第 38 卷 第 1 期
2019 年 1 月

应用声学
Journal of Applied Acoustics

Vol. 38, No. 1
January, 2019

◇ 研究报告 ◇

连续音素的改进深信度网络的识别算法*

阴法明[1†] 赵 焱[2] 赵 力[2]

(1 南京信息职业技术学院通信学院 南京 210023)

(2 东南大学信息科学工程学院 南京 210096)

摘要 为提高连续语音识别中的音素识别率，提出一种基于改进并行回火训练的受限波尔兹曼机的音素识别算法。首先，利用经过等能量划分后的改进并行回火算法来训练受限玻尔兹曼机，接着将受限玻尔兹曼机堆叠组成一个深信度网络，从而作为深度神经网络预训练的基础模型，然后通过 softmax 层输出，得到用于音素状态后验概率检测的深度神经网络。接着，利用少量的标签数据，根据反向传播算法对网络权重进行微调。最后，将所得后验概率作为隐马尔科夫的发射概率，然后利用 Viterbi 解码器实现音素识别。在 TIMIT 语料库上的实验表明，识别率相比于传统的对比散度类算法提高了约 4.5%，在不增加计算量的情况下比原始并行回火算法提高约 1%。

关键词 并行回火，受限玻尔兹曼机，深信度网络，音素识别

中图法分类号: TP18 **文献标识码:** A **文章编号:** 1000-310X(2019)01-0039-06

DOI: 10.11684/j.issn.1000-310X.2019.01.006

Phoneme recognition based on deep belief network

YIN Faming[1] ZHAO Yan[2] ZHAO Li[2]

(1 *Nanjing College of Information Technology, Nanjing* 210023, *China*)

(2 *School of Information Science and Engineering, Southeast University, Nanjing* 210096, *China*)

Abstract In order to improve the accuracy of phoneme recognition in continuous speech recognition, in this paper, a modified parallel tempering (PT) algorithm applied to train the restricted Boltzmann machine (RBM) is proposed. Firstly, RBM is trained in light of Metropolis-Hasting for parallel tempering sampling, then stacking up RBMs to form a deep belief network (DBN) as the basis for deep neural network (DNN) pre-training, then by adding an output layer called "softmax" to the network, a DNN detecting the posterior probability of phoneme can be created. Subsequently, backward propagation algorithm is applied to fine-tune the weights discriminatively with less label data. Finally, the sequence of the predicted probability distribution is fed into a standard Viterbi decoder. The experiments show that the proposed method has a better performance on the TIMIT dataset than traditional ways. Its recognition rate is higher 4.5% than contrastive divergence (CD), and 1% than original PT without more computation.

Key words Parallel tempering, Restricted Boltzmann machine, Deep belief network, Phoneme recognition

2018-04-25 收稿; 2018-08-02 定稿

*国家自然科学基金项目 (61571106)

作者简介: 阴法明 (1980-), 男, 江苏南京人, 硕士研究生, 研究方向: 电子与信息处理。

†通讯作者 E-mail: yinfm@njcit.cn

0 引言

音素识别指的是对给定的语音特征向量，估计语音标签序列的过程，在诸多语音识别系统中具有广泛的应用[1−2]，如关键字识别、语言分类、说话人识别等。有效的音素识别是提高语音识别的关键。

目前语音识别系统常用隐马尔科夫模型(Hidden Markov models, HMM)来处理语音中的时域变量，用高斯混合模型(Gaussian mixture models, GMM)来确定每一个HMM状态是如何对应于一帧输入语音参数[3]。但是这种方法还存在一些缺点：在模拟数据空间中非线性样本时，其统计无效。例如对球面上的点集进行建模时，GMM就需要使用大量的对角高斯或协方差高斯[4]。此外这种方法的语音是通过调制动态系统中相对较少的参数产生的，这意味着它真实的底层结构是用了一组低维数据来表示一帧包含了上百参数的语音。所以如果能充分挖掘帧中的信息，就有可能找到一种比GMM更好的方法来进行语音建模。

为克服上述缺点，有学者提出将深度神经网络应用于声学建模中，用深信度网络(Deep belief network, DBN)/隐马尔科夫模型(DBN/HMM)结构来提高最终的识别率[5−6]。Google 与YouTube的相关实验也表明DBN/HMM在语音识别效果上要远远优于传统的GMM/HMM[4]。而DBN是通过将多个受限玻尔兹曼机(Restricted Boltzmann machine, RBM)堆叠而成，所以RBM的训练成为整个结构的关键。Hinton[7]在2010 年提出了对比散度(Contrastive divergence, CD)用来训练RBM，之后又出现了持续对比散度(Persistent contrastive divergence, PCD)[8]。但是这两种方法都是对单条马尔可夫链进行采样，且在初始化数据上也较为粗糙，导致其在计算模型期望时存在较大误差。

为此本文在并行回火(Parallel tempering, PT)算法的基础上，根据来自多条吉布斯链样本的状态能量，进行等能量划分，构建多个能量环，提高相邻温度链之间的交换率，进而优化RBM的训练，并将训练好的RBM堆叠成DBN进行音素识别。在TIMIT语料库上，由改进的并行回火算法所获得的识别率明显高于对比散度类算法。

1 受限玻尔兹曼机

受限玻尔兹曼机(RBM)是一种特殊的马尔科夫随机域，一个RBM包含一个由随机的隐层单元构成的隐层和一个由随机的可见单元构成的显层，其中隐层一般为伯努利分布，显层一般是高斯分布或伯努利分布[9]。RBM可以表示成双向图，只有不同层之间的单元才会存在边，同层单元之间都不会有边连接，即层间全连接，层内无连接。

RBM是一种基于能量的模型，其可见矢量$\boldsymbol{v}$和隐层矢量$\boldsymbol{h}$的联合配置能量由公式(1)给出。

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i\in\text{visible}} a_i v_i - \sum_{j\in\text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}, \tag{1}$$

其中，v_i是可见单元的二值状态，h_j是隐层单元的二值状态，a_i和b_j分别是可见单元i和隐层单元j的偏置值，w_{ij}是链接权值。通过E可以定义可见单元和隐层单元状态的联合分布概率：

$$p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{w}) = \frac{1}{Z}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{w})}, \tag{2}$$

其中Z是配分函数或归一化项，$Z=\sum_{\boldsymbol{v}',\boldsymbol{h}'}\mathrm{e}^{-E(\boldsymbol{v}',\boldsymbol{h}';\boldsymbol{w})}$。模型中可见矢量$\boldsymbol{v}$的概率计算公式如下：

$$p(\boldsymbol{v}) = \frac{1}{Z}\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{w})}. \tag{3}$$

因为RBM层内无连接，所以隐层单元之间是独立的，所以可见矢量$\boldsymbol{v}$的概率是对隐层单元的求和。RBM中的权值更新算法依据梯度下降法[7]：

$$\frac{1}{N}\sum_{n=1}^{n=N}\frac{\partial \lg p(\boldsymbol{v}^n)}{\partial w_{ij}} = \langle v_i h_j\rangle_{\text{data}} - \langle v_i h_j\rangle_{\text{model}}, \tag{4}$$

式(4)表示由输入数据所确定的期望$\langle v_i h_j\rangle_{\text{data}}$与模型获取的期望$\langle v_i h_j\rangle_{\text{model}}$之间的差异。最终，可以得到RBM 的权值每次更新的大小为

$$\Delta w_{ij} = \epsilon\left(\langle v_i h_j\rangle_{\text{data}} - \langle v_i h_j\rangle_{\text{model}}\right). \tag{5}$$

2 改进的RBM的训练算法

对于RBM而言，由于隐层单元之间没有连接，无偏样本$\langle v_i h_j\rangle_{\text{data}}$是很容易得到的，而且条件分布

$p(\boldsymbol{h}|\boldsymbol{v})=\prod_j p(h_j|\boldsymbol{v})$，给定一个可见矢量$\boldsymbol{v}$，隐层单元$h_j$的状态为1的概率为

$$p\left(h_j=1|\boldsymbol{v}\right)=\text{logistic}\left(b_j+\sum_i v_i w_{ij}\right). \quad (6)$$

同理可得给定一个隐层矢量$\boldsymbol{h}$，可见单元v_i的状态为1的概率为

$$p\left(v_i=1|\boldsymbol{h}\right)=\text{logistic}\left(a_i+\sum_j h_j w_{ij}\right). \quad (7)$$

无偏样本$\langle v_i h_j\rangle_{\text{model}}$的获得是很困难的。传统算法采用对比散度来近似计算该模型的期望，步骤总结如下：(1) 初始化可见矢量$\boldsymbol{v}_0$；(2) 采样$h_0:p(\boldsymbol{h}|\boldsymbol{v}_0)$；(3) 采样$v_1:p(\boldsymbol{v}|\boldsymbol{h}_0)$；(4) 采样$h_1:p(\boldsymbol{h}|\boldsymbol{v}_1)$；如此交替进行采样来训练RBM。由此可知，该算法的复杂度是指数级增加的。

为解决RBM的训练效率问题，目前提出了对比散度(CD)、持续对比散度(PCD)和并行回火(PT)等方法[10]。对比散度是训练RBM的标准方法，它通过训练数据来初始化吉布斯链，然后交替执行CD-1算法，所以实际上它并没有依据模型分布来计算对数概率的梯度[7]。持续对比散度是通过对一条持续马尔科夫链进行吉布斯采样来计算模型梯度，其初始吉布斯的状态来源于前一次的更新参数，而不是训练数据[8]。这两种方法都仅使用单一的马尔科夫链来计算$\langle v_i h_j\rangle_{\text{model}}$，这会引起训练退化。尤其是对含有多个峰值的目标分布，这种使用对比散度或持续对比散度的吉布斯采样会容易陷入局部最优。

“回火”作为一种通用策略，它通过从$1/t<1$的模型中采样来实现不同峰值之间的快速混合。本文使用并行回火采样对RBM训练(RBM-PT)，并行回火引入了增补吉布斯链，它能够从渐进平滑的原始分布中采样[11−12]。RBM-PT在训练过程中，每个温度对应一条吉布斯链并使用并行回火的方法采样。每条吉布斯链对应一个不同的温度t_i，t_i满足$1=t_1<\cdots<t_i<\cdots t_{M-1}<t_M$，不同温度链之间根据一定的条件决定是否交换采样值。

根据式(2)，在不同的温度下，并行回火RBM联合概率为

$$p(\boldsymbol{v},\boldsymbol{h})=\frac{1}{Z(t_i)}\exp\left(-\frac{1}{t_i}E\left(\boldsymbol{v},\boldsymbol{h};\theta\right)\right),\quad i=1,2,\cdots,M. \quad (8)$$

通过将式(1)的RBM参数$\theta_{\text{RBM}}=\{W,a,b\}$中的显层单元与隐层单元之间的连接权重$W$乘以温度$\beta$，整个模型的参数变为$\theta_{\text{RBM-PT}}=\{\beta W,a,b\}$，对于偏置值$a$和$b$并没有改变。此时，并行回火算法可与受限波尔兹曼机结合，改善训练效率。公式(8)中的参数t指“温度”，该参数反映了基于能量模型的统计物理起源。当温度趋于0时，$1/t$则趋于无穷，此时的基于能量的模型是确定性的。反之，基于能量的模型成了均匀分布。

并行回火蒙特卡罗算法包括两个阶段：

(1) Metropolis-Hastings采样[13]阶段：根据已有的采样值计算当前温度的下一个采样点，基本采样计算公式为

$$x^{i+1}|x^i=\text{Metropolis}-\text{Hastings}\left(x^i+N\left(0,\frac{\sigma_i^2}{t_k}\right)\right), \quad (9)$$

其中，$N\left(0,\frac{\sigma_i^2}{t_k}\right)$是均值为0、方差为$\frac{\sigma_i^2}{t_k}$的正态分布，$t_k$表示温度，$x^i$表示第$i$条链的显层与隐层状态。

(2) 交换：并行回火RBM模型的交换条件如下：

$$\min\left\{1,\exp\left(\left(\frac{1}{t_\gamma}-\frac{1}{t_{\gamma-1}}\right)*\left(E(v_\gamma,h_\gamma)-E(v_{\gamma-1},h_{\gamma-1})\right)\right)\right\}, \quad (10)$$

其中，t_γ与$t_{\gamma-1}$是两个相邻的温度，$E(v_\gamma,h_\gamma)$与$E(v_{\gamma-1},h_{\gamma-1})$是其对应的隐层期望。如果满足该条件，就把相邻的温度链下的采样点交换，否则不交换。为了提高这种交换率，本文提出了如下改进方法：由公式(10)可得，当温度固定时，交换率取决于两个状态能量之差，且差值越小交换的可能就越大。本文根据所有链的状态能量，将状态空间分为几个等能量集合，促使当前状态向等能量集中的其他状态转移。具体算法如下：

首先引入$d+1$个能量水平：

$$H_1<H_2<\cdots<H_{d+1}=\infty, \quad (11)$$

理论上H_1应小于最小能量，但在本文中H_1被设为最小能量，而H_d等于最大能量值。因为这样也能包含模型中的所有状态能量。$H_2,\cdots,H_{d-1}$通过均分(H_d-H_1)获得。

其次根据这$d+1$个能量水平，要将N个马尔可

夫链划分为多个能量环，每个能量环 D_j 定义如下：

$$D_j = \{(\boldsymbol{v},\boldsymbol{h}) : E(\boldsymbol{v},\boldsymbol{h}) \in [H_j, H_{j+1}]\},$$
$$j = 1, \cdots, d. \tag{12}$$

接着在能量环内执行交换，而是否交换的依据类似于公式(10)，不同的是此处的能量差应为同一能量环内的两条链的能量差。实际中交换的次序是从高温向低温执行的。此外由于在训练时RBM的参数是动态改变的，所以这些状态能量也是动态的，实际操作中我们只要在训练RBM前设定好能量环的数量 d 即可。

最后经过多次循环采样、交换，最终将 $t_1 = 1$ 温度下的采样值用于RBM预训练模型参数 θ，并采用并行回火获取的目标采样值可使RBM训练获得较好的应用效果。

3 基于RBM的深信度网络

在训练好一个RBM后，其隐层单元状态可以作为训练下一个RBM的数据，所以该RBM能够学习到第一个RBM隐层单元之间的依赖性。这一过程可以一直重复下去，直到产生所需要的非线性特征检测器的层数，层数越多统计数据结构也就越复杂。将多个RBM堆叠起来就能产生一个多层生成模型——深信度网络(DBN)。虽然单个RBM是间接模型，但由它产生的DBN是一个混合生成模型。DBN的最上面2层是无向链接，其他层是自顶向下的有向链接。获得DBN之后，在其顶层之上，再增加一个softmax输出层，输出每种音素对应的概率值。此时的网络称为DBN-DNN，如图1所示。

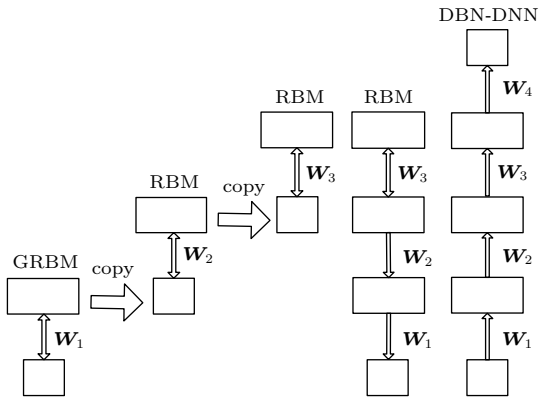

图1 利用RBM堆叠产生用于音素识别的DBN

Fig. 1 Stacking up RBMs to form DBN for phoneme recognition

RBM的预训练仅仅为了使得DBN获得一个较好的初始权重，避免训练时陷入局部最优[14]。为了使得DBN能更好地应用于音素识别，还需要针对目标输出进行监督训练。其输出目标为语音内的中间帧所对应的HMM状态。训练的损失函数为交叉熵，通过方向传播算法获得网络的最终权值。

4 实验结果分析

4.1 实验配置

本文实验在TIMIT语料库上进行，选择462个说话人的3296个语句为训练集，选择TIMIT的核心测试集(24个说话人的192个语句)作为测试集。语音信号使用Hamming窗处理，帧长25 ms，帧移10 ms，预加重系数为0.97。声学特征参数使用13阶梅尔频率倒谱系数(Mel-frequency cepstrum coefficients, MFCC)，以及其一阶、二阶差分系数，最终使得每帧语音含有39维特征参数。RBM的训练使用8条吉布斯链。预训练时的学习率为0.001。监督学习中的学习率为0.0001，以Adam为优化器。

4.2 参数性能分析实验

图2给出了隐层单元数为1024时，隐层数与帧数对识别结果的影响。从图2中可以看出，随着隐层数量和输入帧数的增加，识别性能有明显改善。其中隐层数量的增加提高了网络对非线性函数的拟合能力，而帧数的增加则代表了输入上下文信息量的增加。当DNN的隐层数为4、输入帧数为15时，取得了最佳识别性能。说明隐层数量的增加并不会

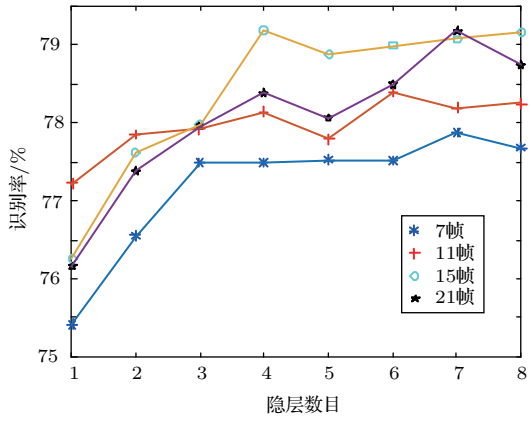

图2 输入帧数变化时的音素识别性能

Fig. 2 The phoneme recognition performance when the input frames numbers change

无限度地提高识别率，因为随着层数的增加，会导致梯度消失等问题[15]。同样输入信息的增加也不会无限度地改善系统性能，一方面是因为时间跨度较大的两帧语音数据之间的相关性较小，甚至有可能从一个音素所在时间蔓延到另一个音素时间，导致识别率下降；另一方面是当网络参数确定后，DNN对于这些特征的区分能力是有限的。如图 2 中 15 帧语音与 21 帧语音所对应的识别率曲线图所示。

图 3 给出了输入帧数固定为 11 帧，隐层单元数对识别结果的影响。从图 3 中可以看出，当隐层数固定时，增加隐层单元数可以提高音素识别性能。当隐层单元数较少时，通过增加隐层数量能有效提高识别性能，但当隐层数过多时，这种改善效果就显得非常有限。这表明隐层单元数在一定程度上决定了网络最终的识别率。实际中，过多的隐层单元数和隐层数会带来庞大的时间开销，而带来的性能改善却是有限的，所以往往需要折中考虑参数配置。

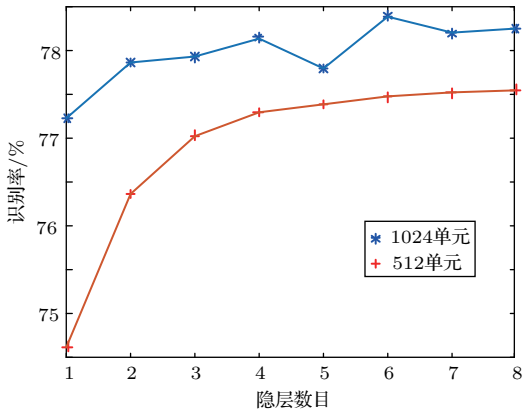

图 3　隐层单元数不同时的音素识别性能

Fig. 3 Phoneme recognition performance with different number of hidden layer nodes

4.3　不同训练算法的对比实验

上文中简述了各种不同 RBM 的训练方法及各自的特点，本实验给出在隐层单元数为 1024、输入帧数为 11 帧时，不同训练算法的识别率对比结果。从图 4 中可以看出，并行回火类算法的识别性能明显优于对比散度类算法。主要原因在于对比散度与持续对比散度仅使用一条马尔可夫链进行梯度估算，而并行回火类算法则依据从原始分布中采样出的多条吉布斯链对公式 (4) 进行计算，其精确度更高。而本文所提的方法的识别率比对比散度算法提高约 4.5%，比原始的并行回火算法识别率高 1% 左右，因为通过等能量划分后，相邻温度下的状态交换率提高了，进而提高了最终的识别率。由此说明在没有增加计算量的情况下，本文对并行回火算法的改进在音素识别应用上是有效的。

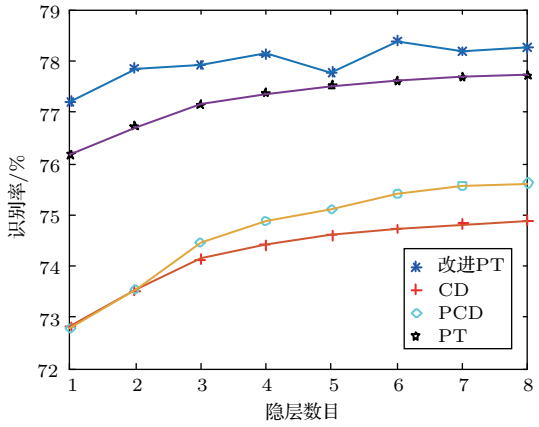

图 4　不同训练算法的音素识别性能

Fig. 4 Phoneme recognition performance of different training algorithms

5　结论

本文首先研究分析了 RBM 的学习原理，在并行回火算法的基础之上，根据模型分布所得的样本能量，进行等能量划分，以提高相邻温度链之间的交换率，进而提高模型期望的计算精度，训练出较好的 RBM。然后将 RBM 组成 DBN 应用于音素识别中，实验表明，由该方法训练所得的 RBM 可以有效提高最终识别率。

参　考　文　献

[1] Saeb A, Razzazi F, Babaei-Zadeh M. A fast phoneme recognition system based on sparse representation of test utterances[C]. Hands-Free Speech Communication and Microphone Arrays, 2014.

[2] Lohrenz T, Fingscheidt T. Turbo fusion of magnitude and phase information for DNN-based phoneme recognition[C]. Automatic Speech Recognition and Understanding Workshop, 2018.

[3] Khelifa M O M, Belkasmi M, Abdellah Y, et al. An accurate HSMM-based system for Arabic phonemes recognition[C]. Ninth International Conference on Advanced Computational Intelligence, 2017.

[4] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82–97.

[5] SreeB R L, Vijaya M S. Building acoustic model for phoneme recognition using PSO-DBN[J]. International Journal of Business Intelligence & Data Mining, 2018, 1(1): 1.

[6] Mohamed A R, Sainath T N, Dahl G, et al. Deep belief networks using discriminative features for phone recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

[7] Hinton G E. A practical guide to training restricted Boltzmann machines[J]. Momentum, 2012, 9(1): 599–619.

[8] Jang H, Choi H, Yi Y, et al. Adiabatic persistent contrastive divergence learning[C]. IEEE International Symposium on Information Theory, 2017.

[9] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527–1554.

[10] Berglund M, Raiko T. Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence[J]. Computer Science, 2014, arXiv: 1312.6002.

[11] Cho K, Raiko T, Llin A. Parallel tempering is efficient for learning restricted Boltzmann machines[C]. Neural Networks (IJCNN), The 2010 International Joint Conference on, IEEE, 2010.

[12] Desjardins G, Courville A, Bengio Y. Adaptive parallel tempering for stochastic maximum likelihood learning of RBMs[J]. Computer Science, 2010, arXiv: 1012.3476.

[13] Koller D, Friedman N. Probabilistic graphical models: principles and techniques-adaptive computation and machine learning[M]. Massachusetts: The MIT Press, 2009: 161–168.

[14] He F, Han Y, Wang H, et al. Deep learning architecture for iris recognition based on optimal Gabor filters and deep belief network[J]. Journal of Electronic Imaging, 2017, 26(2): 023005.

[15] Deng J, Xu X, Zhang Z, et al. Semi-supervised autoencoders for speech emotion recognition[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, 26(1): 31–43.