基于递归神经网络的语音识别快速解码算法

张 舸[①②] 张鹏远*[①②] 潘接林[①] 颜永红[①②③]
[①](中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)
[②](中国科学院大学 北京 100190)
[③](中国科学院新疆理化技术研究所新疆民族语音语言信息处理实验室 乌鲁木齐 830011)

摘 要：递归神经网络(Recurrent Neural Network, RNN)如今已经广泛用于自动语音识别(Automatic Speech Recognition, ASR)的声学建模。虽然其较传统的声学建模方法有很大优势，但相对较高的计算复杂度限制了这种神经网络的应用，特别是在实时应用场景中。由于递归神经网络采用的输入特征通常有较长的上下文，因此利用重叠信息来同时降低声学后验和令牌传递的时间复杂度成为可能。该文介绍了一种新的解码器结构，通过有规律抛弃存在重叠的帧来获得解码过程中的计算开销降低。特别地，这种方法可以直接用于原始的递归神经网络模型，只需对隐马尔可夫模型(Hidden Markov Model, HMM)结构做小的变动，这使得这种方法具有很高的灵活性。该文以时延神经网络为例验证了所提出的方法，证明该方法能够在精度损失相对较小的情况下取得 2~4 倍的加速比。

关键词：语音识别；递归神经网络；解码器；跳帧计算
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2017)04-0930-08
DOI: 10.11999/JEIT160543

Fast Decoding Algorithm for Automatic Speech Recognition Based on Recurrent Neural Networks

ZHANG Ge[①②] ZHANG Pengyuan[①②] PAN Jielin[①] YAN Yonghong[①②③]
[①](*The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)
[②](*University of Chinese Academy of Sciences, Beijing* 100190, *China*)
[③](*Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi* 830011, *China*)

Abstract: Recurrent Neural Networks (RNN) are widely used for acoustic modeling in Automatic Speech Recognition (ASR). Although RNNs show many advantages over traditional acoustic modeling methods, the inherent higher computational cost limits its usage, especially in real-time applications. Noticing that the features used by RNNs usually have relatively long acoustic contexts, it is possible to lower the computational complexity of both posterior calculation and token passing process with overlapped information. This paper introduces a novel decoder structure that drops the overlapped acoustic frames regularly, which leads to a significant computational cost reduction in the decoding process. Especially, the new approach can directly use the original RNNs with minor modifications on the HMM topology, which makes it flexible. In experiments on conversation telephone speech datasets, this approach achieves 2 to 4 times speedup with little relative accuracy reduction.

Key words: Speech recognition; Recurrent Neural Network (RNN); Decoder; Frame skipping

1 引言

近年来，不同种类的递归神经网络(Recurrent Neural Network, RNN)，例如长短时记忆神经网络(Long Short Term Memory, LSTM)[1−4]和时延神经网络(Time Delay Neural Network, TDNN)[5−7]开始被用于自动语音识别的声学模型建模中，并取得了相对前馈神经网络更好的性能。产生这种优势的主要原因是递归神经网络的记忆能力能够涵盖整个语音序列的历史，而前馈神经网络则仅利用一个有限长度的窗内的上下文信息。

时延神经网络是一种结构较为简单的递归神经网络。这种网络在保持与前馈神经网络相近的计算方法的同时，在隐层利用了更长的上下文，从而使

收稿日期：2016-05-26；改回日期：2017-01-09；网络出版：2017-02-24
*通信作者：张鹏远 zhangpengyuan@hccl.ioa.ac.cn
基金项目：国家自然科学基金(U1536117, 11590770-4)，国家重点研发计划重点专项(2016YFB0801200, 2016YFB0801203)，新疆维吾尔自治区科技重大专项(2016A03007-1)
Foundation Items: The National Natural Science Foundation of China (U1536117, 11590770-4), The National Key Research and Development Plan of China (2016YFB0801200, 2016YFB0801203), The Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (2016A03007-1)

网络获得利用更多上下文信息的能力。具体地说，前馈神经网络的隐层输入是当前帧上一层的输出，而时延神经网络的隐层输入是与当前帧相对应的一个帧序列在上一层的输出的顺序拼接。时延神经网络可以用更小的网络参数量获得超过前馈神经网络的性能，从而同时提高语音识别的精度和效率[5]。

然而，递归神经网络的连接比前馈神经网络更复杂，因而带来了更大的计算量，导致解码速度更慢。这限制了递归神经网络在对实时性要求较高的任务中的应用。解决这一问题的直接方法是减小神经网络的尺寸，例如减少网络层数或各层的结点数。然而采用更小的神经网络进行解码将显然影响到系统的准确性，所以将网络保持在一个合理的尺寸是必要的[8, 9]。另一类方法是修改神经网络的结构，例如，时延神经网络的隐层采用不连续的上下文来减小计算量[5]。这种方法能够将计算复杂度降低到接近普通前馈神经网络。这些加速方法都通过对神经网络的修改来达到减小计算量的目的，因此不能使同一个网络满足不同的应用场景对精度和速度的要求。为了达到加速计算的目的，本文提出了一种跳帧计算的方法，从而直接减少神经网络需要计算的帧数，并保持神经网络本身的结构不变。这种方法能够在加速计算的同时保持递归神经网络在精度上的优势，同时由于不需要改变神经网络的结构，因而具有更好的灵活性。

由于在语音识别中，语音信号的帧移通常足够短，用当前帧的相邻帧的声学后验概率来预测当前帧的声学后验概率是合理的[10]。实验证明，按照一个较小的比例对语音序列的帧进行顺序采样，就足以表达整个语音序列的声学信息。进一步，我们研究了在解码的令牌传递过程中同样省略一些帧的可能性，并证明这种方法可以在保持识别性能基本不变的情况下将系统整体速度提升若干倍。在采用这种方法时，虽然语音序列的声学信息得到足够的保存，但省略一些帧会导致声学模型得分的动态范围发生变化，从而影响到系统的其他配置参数。此外，实验中我们发现，声学模型单元对应的隐马尔可夫模型(Hidden Markov Model, HMM)结构也需要做出调整，以应对跳帧对令牌传递路径的要求。

本文在以时延神经网络为例的递归神经网络上进行了实验，证明跳帧方法可以简单地应用于递归神经网络，并根据任务需要调节系统精度和效率之间的平衡。本文的其它部分是这样安排的：第 2 节首先介绍了所采用的解码系统基本结构；第 3 节描述了跳帧方法的基础版本帧异步方法；第 4 节重点叙述对跳帧方法的分析和得分修正，以及对 HMM 结构的调整；第 5 节提出了在跳帧系统的基础上进行可变帧率的令牌传递的方法；最后给出了实验结果及总结。

2　基线系统结构

语音识别的基本方程可以表示为

$$\widehat{W} = \arg\max_W \left\{P(X \mid W)P(W)\right\} \tag{1}$$

其中 $P(W)$ 表示语言模型分数，$P(X \mid W)$ 表示声学模型分数[11]。对于本文中采用的解码器结构，声学模型分数由深度神经网络给出，而语言模型分数由加权有限状态转换器 (Weighted Finite-State Transducer, WFST)提供。

本文采用的解码器解码过程如下：输入的语音信号被转换为由若干帧组成的特征序列，然后这些特征被输入到时延神经网络中，获得每一帧对应的声学模型后验概率。根据当前帧的声学模型后验概率和语言模型后验概率，解码器在动态加载 HMM 的 WFST 上进行基于令牌传递算法的维特比剪枝搜索[12, 13]，得到的最优路径表示了输入语音对应的识别结果。我们使用的 WFST 输入是上下文相关的三音子，输出是识别词，神经网络的输出与 HMM 状态一一对应，HMM 状态到相应三音子的映射在解码过程中动态进行[11]。为了在解码过程中控制搜索空间的规模，以调节解码速度和精度的平衡，我们采用多种剪枝策略以去除可能性较低的搜索路径[13]。

3　帧异步方法

在一般的基于递归神经网络的语音识别系统中，神经网络的输入通常具有一个较长的上下文，用于计算输入窗口中间帧的声学模型后验概率。因此，在神经网络的计算中，相邻帧的计算窗口有很长的重叠。从而，有可能通过当前帧 t 的前若干帧 $t-K, K=1,2,\cdots$ 来预测当前帧的声学模型后验概率。最直接的概率预测方式是对相邻的两帧，其奇数帧的后验概率采用前一个偶数帧的后验概率，公式为[10]：

$$p(2t+1) = p(2t), \quad t = 0,1,\cdots \tag{2}$$

考虑到语音信号是短时平稳的，相邻帧的声学模型后验概率应该相近，因此这种预测方式是合理的。进一步，由于输入窗口足够长，可以认为从每 3 帧或每 4 帧中选取一帧，以这一帧的概率为这些帧的概率，也是合理的。实验证明这种方法可以获得和一般的系统相近的精度。下文中我们将从连续的 n 帧中选取第 1 帧用于计算声学模型后验概率，并复制作为其余 $n-1$ 帧概率的方法称作 n-帧异步方

法，相应的系统称作 n-帧异步系统。

显然，帧异步方法可以与选取帧的比例成正比地减小声学模型后验概率的计算量。然而，这种方法将保持令牌传递过程的计算量不变。实验证明在计算声学模型后验概率时，我们的系统最多可以忽略总帧数的 3/4 而不造成显著的精度损失。因此，我们期望一个设计良好的跳帧系统可以达到比基线系统快接近 4 倍的加速比。为此，解码过程中的令牌传递过程也需要做相应的改动。

3.1 时延神经网络前向计算中的近似方法

在前馈神经网络中，当前帧的概率计算仅依赖于当前帧的输入，因此在前馈神经网络上应用帧异步方法时，每 n 帧中只有 1 帧参与计算，计算量将减小到逐帧方法的 $1/n$。

注意到递归神经网络在计算当前帧的概率时，需要依赖前后若干帧的隐层计算结果。在时延神经网络中，隐层的输入是上一隐层输出的一个帧序列。如果我们从 n 帧中选取 1 帧，则被选取的帧对应的帧序列可能没有被计算。例如，当隐层的输入为

$$I_l(t)=\{O_{l-1}(t-3),\ O_{l-1}(t+3)\} \tag{3}$$

当我们采用每 2 帧取 1 帧的方法时，如果采用和前馈神经网络一致的计算方法，上一层对应 t-3 和 t+3 的激活值将不被计算，因此这一层的输入也无法得到。一个直观的办法是计算这些帧对应的部分隐层输出，然而这会减少帧异步方法的加速程度，并且计算的不连续性会对系统实现带来困难。因此，我们选择在隐层的输入序列中按照与输入相同的规则从 n 帧中选取第 1 帧，并作为其余 n −1 帧的隐层输出。此时前述的公式变为

$$I_l(t)=\{O_{l-1}(t-4),\ O_{l-1}(t+2)\} \tag{4}$$

这样，时延神经网络的计算量减小到 $1/n$。实验证明，这种近似方法会对系统引入一些误差，但尚不会造成显著的精度损失。

3.2 选取计算帧的方法

本文所述的 n-帧异步方法是从连续的 n 帧中选取第 1 帧用于计算声学模型后验概率，并复制作为其余 n−1 帧概率。我们同时考虑了根据帧的特性选取关键帧的方法。然而，时延神经网络的层间依赖关系影响了这种方法的可行性。在 n 较小的情况下，选取关键帧和选取第 1 帧的区别并不明显，而 n 较大的情况下，由于关键帧之间的距离并不确定，时延神经网络的计算近似误差将变得更大且不稳定，因此无法预期得到可靠的结果。本文的实验证明，选取第 1 帧作为计算帧的方法已经能够得到接近逐帧系统的精度，因此，本文将不在时延神经网络上采用更复杂的选取关键帧的方法。

4 跳帧方法

当使用 WFST-RNN 系统进行基于令牌传递算法的语音序列解码时，每个 HMM 状态上的对应令牌在每一帧会沿着 HMM 结构所允许的路径向前传递一次[11]。下文中我们将从连续的 n 帧中选取第 1 帧用于计算声学模型后验概率，并忽略其余 n−1 帧的方法称作 n-帧跳帧方法，相应的系统称作 n-帧跳帧系统。在一个 n-帧跳帧系统中，每 n 帧中有 n−1 帧被忽略，因此需要重新考察令牌传递的路径和步长。为了保持系统的行为完全一致，最直接的方法是在每一帧都把令牌向前传递 n 次，这样将得到和帧异步系统完全一致的结果。然而这样的系统具有和帧异步系统完全一致的令牌转递次数，从而具有一致的时间复杂度，因此无法得到加速计算的效果。为了能够减少令牌传递步骤的时间开销，需要一种能够仍然保持每帧进行一次令牌传递的方法。

4.1 搜索空间的等价性

我们首先研究 2-帧跳帧系统。首先，我们尝试直接删去奇数帧，并不对令牌传递的方法做改动。这时，WFST 解码器的算法保持不变，而解码配置需要做出调整。当有一半帧在解码过程中被跳过时，声学模型后验概率的累积分数也会减半，这意味着剪枝宽度需要同样减半，从而可以获得等价的搜索空间。此外，为了获得正确的解码结果，解码过程中语言模型得分的权重也需要进行调整。

在主流的 WFST 解码器结构中，声学模型得分和语言模型得分在对数域上相加以获得总的分数，从而得到最优路径[11]。因此，声学模型得分和语言模型得分的动态范围应该相匹配，从而分数累加才有意义[14]。注意到声学模型得分在每一帧累加，而语言模型得分仅在词尾累加，这就造成了两者的动态范围差异较大。为了解决这个问题，WFST 解码器引入一个语言模型权重因子来平衡声学模型得分和语言模型得分，这个因子被乘到语言模型得分上，以使之与声学模型得分相匹配[14]。在我们的实践中，语言模型权重因子取 9 时可以得到最佳性能。显然，跳帧系统减小了声学模型累积得分的动态范围，并且由于跳过了一半的帧，每个词对应的平均帧数也被减半了。因此，为了保持语言模型得分和声学模型得分的平衡，在 2-帧跳帧系统中，我们取语言模型权重因子为 4.5。在跳过更多帧的系统中，语言模型权重因子还需要相应按比例减小。

4.2 跨状态 HMM

自从隐马尔可夫模型用于语音识别之初，多种

表达声学模型的 HMM 结构都得到了讨论[15]。其中，最常用的结构是由左向右的逐状态 HMM[16]。如图 1 所示，在这种最简单的结构中，每个上下文相关音素被表示为 3 个 HMM 状态，解码过程中的令牌在一条从左向右经过每一个状态的路径上传递。根据当前帧声学模型后验概率，每个状态上的令牌可以停留在当前状态或前进到下一个相邻状态。因此，当帧率为每帧 10 ms 时，每个音素的持续时间不能短于 30 ms。对 2-帧跳帧系统，由于有一半的帧被删去，音素的最短持续时间将延长到 60 ms。如果不同音素的持续时间相近，这个持续时间可以支持每秒 16 个音素的语速。然而，音素持续时间往往变化范围很大，特别是辅音音素通常短得多。在我们的实验中，对于逐帧系统，有多达 72%的音素存在至少 1 个仅占 1 帧的状态。如果对应的这一帧被跳过，这个状态将被其相邻的状态所表示，这样将迫使音素的相邻音素让出至少 1 帧用于填补当前音素的空缺，从而导致解码过程中对状态的预测不准确。进一步地，我们发现 1 个音素的所有 3 个状态都仅在逐帧系统中占 1 帧的情况十分罕见，这导致音素在 2-跳帧系统中往往只存在 1 个缺失了对应帧的状态。如果这个状态在令牌传递的路径上可以被跳过，那么 HMM 结构仍然能够在跳帧系统中正确描述这个音素。这样的 HMM 结构如图 2 所示。

在 HMM 结构中添加跨状态转移路径会略微增加令牌传递的计算开销。在 n 个状态的逐状态 HMM 中，状态转移矩阵的转移概率分布满足：

$$\left.\begin{array}{l} a_{ij} \neq 0, \quad i=j \text{或} i=j-1 \\ a_{ij}=0, \text{ 其他} \end{array}\right\} \tag{5}$$

从而对每个音素有 $2n$+1 条转移路径，对 3 状态 HMM 而言，即 7 条转移路径。加入跨状态转移之后，每个音素有 $3n$+1 条转移路径，对 3 状态 HMM 而言，即 10 条转移路径。转移路径数的增加会增加解码过程中的时间复杂度，但由于增加的比例有限，跳帧解码系统仍然能够获得相对于帧异步系统的速度优势。

在跳过更多帧的 n-跳帧系统中，显然 HMM 结构需要允许更长的跨状态转移。然而为了获取音素的有效信息，我们应该保证每个音素至少有 1 个观测值，即对 n 状态的 HMM，最多只能允许跨过 n −1 个状态的转移。这样，3 状态的 HMM 有 12 条转移路径。这样的 HMM 结构如图 3 所示。

为了保持系统的灵活性，即保证同一个模型可以用于帧同步、帧异步和跳帧系统，我们选择不修改存储 HMM 的声学模型文件，这样在更改系统配置时就不需要重新训练声学模型。作为替代，我们在解码过程中为 HMM 添加必要的跨状态转移。注意到常用的逐状态 HMM 仅包含状态到自身的转移和状态到下一个相邻状态的转移，为了添加跨状态转移，我们需要对应的跨状态转移概率。仍然是为了避免重新训练声学模型，我们选择将跨状态转移概率设置为常数。实验证明，状态间转移概率的数值对语音识别的精度并不造成显著的影响，因此这样设置是合理的。这种做法充分保持了声学模型的灵活性，使语音识别系统可以对精度和效率进行按需配置，而不需要更换声学模型。

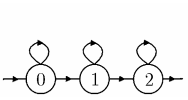

图 1 逐状态的 HMM

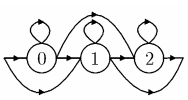

图 2 跨单个状态的 HMM

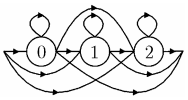

图 3 跨 n−1 个状态的 HMM

4.3 跳帧系统的等价帧移

注意到在跳帧方法中，我们按照给定的间隔从连续的 n 帧中选取 1 帧。这一操作等价于在进行特征提取时采用 n 倍的帧移，即单位时间内的帧数变为 $1/n$。显然，前文描述的 n-帧跳帧系统在性能上应等价于帧移 n 倍的逐帧系统。

然而，改变帧移意味着新的特征将与已有的声学模型不匹配，因而需要重新训练声学模型。这样，对每一个不同的帧移设定，就需要训练一个独立的声学模型。这将使调节识别精度和效率的平衡变得更加耗时。与之相比，跳帧系统可以不更换声学模型就对精度和效率进行配置，从而增加了系统应用中的灵活性。

5　可变帧率方法

前述的跳帧方法在选择参与令牌传递的帧的时候，仅按照时间序列的给定间隔进行选取。本节提出可变帧率方法，以进一步利用相邻帧的信息冗余，减少令牌传递的时间。

如前文所述，跳帧的合理性来自语音信号中相邻帧的相似性，因此，根据相邻帧的相似性进行令牌传递的帧的选取，可以得到更好的效果。另一方面，基于神经网络的泛化能力，对神经网络的输出进行相似性衡量，比对音频信息进行相似性衡量，具有更好的鲁棒性。我们根据相邻帧的神经网络输出的相似性进行帧选取。具体方法是计算当前帧 t 与上一个参与令牌传递的帧 k 的神经网络输出向量之间的距离 $d(t,k)$，当这一距离小于设定的门限时，认为两帧足够相似，从而不加入令牌传递。通过这一方法，可以得到用于令牌传递阶段的帧的最小集

合，从而加速令牌传递过程。这一方法可以应用于帧同步系统，也可以应用于跳帧系统，当应用于跳帧系统时，参与相似性计算的神经网络输出是跳帧之后的相邻输出。

如前所述，更少的帧加入令牌传递意味着需要更长的跨状态转移，因此采用与跳帧方法相同的对 HMM 的修改。在跳帧方法的基础上，可变帧率方法能够进一步减少令牌传递环节的时间开销，同时也不带来显著的精度损失。

6 实验与分析

我们在一个 1.5 h 中文电话对话语音测试集上测试了上述方法。这个测试集是 HKUST 普通话电话语音数据集(LDC2005S15/LDC2005T32)的一部分。实验中使用的语言模型是采用中文对话语料训练得到的 200 MB 3 阶 n-gram 语言模型，包含 46591 个 1 阶项，2549800 个 2 阶项和 2360752 个 3 阶项。输入特征是帧移为 10 ms 的 60 维 PLP 特征[17]。时延神经网络的输入窗长是前 2 帧后 2 帧共 5 帧，含 4 个隐层，隐层的上下文帧序列分别是{−1, 2}, {−3, 4}, {−7, 2}, {0}。

实验在一台 Dell R720 服务器上运行，配置为 2 颗 6 核心 Intel Xeon E5-2620 处理器和 128 GB RAM，解码器以 12 线程运行，结果中描述的实时率是各线程单线实时率的平均值。

6.1 帧同步方法和帧异步方法

表 1 显示了作为基线的帧同步系统和帧异步系统采用同一个时延神经网络的性能。在这组实验中，所有配置参数保持不变。在帧异步系统中，隐层的上下文帧序列取了近似，2-帧异步系统是{−2, 2}, {−4, 4}, {−8, 2}, {0}；3-帧异步系统是{−3, 0}, {−3, 3}, {−9, 0}, {0}; 4-帧异步系统是{−4, 0}, {−4, 4}, {−8, 0}, {0}。如前文所预期，2-帧异步系统得到了与基线系统相近的精度，并且加速大约 40%。这组结果显示语音信号中一部分的帧已经足够承载语音信号的声学信息。

6.2 跳帧方法

如前文所述，在跳帧系统中，帧数减少导致声学模型得分的动态范围相应变小，因此本文对语言模型的得分做了相应的修正，以使之与声学模型得分相匹配。

表 2 显示了 2-帧跳帧系统在不同配置下的结果。直接在基线系统中去除一半帧的做法导致了相当大的精度损失。另一方面，调整了语言模型权重因子和剪枝宽度之后，系统的精度损失变小，但与帧异步系统相比仍然有较大的差距。基于这个配置，通过引入跨单状态的 HMM 模型，系统获得了明显的精度增益，即与帧异步方法精度接近，且相对基线系统加速约 50%。

作为对以上结果的推广，我们尝试从语音序列中抛弃更多的帧。在这种推广中，系统每 n 帧取 1 帧用于计算。表 3 显示了系统采用跳帧帧数和 HMM 跨状态数的结果。可以看出，跨单状态的 HMM 模型结构已经不能满足超过 2 帧的跳帧系统的需要，在 3-帧跳帧系统中，需要采用跨 2 状态的 HMM 模型结构。另一方面，我们的 4-帧跳帧系统无法得到理想的结果。我们认为造成这一结果的原因主要是由于随着跳帧数的增长，即使采用更大的跨状态数，为了保证每个音素至少占有一次观测而要求的音素最短长度仍然会上升，例如在 4-帧跳帧系统中上升到了 40 ms。因此，最大跳帧数与音素的持续时间有关，从而不能是无上限的。

表 1 帧同步系统和帧异步系统的性能

	字错误率(%)	实时率	相对帧同步系统的精度损失(%)	相对帧同步系统的时间开销(%)
帧同步系统	30.0	0.436	0	100
2-帧异步系统	30.5	0.270	1.67	61.9
3-帧异步系统	30.6	0.208	2.00	47.7
4-帧异步系统	31.2	0.181	4.00	41.5

表 2 2-帧跳帧系统的性能

	字错误率(%)	实时率	相对帧同步系统的精度损失(%)	相对帧同步系统的时间开销(%)
2-帧跳帧系统，参数不变	35.2	0.258	17.30	59.2
2-帧跳帧系统，语言模型权重因子和剪枝宽度减半	33.7	0.218	12.30	50.0
2-帧跳帧系统，跨单状态 HMM，语言模型权重因子和剪枝宽度减半	31.1	0.219	3.67	50.2

6.3 可变帧率方法

在基线系统和 n-帧跳帧系统的基础上，在令牌传递环节采用可变帧率方法的结果如表 4 所示。表中令牌传递环节的跳帧比指令牌传递中被跳过的帧数占神经网络输出的帧数的比例。

表 4 结果显示，可变帧率方法相对于固定帧率的基线方法和跳帧方法，提高了计算速度，主要是提高了令牌传递环节的速度。另一方面，随着跳帧数的增加，令牌传递环节的加速比也减小，这是由于跳帧之后的神经网络输出相似性比逐帧的小，从而冗余信息量少，可以抛弃的帧变少。注意到令牌传递环节的跳帧比与神经网络计算环节的跳帧数大致成反比，这表明本方法所保留的帧数在不同的神经网络计算配置下是稳定的，即对应令牌传递过程中所需的最小帧数。这一结果符合上文对可变帧率方法的分析。

6.4 前向计算中近似算法的评价

在时延神经网络的帧异步方法和跳帧方法计算中，本文采用了上下文帧序列近似的方法。为了验证这一方法对计算精度的影响，下面的实验比较了不采用帧序列近似和采用帧序列近似时帧异步方法和跳帧方法的精度差异。在这组实验中，时延神经网络的前向计算是逐帧进行的，因此非帧序列近似系统应该有和逐帧系统近似的速度。

表 5 中的结果表明，帧序列近似相对非帧序列近似，带来了 1%~2%的精度损失。考虑到采用帧序列近似之后带来的计算速度提升，我们认为这种精度损失是可以接受的。

6.5 与前馈神经网络的比较

如前文所述，时延神经网络的每个隐层在帧异步系统和跳帧系统中都需要对上下文帧序列取近似。当系统采用前馈神经网络时，由于只有输出概率取近似，系统的精度有可能更高。其实验结果见表 6。可以看出，帧异步系统和跳帧系统采用前馈神经网络时的加速比与采用时延神经网络相近，而性能损失更小。正如我们在前文中所分析的，时延神经网络在进行跳帧计算时进行了多次近似，在我们的实验中每一层都有 1/2 的帧采用了相邻帧来代

表 3 n-帧跳帧系统的性能

	字错误率(%)	实时率	相对帧同步系统的精度损失(%)	相对帧同步系统的时间开销(%)
2-帧跳帧系统，跨单状态 HMM	31.1	0.219	3.67	50.2
3-帧跳帧系统，跨单状态 HMM	35.5	0.152	18.3	34.8
3-帧跳帧系统，跨多状态 HMM	31.0	0.173	3.33	39.7
4-帧跳帧系统，跨多状态 HMM	32.4	0.123	8.00	28.2

表 4 n-帧可变帧率系统的性能

	字错误率(%)	令牌传递环节的跳帧比(%)	实时率	相对帧同步系统的精度损失(%)	相对帧同步系统的时间开销(%)
1-帧可变帧率系统	30.2	58.2	0.387	0.67	88.7
2-帧可变帧率系统	31.3	30.4	0.208	4.33	47.7
3-帧可变帧率系统	31.5	18.0	0.154	5.00	35.3
4-帧可变帧率系统	33.6	13.1	0.121	12.00	27.8

表 5 采用和不采用帧序列近似的精度比较(%)

	帧序列近似系统字错误率	相对帧同步系统的精度损失	非帧序列近似系统字错误率	相对帧同步系统的精度损失
2-帧帧异步系统	30.5	1.67	30.2	0.67
3-帧帧异步系统	30.6	2.00	30.3	1.00
4-帧帧异步系统	31.2	4.00	30.6	2.00
2-帧跳帧系统	31.3	3.67	30.7	2.33
3-帧跳帧系统	31.0	3.33	30.8	2.67
4-帧跳帧系统	32.4	8.00	32.1	7.00

表 6 前馈神经网络用于帧同步系统、帧异步系统和跳帧系统的性能

	字错误率(%)	实时率	相对帧同步系统的精度损失(%)	相对帧同步系统的时间开销(%)
帧同步系统	43.4	0.713	0	100
2 帧异步系统	43.5	0.511	0.23	71.7
2-帧跳帧系统，跨单状态 HMM	43.6	0.377	0.46	52.9
3-帧异步系统	43.7	0.449	0.69	63.0
3-帧跳帧系统，跨多状态 HMM	43.8	0.292	0.92	41.0
4-帧异步系统	44.1	0.421	1.61	59.0
4-帧跳帧系统，跨多状态 HMM	43.9	0.208	1.15	29.2

替，因此神经网络对声学模型后验概率的预测会有一定的偏差。我们认为，考虑到加速比以及时延神经网络相较与前馈神经网络更高的精度，2-帧跳帧系统的相对 3%性能损失是可以接受的。在 3-帧异步系统和 3-帧跳帧系统的实验也得到了类似的结果。

7 结束语

本文提出了一种适用于递归神经网络的语音识别快速解码算法，其可以显著减小语音识别系统的计算时间复杂度，同时保持精度接近不变。通过有规律地抛弃输入中有重叠的声学特征帧和神经网络输出相似的相邻令牌传递帧，并简单改变 HMM 模型结构，这种方法可以加速解码过程而不需要对递归神经网络模型和 WFST 网络做改动。实验证明，方法在时延神经网络上的最佳性能是：相对精度损失 5.00%时解码时间缩短 64.7%。我们认为，相对于解码速度的增益，精度损失是可以接受的。此外，这种方法被证明并不依赖于神经网络自身的结构，因此对于不同的神经网络具有良好的兼容性，并且有望与其他的计算加速方法兼容。

参 考 文 献

[1] GRAVES Alex, JAITLY Navdeep, and MOHAMED Abdel-rahman. Hybrid speech recognition with deep bidirectional LSTM[C]. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Olomouc, Czech Republic, 2013: 273–278.

[2] SAK Hasim, SENIOR Andrew, and BEAUFAYS Françoise. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]. 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), Singapore, 2014: 338–342.

[3] NARAYANAN Arun, MISRA Ananya, and CHIN Kean. Large-scale, sequence-discriminative, joint adaptive training for masking-based robust ASR[C]. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, 2015: 3571–3575.

[4] LI Jinyu, MOHAMED Abdelrahman, ZWEIG Geoffrey, *et al.* Exploring multidimensional LSTMs for large vocabulary ASR[C]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016: 4940–4944.

[5] PEDDINTI Vijayaditya, POVEY Daniel, and KHUDANPUR Sanjeev. A time delay neural network architecture for efficient modeling of long temporal contexts[C]. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, 2015: 3214–3218.

[6] SNYDER David, GARCIA-ROMERO Daniel, and POVEY Daniel. Time delay deep neural network-based universal background models for speaker recognition[C]. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, USA, 2015: 92–97.

[7] PEDDINTI Vijayaditya, CHEN Guoguo, MANOHAR Vimal, *et al.* JHU ASpIRE system: robust LVCSR with TDNNs, i-vector adaptation, and RNN-LMs[C]. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, USA, 2015: 539–546.

[8] SEIDE Frank, LI Gang, and YU Dong. Conversational speech transcription using context-dependent deep neural networks[C]. 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), Florence, Italy, 2011: 437–440.

[9] SELTZER Michael L, YU Dong, and WANG Yongqiang. An investigation of deep neural networks for noise robust speech recognition[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013: 7398–7402.

[10] VANHOUCKE Vincent, DEVIN Matthieu, and HEIGOLD

Georg. Multiframe deep neural networks for acoustic modeling[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013: 7582–7585.

[11] MOORE Darren, DINES John, DOSS Mathew Magimai, *et al.* Juicer: A Weighted Finite-State Transducer Speech Decoder[M]. Berlin, Heidelberg, Springer, 2006: 285–296.

[12] YOUNG S J, RUSSELL N H, and THORNTON J H S. Token passing: A simple conceptual model for connected speech recognition systems[R]. CUED/F-INFENG/TR38, Engineering Department, Cambridge University, 1989.

[13] NOLDEN David, SCHLÜTER Ralf, and NEY Hermann. Extended search space pruning in LVCSR[C]. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012: 4429–4432.

[14] 郭宇弘. 基于加权有限状态转换机的语音识别系统研究[D]. [博士论文], 中国科学院大学, 2013: 1-20.
GUO Yuhong. Automatic speech recognition system based on weighted finite-state transducers[D]. [Ph.D. dissertation], University of Chinese Academy of Sciences, 2013: 1–20.

[15] RABINER Lawrence R and JUANG Biinghwang. An introduction to hidden Markov models[J]. *IEEE ASSP Magazine*, 1986, 3(1): 4–16. doi: 10.1109/MASSP.1986.1165342

[16] YOUNG Steve, EVERMANN Gunnar, GALES Mark, *et al.* The HTK Book Vol. 2[M]. Cambridge, Entropic Cambridge Research Laboratory, 1997: 59–210.

[17] ZHANG Qingqing, SOONG Frank, QIAN Yao, *et, al.* Improved modeling for F0 generation and V/U decision in HMM-based TTS[C]. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, USA, 2010: 4606–4609.

张　舸：　男，1991 年生，博士生，研究方向为语音识别.

张鹏远：　男，1978 年生，研究员，研究方向为大词表非特定人连续语音识别、关键词检索、声学模型、鲁棒语音识别等.

潘接林：　男，1965 年生，研究员，博士生导师，研究方向为大词表非特定人连续语音识别、语音分析、声学模型、环境噪声、快速搜索算法等.

颜永红：　男，1967 年生，研究员、博士生导师，研究方向为大词表非特定人连续语音识别、语音信号前端处理、多媒体数据检索、言语生成与听觉感知等.