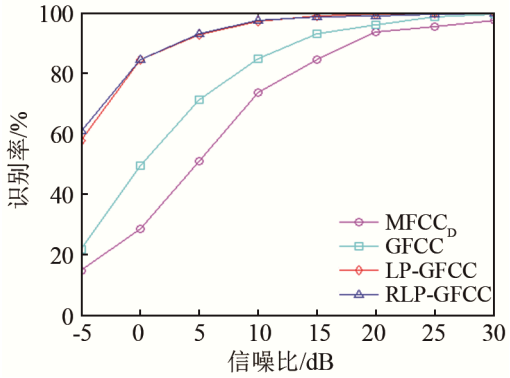


图 5 机枪噪声环境下的特征识别率

Fig.5 Feature recognition rate in machinegun noise environment

噪声环境下，LP-GFCC 平均识别率比传统特征 $MFCC_D$ 和 GFCC 分别高出 39.48%和 26.80%。由于本文实验在求取特征 RLP-GFCC 时，参数 λ 取固定值 10^{-10}，特征 RLP-GFCC 系统识别率稍微高于特征 LP-GFCC。文献[7]关于正规化线性预测功率谱，对参数 λ 提出了一种自适应方法，参数 λ 是随基音变化的数，能够减少传统线性预测对语音信号造成的失真。

3.4 实验 4 说话人识别特征计算时间对比

表 3 列举了特征 $MFCC_D$、GFCC、LP-GFCC、RLP-GFCC 的平均计算时间，测试语音时长为 5 s，每种特征测试 20 次，最后取平均时间。实验仿真软件平台为 Matlab R2014a，计算机 CPU 为酷睿 i3-2310，主频为 2.1 GHz。虽然特征 LP-GFCC 和 RLP-GFCC 的计算时间较 MFCC、GFCC 稍长，但在性能好的计算机实验平台上计算时间还会缩短，能够满足一定的实时性。在下一步的研究工作中，需要改进特征的计算复杂度，期望能够有更好的实时性能。

表 3 说话人识别特征计算时间对比结果

Table 3 Comparison of computation time between different speaker recognition features

特征参数	平均计算时间/s
$MFCC_D$	0.1689
GFCC	0.2762
LP-GFCC	1.4835
RLP-GFCC	2.0794

4 结 论

环境噪声对语音信号影响很大，不仅影响语音质量以及可懂度，而且造成语音识别和说话人识别系统识别率的迅速下降。本文通过结合线性预测分析理论和伽马通滤波器的特殊性质，提出了说话人

识别特征 LP-GFCC 和 RLP-GFCC，利用 TIMIT 语音库和 noisex-92 噪声库，Matlab 仿真实验表明，这两种特征在说话人识别系统中性能优于传统特征 MFCC 和 GFCC，提高了系统的说话人识别率和对噪声环境的鲁棒性。但 RLP-GFCC 的识别性能稍微优于特征 LP-GFCC，补偿参数 λ 对说话人识别系统的识别率影响较大，因此在后续的说话人识别研究工作中，可以引入相关文献中的自适应方法。

参 考 文 献

[1] 吴朝晖. 说话人识别模型与方法[M]. 北京：清华大学出版社, 2009.
WU Chaohui. The model and method of speaker recognition[M]. Beijing: Tsinghua University Press, 2009.

[2] 蒋晔. 基于短语音和信道变化的说话人识别研究[D]. 南京：南京理工大学, 2013.
JIANG Ye. Research on speaker recognition over short utterance and varying channels[D]. Nanjing: Nanjing University of Science and Technology, 2013.

[3] Pati D, Prasanna S R M. Processing of linear prediction residual in spectral and cepstral domains for speaker information[J]. International Journal of Speech Technology, 2015, **18**(3):1-18.

[4] 周燕, 胡志峰. 基于免疫聚类的 RBF 网络在说话人识别中的应用[J]. 声学技术, 2010, **29**(2): 184-187.
ZHOU Yan, HU Zhifeng. Application of immune algorithm based RBF network to human speaker recognition[J]. Technical Acoustics, 2010, **29**(2): 184-187.

[5] 林琳, 陈虹, 陈建. 基于鲁棒听觉特征的说话人识别[J]. 电子学报, 2013, **41**(3): 619-624.
LIN Lin, CHEN Hong, CHEN Jian. Speaker recognition based on robust auditory feature[J]. Acta Electronica Sinica, 2013, **41**(3): 619-624.

[6] 王玥, 钱志鸿, 王雪, 等. 基于伽马通滤波器组的听觉特征提取算法研究[J]. 电子学报, 2010, **38**(3): 525-528.
WANG Yue, QIAN Zhihong, WANG Xue, et al. An auditory feature extraction algorithm based on γ-tone filter-banks[J]. Acta Electronica Sinica, 2010, **38** (3): 525-528.

[7] Ekman L A, Kleijn W B, Murthi M N. Regularized linear prediction of speech[J]. IEEE Transactions on Audio Speech & Language Processing, 2008, **16**(1): 65-73.

[8] Bastys A, Kisel A, Alna B. The use of group delay features of linear prediction model for speaker recognition[J]. Informatica, 2010, **21**(1): 1-12.

[9] Bastys A, Kisel A, Alna B. The use of group delay features of linear prediction model for speaker recognition[J]. Informatica, 2010, **21**(1): 1-12.

[10] Saeidi R, Alku P, Backstrom T. Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch[J]. Audio Speech & Language Processing IEEE/ACM Transactions on, 2016, **24**(1): 42-53.

[11] Makhoul J. Linear prediction: a tutorial review. Proc IEEE 63: 561-580[J]. Proceedings of the IEEE, 1975, **63**(4): 561-580.

[12] 宋知用. MATLAB 在语音信号分析与合成中的应用[M]. 北京：北京航空航天大学出版社, 2013.
SONG Zhiyong. Application of MATLAB in speech signal analysis and synthesis[M]. Beijing: Beihang University Press, 2013.

[13] Shimamura T, Nguyen N D. Autocorrelation and double autocorrelation based spectral representations for a noisy word recognition system[C]// INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September. 2010.

[14] Hanilçi C, Kinnunen T, Ertaş F, et al. Regularized all-pole models for speaker verification under noisy environments[J]. IEEE Signal Processing Letters, 2012, **19**(3): 163-166.

[15] D. P. W. Ellis (2009). Gammatone-like spectrograms. http://www.ee.co-lumbia.edu/~dpwe/resources/matlab/gamatonegram/.

[16] Li Q, Reynolds D A. Corpora for the evaluation of speaker recognition systems[C]// Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference. IEEE Computer Society, 1999: 829-832.