目录

        - Deep autoencoder
3. 性能评估
    - accu
    - precision
    - recall
    - f-measure
    - equal error rate
    - receiver operating characteristics
    - rmse
    - 深度学习实现框架

图1：说话人识别应用的分类

开放集、闭合集、文本独立、文本依赖

图2：传统说话人识别过程

- data preprocessing
    - silence removal
    - pre-emphasis
    - spectrograms
    - endpoint detection
- dimensionality reduction
    - PCA
    - Sequential Forward Generation
    - Sequential Backward Generation
- feature extraction
    - DL feature
        - cnn
        - rnn
        - Restricted Boltzmann Machine
        - deep autoencoder
    - Shallow feature
        - time domain
        - frequency domain
- toolbox
    - opensmile, sidekit, jaudio, kaldi, yaafe, libxtract, librosa, essentia, pyaudioanalysis
- model
    - svm,gaussian mixture model,naive bayes,knn,ann
- measure
    - accuracy
    - precision
    - reclal
    - f-measure
    - auroc

LibriSpeech包括与英语演讲相关的音频文件，属于男性和女性英语使用者。该数据集中的所有话语都以16 kHz频率和16位比特率采样。该语料库包括五个不同的训练和测试集，用于开发SI模型。

图3：自动说话人识别系统的分类

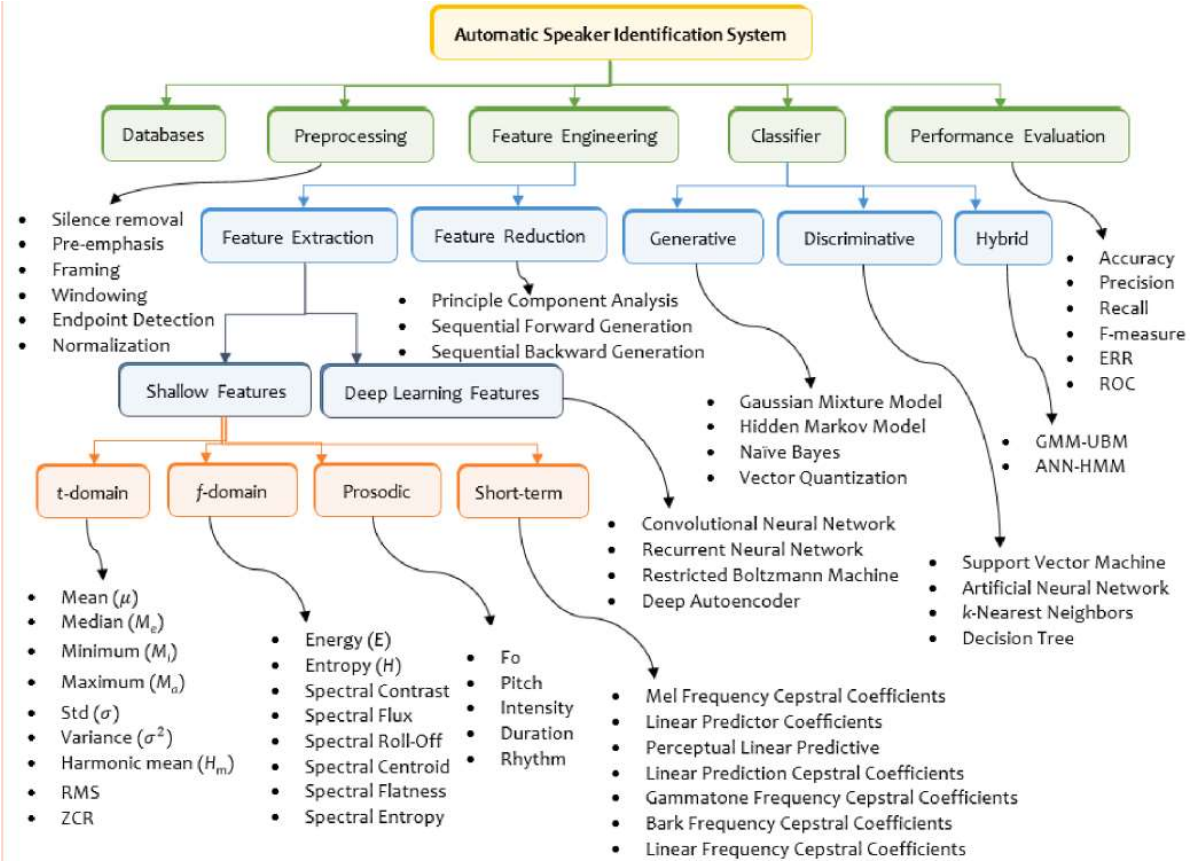

表1：现有review

标2：缩写

表3：语音数据库的特征

英语、size=11373，没有年龄信息，16bits/sample，1channel，sample rate=16khz，格式=flac，音源=安静

表4：各个研究中的预处理方法

Methods	Strengths	Weaknesses	Reference
Silence Removal	Reduce processing time and enhance the performance of SI system	Only suitable for white Gaussian noise	Avci (2009), Daqrouq (2011), Daqrouq and Tutunji (2015), Fan and Hansen (2010), Jawarkar et al. (2015), Lukic et al. (2016), Medikonda and Madasu (2018), Novotný et al. (2019), Sarma and Sarma (2013), Wu and Tsai (2011), Zhao et al. (2014)
Pre-emphasis	Negative spectral slope of the various voiced parts is improved which enhance the performance.	gain of low frequencies also become high.	An et al. (2019a), Avci (2009), Faragallah (2018), Liu, Wu, Li, Li, & Shen (2018), Medikonda and Madasu (2018), Renisha and Jayasree (2019), Sun, Gu, Xie, and Chen (2019)
Hamming Window	reduces errors in the estimation of distortion	huge variance of spectral estimation	Al-Rawahy et al. (2012a, 2012b), Avci (2009), Faragallah (2018), Liu et al. (2018), Mporas et al. (2016)
End-Point Detection	incorporate spectral information and minimize the computational resources	In low SNR and non-stationary environments, end-point detection often fails and performance of SI degrade dramatically	Avci (2009), Renisha and Jayasree (2019)
Speech Signals Normalization	Prevent an error estimate caused by a change in the volume of speakers.	increase in the magnitude of jitter (precision measure for displacement estimation)	Avci (2009), Daqrouq (2011), Daqrouq and Tutunji (2015), Renisha and Jayasree (2019), Wu and Lin (2009a)
Median Filtering	Reduces random noise, particularly when noise amplitude likelihood density has large tails.	For multidimensional signals it does not produce satisfactory results as compared to one-dimensional	Sarma and Sarma (2013)
Spectrogram	Express speech signal by combining the benefit of both time and frequency domains and represents the relationship of frequency, time, and energy amplitude directly.	poor frequency and time resolution	An et al. (2019a), Bunrit, Inkian, Kerdprasop, and Kerdprasop (2019), Imran, Haflan, Shahrebabaki, Olfati, and Svendsen (2019), Liu et al. (2018), Lukic et al. (2016), Yadav and Rai (2018)
Recursive Least Squares	small memory requirement, converge faster and easily changed into real-time systems	computationally intensive, prevent the classifier from going to sleep;	Dhakal et al. (2019)

图4：Number of subjects per gender comparison.

图5：综述的文章中的数据库出现频率

图6：语音信号框架和窗口(Jahangir等人，2020)

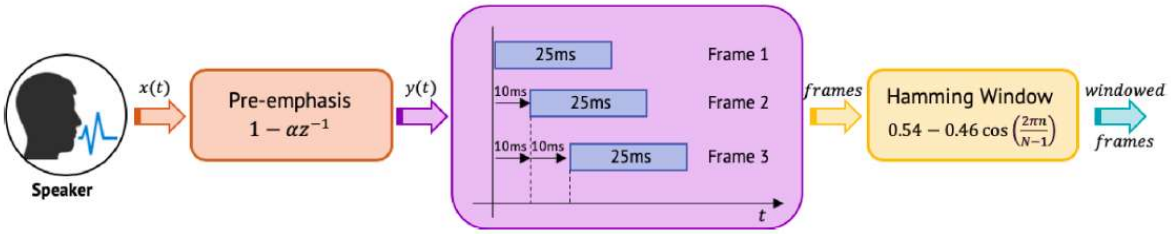

Fig. 6. Speech Signal Framing and Windowing (Jahangir et al., 2020).

表5：时域特征列表

Feature	Formula	Feature	Formula
Mean (μ)	$\bar{s} = 1/N\sum_{i=1}^{N} s_i$	Root mean square (R_{ms})	$rms = \sqrt{1/N\sum_{i=1}^{N}(s_i)^2}$
Median (M_e)	$median_i(s_i)$	Peak amplitude (P_a)	$\max(s_i) - \min(s_i)$
Minimum (M_i)	$min_i(s_i)$	Pitch angle (P_k)	$arctan\left(x_i/\sqrt{y^2 - x_i^2}\right)$
Maximum (M_a)	$max_i(s_i)$	Kurtosis (K_r)	$E\left[\left(s_i - \bar{s}\right)^4\right] \Big/ E\left[\left(s_i - \bar{s}\right)^2\right]^2$
Standard deviation (σ)	$\sigma = \sqrt{1/N\sum_{i=1}^{N}(s_i - \bar{s})^2}$	Skewness (S_k)	$E\left[\left(s_i - \bar{s}/\sigma\right)^3\right]$
Variance (σ^2)	$\sigma^2 = \frac{\sum(s_i - \bar{s})^2}{N}$	Signal power (S_p)	$\sum_{i=1}^{N} s_i^2$
Harmonic mean (H_m)	$1/N\sum_{i=1}^{N} 1/s_i$	Coefficient of variation (C_v)	σ_{si}/μ_{si}
Interquartile range (I_r)	$Q_3(s_i) - Q_1(s_i)$	Zero Cross Rate (ZCR)	$1/N\sum_{i=1}^{N}\lvert s_i - s_{i-1}\rvert$

表6：频域特征列表

Feature	Description	Formula	References
Energy (E)	The energy of the signal reflects the area under the square magnitude.	$\sum_{i=1}^{N}[s_i]^2/\boldsymbol{length}(s_i)$	(Yakovenko & Malychina, 2016)
Mean frequency (μF)	Mean frequency measures the mean normalized frequency of the power spectrum of a speech signal.	$\sum_{i=1}^{N}(\boldsymbol{is}_i(\boldsymbol{F}))/\sum_{j=1}^{N}\boldsymbol{s}_j(\boldsymbol{F})$	(Almaadeed et al., 2016)
Spectral Flux	Spectral Flux differentiate between normalized spectral magnitudes.	$\sum_{i=1}^{N}\left(Z^{(k)}[i] - Z^{(k-1)}[i]\right)^2 Z^{(k)}$ and $Z^{(k-1)}$ are the normalized magnitudes of Fourier transform at k and $k-1$ frames.	(Mannepalli, Sastry, & Suman, 2017)
Spectral Roll-Off	Spectral Roll-Off measures the spectral concentration less than threshold.	$\lambda\sum_{i=1}^{N}\lvert Z[i]\rvert\lambda \simeq 0.85$ (frequency where 85 percent of the speech signal power resides)	(Thoman, 2009)
Spectral Centroid	Spectral Centroid is the average frequency of speech signal weighted by magnitude.	$\sum_{i=1}^{N} i\times\lvert Z[i]\rvert/\sum_{i=1}^{N}\lvert Z[i]\rvert$	(Thoman, 2009)
Spectral Flatness	Spectral Flatness shows whether the distribution is spike or smooth.	$\left(\prod_{i=1}^{N}\boldsymbol{Z}[i]\right)^{1/N}\left(\sum_{i=1}^{N}\boldsymbol{Z}[i]\right)^{-1}$	(Kadiri, Prasad, & Yegnanarayana, 2020)
Spectral Entropy	Spectral Entropy calculate the regularity of power spectrum of speech signal.	$1/\log N\left(\sum_{i=1}^{N} P(Z[i])\log P(Z[i])\right)$	(Luque-Suárez, Camarena-Ibarrola, & Chávez, 2019, Shannon, 2001)
Spectral Contrast	Spectral Contrast represents the relative distribution of frequency instead of average frequency of speech signal.	$\log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} s_{k,i}\right) - \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} s_{k,N-i+1}\right)$ where N is a number in kth - sub-band, k ∈ [1,6]and$\alpha = 0.02$	(Jiang, Lu, Zhang, Tao, & Cai, 2002)

表7：韵律特征列表

Feature	Description	References
Fundamentalfrequency (F0)	F0 is the reciprocal of time interval between two consecutive glottal cycles.	(Ajmera et al., 2011; Sekkate et al., 2019; Tirumala et al., 2017)
Pitch	It is a perceptual property of the speech signal with physical characteristics denoted by the F0.	(Ajmera et al., 2011; Sekkate et al., 2019; Tirumala et al., 2017)
Intensity	Intensity is the measure of loudness or energy of a signal and related to amplitude square.	(Sekkate et al., 2019)

图7：MFCC特征提取技术

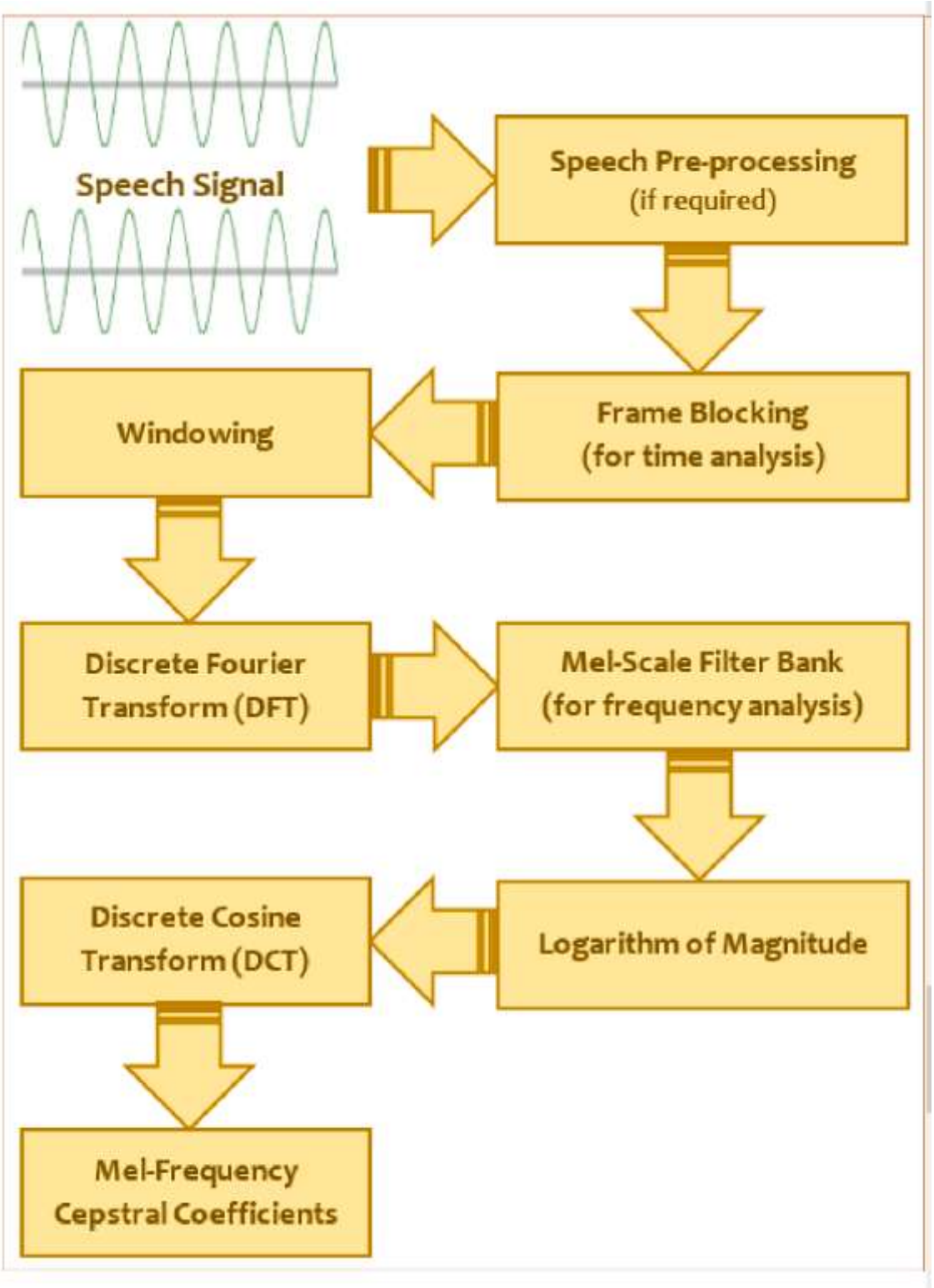

表8：文献中使用的特征提取方法

Feature extraction methods	Reference
MFCC + Spectrogram + log-mel Filterbank	An et al. (2019a)
MFCC + Spectrogram + MFSC	Imran et al. (2019)
MFCC + Spectrogram	Bunrit et al. (2019), Liu et al. (2018)
MFCC + Delta MFCC + Delta–Delta MFCC	Ali et al. (2018), Bisio et al. (2018), Dovydaitis and Rudžionis (2018), Soleymanpour and Marvi (2017)
MFCC + ZCR + Spectral Roll-off + Roughness + Brightness + Irregularity	Sardar and Shirbahadurkar (2018b)
MFCC + LFCC + LPC + ZCR + Spectral Roll-off + Brightness + Irregularity + Roughness	Sardar and Shirbahadurkar (2019)
MFCC + Delta MFCC + Delta–Delta MFCC + GFCC	Medikonda and Madasu (2018), Zhang et al. (2018)
MFCC + LPCC + DGS + DGCS	Sun et al. (2019)
MFCC + LPC + LBP	Abdul (2019)
MFCC + Spectral Roll-off + Brightness + Roughness + Irregularity	Sardar and Shirbahadurkar (2018a)
MFCC + ZCR + Spectral Centroid + Spectral Entropy + Spectral Flux + Spectral Spread + Spectral Roll-off + Energy + Entropy of Energy + Chroma Deviation + Chroma Vector	Mokgonyane et al. (2019)
MFCC + LPC	Almaadeed et al. (2016)
MFCC + DWT + WPT + WSBC	Almaadeed, Aggoun, and Amira (2015)
MFCC + LPCC + LPC residual + phase	Kawakami et al. (2014)
MFCC	Mporas et al. (2016), Tirumala and Shahamiri (2017)
MFCC + GFCC	Jawarkar et al. (2015), Zhao et al. (2014)
MFCC + LFCC	Fan and Hansen (2010)
MFCC + Spectral and Cepstrum periodicities	Al-Rawahy et al. (2012a, 2012b)
MFCC + Delta MFCC + Delta–Delta MFCC + Spectrogram	Chunlei Zhang, Koishida, and Hansen (2018a)
MFCC + CFCC + GFCC + RASTA + RASTA-PLP	Li and Gao (2016)
MFCC + Delta MFCC	Michalevsky et al. (2011)
MFCC + BFCC + PLP + RASTA-PLP	Abdalmalak and Gallardo-Antolín (2018)
MFCC + LPCC	Sadıç and Bilginer Gülmezoğlu (2011)
MFCC + IMMFC	Chakroborty and Saha (2009)
MKMFCCs	Faragallah (2018)
Hanman Transform (HT) + Type-2 Information Set (T2IS)	Medikonda and Madasu (2018)
Frame based linear predictive coding spectrum (FBLPCS)	Wu and Lin (2009a)
Spectrogram	Lukic et al. (2016), Yadav and Rai (2018)
Statistical Features (mean, median, interquartile range, standard deviation) + Gabor Filter + Spectrogram	Dhakal et al. (2019)
WPT + DWT	Wu and Lin (2009b)
LPC + DWT	Sarma and Sarma (2013)
Spectrogram + Radon transform	Ajmera et al. (2011)
Vocal resonant frequencies (F1, F2, F3) + Wavelet Packet Transform (WPT)	Daqrouq (2011)
Empirical mode decomposition (EMD)	Wu and Tsai (2011)
PWCC + PWPCC + SPWPCC	Renisha and Jayasree (2019)
Coiflet Wavelet	Indumathi and Chandra (2015)
Statistical Features (Min, Max, Mean, Median, Mode, Variance, Standard Deviation) + ZCR + RMS	Jahangir, Teh, Ishtiaq, Mujtaba, and Nweke (2018)
Pitch + intonation + timing + loudness	Manikandan and Chandra (2016)
Vocal resonant frequencies (F1, F2, F3, F4, F5) + Wavelet Packet Transform (WPT)	Daqrouq and Tutunji (2015)

表9：SID中的特征提取工具箱

	HTK	SIDEKIT	Open SMILE	Kaldi	pyAA	jAudio	YAAFE	Essentia	LibXtract	Librosa
Organization	Cambridge University	LIUM, Universite du Mans	audEERING GmbH	Microsoft Research	Behavioral Signals	QMUL	Telecom ParisTech	Pompeu Fabra University	–	NYU
Platform	Windows LINUX	Windows MacOS LINUX	Windows MacOS LINUX	Windows MacOS LINUX	Windows MacOS LINUX	Windows MacOS LINUX	MacOS LINUX	Windows MacOS LINUX	Cross-Platform	Cross-Platform
Open Source	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Access	Free[a]	Free[b]	Free[c]	Free[d]	Free[e]	Free[f]	Free[g]	Free[h]	Free[i]	Free[j]
Language Support	Python MATLAB C	Python	C++	C++	Python	Java	Python MATLAB C++	Python C++	C Java	Python
MFCC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LPC	✓	×	✓	×	×	×	✓	✓	×	×
LPCC	×	×	×	×	×	×	×	×	×	×
LSF	×	×	×	×	×	×	✓	×	×	×
PLP	✓	×	✓	✓	×	×	×	×	×	×
LFCC	×	✓	×	×	×	×	×	×	×	×
ZCR	×	×	✓	×	✓	✓	✓	✓	✓	✓
HNR	×	×	✓	×	×	×	×	✓	×	×
RMS	×	×	×	×	×	✓	×	×	×	✓
MoM	×	×	×	×	×	✓	×	×	×	×
Energy	×	×	✓	×	✓	×	✓	✓	×	×
Formant	×	×	✓	×	×	×	×	×	✓	×
Tonal	×	×	✓	×	×	×	×	✓	×	✓
Spectral Flux	×	×	×	×	✓	✓	✓	✓	×	✓
Spectral ROF	×	×	×	×	✓	✓	✓	✓	×	✓
Spectral BW	×	×	×	×	×	×	×	×	×	✓
Mean	×	×	×	×	×	×	×	×	✓	×
Kurtosis	×	×	×	×	×	×	×	×	✓	×
S. Centroid	×	×	×	×	✓	×	×	✓	✓	✓
Spectrogram	×	×	×	×	✓	×	×	×	×	✓
Chroma	×	×	×	×	✓	×	×	×	×	✓
Output	HTK	SPRO4 HTK	Matrix CSV HTK ARFF	Matrix	CSV Matrix	XML ARFF	CSV HDF5	YAML JSON	VAMP XML	CSV TSV

**MoM = Method of Moments, Spectral ROF = Spectral Rolloff, S. Centr = Spectral Centroid, Spectral BW = Spectral Bandwidth, pyAA = pyAudioAnalysis, QMUL = Queen Mary University of London, NYU = New York University.

表10：文献中用的特征选择方法

Methods	Advantage	Disadvantage	Reference
PCA	reduce the number of dimensions without losing information	not scale-invariant	Ali et al. (2018), Indumathi and Chandra (2015), Sardar and Shirbahadurkar (2018a), Zhang et al. (2018a), Zhang et al. (2015)
SFG	best performs when the optimum feature subset is small	unable to subtract features that become redundant after adding other features	Sardar and Shirbahadurkar (2018a)
SBG	best perform when the optimum feature subset is big, as it spends most of time exploring large subsets	unable to re-evaluate the usefulness of a feature after its removal	Sardar and Shirbahadurkar (2018a)
UFSM	simple to understand, run and provides a better data understanding	Not always give optimized feature set	Dhakal et al. (2019)

图9：用DNN进行SI

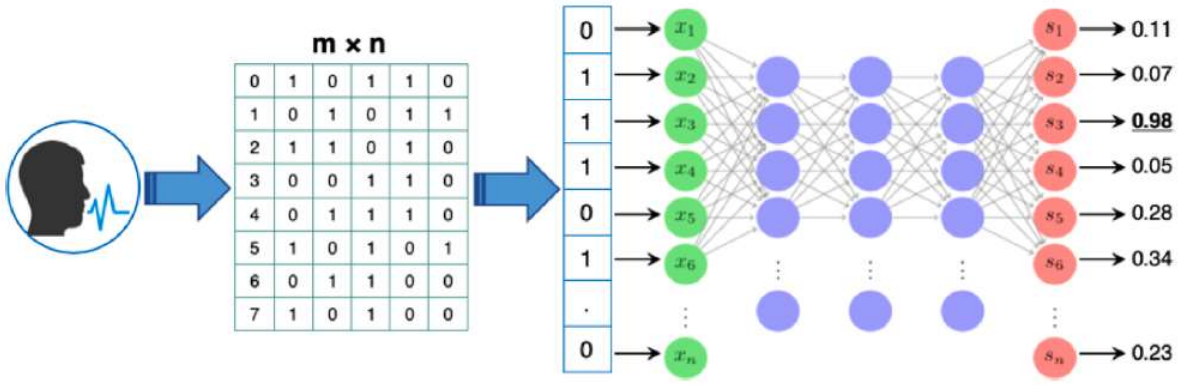

Fig. 9. Deep Neural Network Architecture for SI (Jahangir et al., 2020).

表12：用深度学习进行SI的技术

Technique	Strengths	Weaknesses	Reference
DBN	Effectively yield discriminative features that estimate the complex non-linear dependency among features in each speech sample.	slow learning process and labeled dataset is required for training	Ali et al. (2018), Sun et al. (2019)
RBM	Automatically identify relationships between features, describing them in its weights which can be translated as basis vectors.	Not suitable for massive databases because of training and computational complexity.	Ghahabi and Hernando (2018)
DAE	Useful for unbalanced speech database and able to learn effective non-linear features.	Overfitting can occur because network parameters are more than input data.	Novotný et al. (2019), Tirumala and Shahamiri (2017), Zhang et al. (2015)
CNN	robust to noise, does not need pre-processing steps and capable to learn local correlation patterns in input features	Require massive amount of data and it is not efficient when modelling temporal dependencies in the speech signals.	Abdul (2019), An et al. (2019a), Bunrit et al. (2019), Hajavi and Etemad (2019), Imran et al. (2019), Liu et al. (2018), Lukic et al. (2016), Yadav and Rai (2018), Zhang et al. (2018a)
LSTM	deal with variable length inputs and effectively deal with vanishing gradient problem	Long processing time and computational cost because of updating a number of parameters.	Dovydaitis and Rudžionis (2018), Jung et al. (2018), Larsson (2014)
DNN	Excellent ability to deeply manipulate high-dimensional data.	Sensitive to the variation of input data and training process is slow	Dhakal et al. (2019), Zhang et al. (2015)

表14：文献中使用评估标准的情况

References	Accuracy	Precision	Recall	F1 Score	ERR	ROC	RMSE
Abdul (2019), Ali et al. (2018), An et al. (2019a), Bisio et al. (2018), Daqrouq (2011), Fan and Hansen (2010), Faragallah (2018), Jawarkar et al. (2015), Krobba et al. (2019), Medikonda and Madasu (2018), Mokgonyane et al. (2019), Renisha and Jayasree (2019), Sardar and Shirbahadurkar (2018a), Sardar and Shirbahadurkar (2018b), Sardar and Shirbahadurkar (2019), Soleymanpour and Marvi (2017), Sun et al. (2019), Wang et al. (2015), Wu and Tsai (2011), Zhang et al. (2015), Zhao et al. (2014)	✓	×	×	×	×	×	×
Imran et al. (2019), Jahangir et al. (2018), Manikandan and Chandra (2016)	✓	✓	✓	✓	×	×	×
Ghahabi and Hernando (2018), Hajavi and Etemad (2019), Jung et al. (2018), Liu et al. (2018), Lukic et al. (2016), Novotný et al. (2019), Yadav and Rai (2018), Zhang et al. (2018a), Zhang et al. (2018)	×	×	×	×	✓	×	×
Daqrouq and Tutunji (2015)	✓	×	×	×	✓	×	×
Indumathi and Chandra (2015)	✓	✓	✓	×	×	×	✓
Almaadeed et al. (2015)	✓	×	×	×	×	✓	×
Ajmera et al. (2011), Tirumala and Shahamiri (2017)	✓	×	×	×	×	×	✓
Abdalmalak and Gallardo-Antolín (2018)	×	×	×	×	×	✓	×

表15： 深度学习软件框架

Software	GPU	Lic. / Access	PLATFORM			Support	Supporting DL Techniques					Optimizer	Activation Function
			❖	LINUX			FCNN	AE	CNN	RNN	RBM		
Theano	✓	NumPy[a]	✓	✓	✓	Python	✓	DAE	✓	×	DBN	SGD	ReLU, Tanh SoftMax
Caffe	✓	BSD[b]	✓	✓	✓	Python MATLAB C++	×	×	✓	×	×	SGD	ReLU, Tanh Sigm, ELU
DeepLearn-Toolbox	✓	BSD[c]	✓	✓	×	MATLAB	✓	SAE, DAE	✓	×	DBN	BP	Sigmoid
MatConvNet	✓	BSD[d]	✓	✓	✓	MATLAB C++	×	×	✓	×	×	SGD	ReLU, Sigmoid
Tensorflow	✓	Apache[e]	✓	✓	✓	Python, C++	×	CAE	✓	✓	×	SGD BP	ReLU, Sigmoid
Keras	✓	MIT[f]	✓	✓	✓	Python	×	×	✓	✓	×	SGD ADAM RMSPrp	ELU, ReLU SoftMax SeLU, Tanh Sigmoid
CNTK	✓	MIT[g]	✓	✓	×	C++, C#, Python, Java	×	×	✓	✓	×	SGD	ReLU, Tanh ELU
PyBrain	✓	BSD[h]	✓	✓	✓	Python	✓	×	×	✓	✓	BP	Sigmoid
Torch	✓	BSD[i]	✓	✓	✓	Lua C	✓	✓	✓		×	SGD ADAM	ReLU, Tanh SoftMax
MDLTB	✓		✓	✓	✓	MATLAB	✓	SAE, DAE	✓	✓	×	ADAM	ReLU, PReLU
DL4J	✓	Apache[j]	✓	✓	✓	Scala Java	✓	✓	✓	✓	×	SGD ADAM RMSPrp	Tanh, ReLU SoftM, ELU Sigmoid
Chainer	✓	MIT[k]	✓	✓	✓	Python	✓	✓	✓	✓	×	SGD	ReLU
Neural Networks	✓	MIT[l]	✓	✓	✓	Java	✓	SAE, DAE	✓	×	RBM DBN	-	ReLU, Tanh SoftM, LRN Sigm
ConvNet	✓	BSD[m]	✓	✓	×	MATLAB	×	×	✓	×	×	IBP	ReLU, Sigm SoftMax
OpenNN	✓	GNU[n]	✓	✓	✓	Python C++	✓	×	×	×	×	-	-
RNNLIB	✓	BSD[o]	×	✓	✓	C++	×	×	×	✓	×	-	-
SimpleDNN	✓	mozilla[p]	✓	✓	✓	Kotlin	✓	×	×	✓	×	-	ELU, SoftMax

** MDLTB = MATLAB Deep Learning Toolbox, DL4J = DeepLearning4J, SoftM = SoftMax, Sigm = Sigmoid.