第 43 卷第 6 期　　声　学　学　报　　Vol. 43, No. 6
2018 年 11 月　　ACTA　ACUSTICA　　Nov., 2018

连接时序分类准则声学建模方法优化*

王智超[1†]　张鹏远[1]　潘接林[1]　颜永红[1,2]

(1　中国科学院语言声学与内容理解重点实验室　北京　100190)

(2　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)

2017 年 1 月 4 日收到

2017 年 6 月 27 日定稿

摘要　对基于连接时序分类准则 (connectionist temporal classification, CTC) 的端到端声学建模方法进行研究和优化。研究分析了不同声学特征、建模单元以及神经网络结构对 CTC 声学模型性能的影响，针对 CTC 模型中 blank 符号共享导致的建模缺陷提出了建模单元相关的非共享 blank 方法进行改进，并引入融合建模单元关联信息的模型初始化方法进一步提高 CTC 模型的性能。在 300 小时标准英文数据集 Switchboard 的实验结果显示，结合非共享 blank、时延神经网络以及融合建模单元关联信息的初始化方法，CTC 声学模型相对于基线系统在词错误率上取得绝对 1.1% 的下降，同时在训练速度上取得 3.3 倍的提高，实验结果证明本文针对端到端声学建模提出的优化方法是有效的。

PACS 数：43.60, 43.72

Optimization of acoustic modeling method with connectionist temporal classification criterion

WANG Zhichao[1]　ZHANG Pengyuan[1]　PAN Jielin[1]　YAN Yonghong[2]

(1　*The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences* Beijing　100190)

(2　*Institute of Acoustics, Chinese Academy of Sciences* Beijing　100190)

Received Jan. 4, 2017

Revised Jun. 27, 2017

Abstract　The end-to-end acoustic modeling method based on connectionist temporal classification (CTC) criterion is studied and optimized in this paper. We study on the performance of CTC acoustic models with different acoustic features, modeling units and architectures. A modeling unit related unshared blank method is proposed to improve the modeling defects caused by the blank sharing in the CTC model. And a model initialization method that put the association information between the modeling units into the neural network is introduced to further improves the performance of the CTC model. Experiments were carried out on the 300-hour Switchboard dataset. Results show that the proposed CTC model trained with non-shared blanks, time-delay neural networks and the initialization method with association information between the modeling units achieves an absolute 1.1% reduction in word error rate as well as a 3.3-time speedup over the baseline system. The experimental results show that the proposed method is effective for end-to-end acoustic modeling.

引言

近年来，深度神经网络 (deep neutal network, DNN) 在众多领域都得到了广泛应用[1−2]，而其被引入到语音识别领域取代传统的混合高斯模型 (Gaussian mixture model, GMM) 用于声学建模，取得了巨大成功。神经网络模型通常采用交叉熵 (cross-

* 国家重点研发计划重点专项 (2016YFB0801203, 2016YFB0801200) 资助

† 通讯作者：王智超，wangzhichao214232@sogou-inc.com

entropy, CE) 准则进行参数训练，要求训练数据的标注对齐到帧。因此首先需要训练一个 GMM 种子模型对训练数据进行强制对齐，由于对齐结果对神经网络模型的最终性能有着重要的影响，为提高标注的对齐精度，GMM 模型需要经过上下文无关的音素建模、上下文相关的三因子音素建模以及鉴别性训练等多个步骤，训练过程复杂且周期较长。因此找到一种端到端的学习方法来简化神经网络的训练流程是当前的一个研究热点 [3−6]。

CTC 准则不同于 CE 准则，直接对网络输入输出序列之间的映射关系进行建模。其优化目标是最大化整句标注序列在网络中的输出概率，而并不关心序列中每个音素的时间信息，让神经网络自动学习输入特征到输出文本序列间的对齐关系，从而实现端到端学习，在语音识别领域也受到了越来越多的关注 [7−11]。

CTC 准则虽然简化了神经网络声学模型的训练步骤，其模型识别精度相对于传统的 CE 模型仍然有一定的差距。为提高 CTC 声学模型的性能，本文在一个公开的 300 小时英文电话语音数据集 Switchboard-1 上进行实验，对 CTC 声学模型在以下几个方面进行了研究探索和改进：

(1) 声学特征和建模单元：首先本文对比了 MFCC 和 fbank 两种常用的声学特征对 CTC 模型性能的影响，找到更适合 CTC 模型的声学特征。然后对比了在不同建模单元下 CTC 模型的表现。由于 CTC 的特性使其可容忍较大的建模单元，因此本文对比了两种音素级别的建模单元：上下文独立的单因子音素和上下文相关的三因子音素。实验结果表明对表达更加精细的三因子音素建模可大大提高 CTC 模型的识别精度，这与文献 11 中的结论一致。

(2) 非共享 blank: CTC 方法在网络输出节点中添加一个额外的 blank 符号用来进行辅助对齐。本文提出一种基于建模单元的独立 blank 方法，取代传统 CTC 中所有建模单元共享一个 blank 的方法来提高 CTC 模型的性能。基于建模单元的独立 blank 可以使每个 blank 符号的表达意义更加清晰明确，并可以让 CTC 模型更加有效的利用训练数据。

(3) 网络结构：CTC 准则常结合具有长时建模能力的神经网络结构进行建模 [3−11]，如递归神经网络 (recurrent neural network, RNN) 或带有长短时记忆 (long short-term memory, LSTM)[12] 单元的 RNN。时延神经网络 (time-delay neural network, TDNN)[13−14] 同样具有长时建模能力，但其结构相比于递归神经网络更加简单。本文对比了分别采用 TDNN 和 LSTM 进行建模的 CTC 模型性能。实验结果表明 TDNN-CTC 模型可以获得与 LSTM-CTC 模型相当的识别精度，但在训练速度上 TDNN-CTC 模型更快，是 LSTM-CTC 模型的 3.3 倍。

(4) 模型初始化：文献 20 中提出了一种神经网络参数初始化方法，将建模单元之间的关联信息在模型初始化时融入到神经网络之中。本文引入该方法对 TDNN-CTC 模型进行初始化，并针对本文提出的非共享 blank 符号在初始化时尝试了两种不同的处理方式：将 blank 视为独立符号和将 blank 关联到其相关的音素符号进行初始化。实验发现将 blank 视为独立符号进行初始化的效果更好，对比随机初始化的 TDNN-CTC 模型在识别词错误率上取得了相对 2.0% 的下降。

1 CTC 模型训练

1.1 CTC 准则

CTC 准则由 Alex Graves 提出 [16]，最初是为解决未分段数据的自动标注问题。在 CTC 模型训练时一个额外的 blank 符号被加入到训练数据标注的符号表中，可以用于表示网络对预测不确定时的输出状态。另外 CTC 准则还定义了一个变换 F，若网络的一个输出序列能通过变换 F 映射到正确标注序列，则称该输出序列为一条 CTC 路径。F 的变换规则为：先去除序列中相邻 blank 符号之间的重复标注符号，然后去除序列中的 blank 符号，如下所示 ($\oslash$ 表示 blank):

$$F(\oslash\mathrm{AA}\oslash\oslash\mathrm{BB}\oslash\mathrm{BC}) = \mathrm{ABBC}. \tag{1}$$

因此标注序列 $\boldsymbol{z}$ 可以表示为可以映射到 $\boldsymbol{z}$ 的所有 CTC 路径的集合。那么在给定输入序列 $\boldsymbol{x}$ 的情况下，标注序列 $\boldsymbol{z}$ 的后验概率的计算如式 (2) 所示：

$$\Pr(\boldsymbol{z}|\boldsymbol{x}) = \sum_{p\in\phi(\boldsymbol{z})} \Pr(p|\boldsymbol{x}), \tag{2}$$

其中，$\phi(\boldsymbol{z})$ 为属于 $\boldsymbol{z}$ 的所有 CTC 路径的集合，p 为 $\phi(\boldsymbol{z})$ 中的一条 CTC 路径。$\Pr(\boldsymbol{z}|\boldsymbol{x})$ 可以通过前后向算法 (forward-backward algorithm) 计算得到：

$$\Pr(\boldsymbol{z}|\boldsymbol{x}) = \sum_{u=1}^{|\boldsymbol{z}'|} \alpha(t,u)\beta(t,u), \tag{3}$$

其中，$\boldsymbol{z}'$ 表示带有 blank 符号的标注序列。为了进行前后向计算，需要对原始标注 $\boldsymbol{z}$ 进行变换：首先将 blank 符号插入 $\boldsymbol{z}$ 的首尾，然后在每两个相邻标注符号间也插入一个 blank 符号，构成新的标注 $\boldsymbol{z}'$,

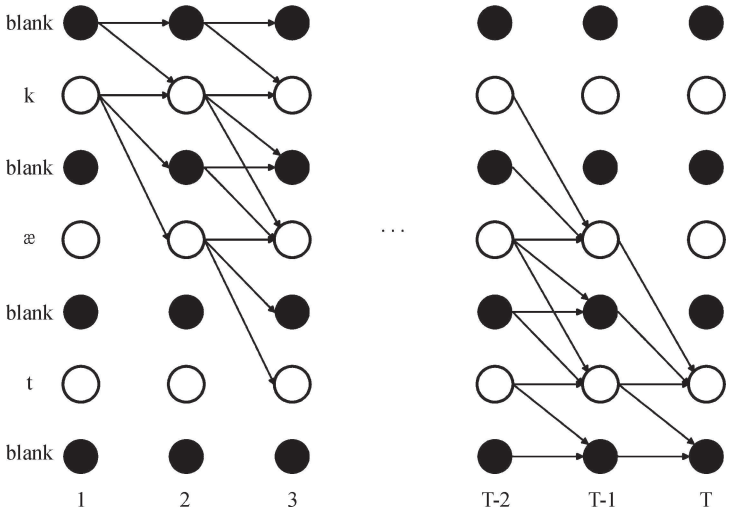

图 1 CTC 前后向计算示意图

因此 $|\boldsymbol{z}'|=2|\boldsymbol{z}|+1$。$\alpha(t,u)$ 表示所有在时刻 t 以符号 u 结尾的 CTC 路径概率之和，并可以根据 $t-1$ 时刻的 α 值递归计算得到；而类似地，$\beta(t,u)$ 表示所有在时刻 t 以符号 u 开始的所有 CTC 路径概率之和，并可以根据 $t+1$ 时刻的 β 值递归计算得到。

为表示前后向计算的过程，我们以 “CAT” 这个单词的音素序列为例，将经过变换后得到的标注 $\boldsymbol{z}'$ 在时间轴上展开，如图 1 所示。图中黑圈代表 blank，白圈代表音素符号，起始时刻状态只能是 blank 符号或者标注 $\boldsymbol{z}$ 中的第一个音素符号，中间时刻状态的跳转规则为：若当前状态为音素，则下一时刻状态可以停留在当前音素或者跳转到 blank 符号，也可直接跳过 blank 到 $\boldsymbol{z}$ 中的下一个音素符号；而如果当前状态是 blank，则下一时刻只能停留在 blank 符号或跳到下个音素符号。

CTC 的训练目标就是最大化后验概率 $\Pr(\boldsymbol{z}|\boldsymbol{x})$，因此定义在所有训练数据集 $\boldsymbol{S}$ 上的损失函数为：

$$L(\boldsymbol{S})=-\sum_{(\boldsymbol{x},\boldsymbol{z})\in\boldsymbol{S}}\ln\Pr(\boldsymbol{z}|\boldsymbol{x}). \tag{4}$$

最终，可以对损失函数求导得出网络的误差，然后通过误差反向回传 (error back-propagation) 算法进行梯度计算，更新神经网络的参数。

1.2 非共享 blank

CTC 中 blank 符号的其中一个作用是可以减轻网络的输出负担，避免网络针对每一帧输入都强制给出一个标注的预测。这样一方面当输入特征为语音中的停顿状态或其他非语音噪声时，网络可以输出 blank 符号；另一方面当输入特征为介于两个相邻发音之间的混淆状态时，网络也可以输出 blank 符号。例如图 2 中所表示的情况，英文单词 “you” 的发音中，两个音素 /j/ 和 /u:/ 之间存在协同发音的情况，/j/ 的结尾部分和 /u:/ 的起始部分存在重叠，没有明确的界限。针对这种情况，在 CE 模型训练时，会根据标注的强制对齐结果进行强制硬判决，而 CTC 模型则可通过输出 blank 符号来表达两个相邻发音之间的混淆状态。

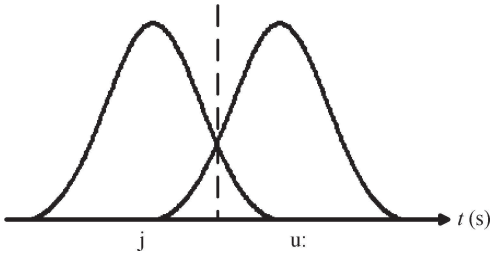

图 2 协同发音示意图

在 CTC 模型的实际训练过程中，由于所有建模单元的边界状态都用同一个 blank 符号表达，使得模型在训练过程中越来越倾向输出 blank，最终导致神经网络的输出在绝大部分情况为 blank，之间夹杂着少量的标注符号。Blank 符号的输出占比过高会导致网络参数的偏离，使得大量训练数据都归属于 blank，导致模型在训练的过程中对训练数据的利用率降低，对模型的性能产生一定影响。

针对这个问题，本文提出了一种基于建模单元的非共享 blank 方法对 CTC 模型进行改进。该方法给每个建模单元分配一个与其相关的 blank 符号，用独立的 blank 取代原先 blank 共享的方式。采用独立 blank 符号进行建模有两点好处：首先，相邻发音单元间的混淆状态可以用音素相关的 blank 符号来表示，这使得不同相邻建模单元间的混淆状态得以有效的区分，使模型建模更加精细；其次，独立的 blank 符号增加了与建模单元之间的相关性，这样即

使在 CTC 训练过程中大量输出 blank 符号，训练数据也会被关联到相应的建模单元上，可以防止大量训练数据被 blank 符号吸收，从而提高了训练数据利用率。

采用非共享 blank 符号建模的 CTC 路径的限制需要进行相应改变：每个建模单元后只能出现与其相关的 blank 符号，直至标注中下一个建模单元符号出现。因此，CTC 路径到标注之间的映射 F 可用式 (5) 表示：

$$F(\mathrm{AA}\oslash_{\mathrm{A}}\oslash_{\mathrm{A}}\mathrm{BB}\oslash_{\mathrm{B}}\mathrm{BC}\oslash_{\mathrm{C}})=\mathrm{ABBC}, \tag{5}$$

其中，$\oslash_{\boldsymbol{x}}$ 表示属于建模单元 $\boldsymbol{x}$ 的 blank 符号。

与此同时，采用非共享 blank 进行建模，给定输入序列 $\boldsymbol{x}$, 标注序列 $\boldsymbol{z}$ 的后验概率 $\Pr(\boldsymbol{z}|\boldsymbol{x})$ 仍然可以通过式 (3) 利用前后向算法计算得到。但前后向的计算过程发生了改变：在进行前后向计算前，不再在原始标注 $\boldsymbol{z}$ 的起始位置插入 blank 符号，只在每个标注符号后插入其相关的 blank 符号，得到 $\boldsymbol{z}'(|\boldsymbol{z}'|=2|\boldsymbol{z}|)$。我们仍以 “CAT” 这个单词的音素序列为例，对基于非共享 blank 建模的 CTC 前后向计算过程进行描述，如图 3 所示。图中黑圈代表 blank, 白圈代表音素符号，起始时刻状态只能是标注 $\boldsymbol{z}$ 中的第一个音素符号，中间时刻状态的跳转规则为：若当前状态为音素，则下一时刻状态可以停留在当前音素或者跳转到其相关的 blank 符号，也可直接跳过 blank 到 $\boldsymbol{z}$ 中的下一个音素符号；而如果当前状态是 blank, 则下一时刻只能停留在与当前时刻一致的 blank 符号或跳到下个音素符号。

另外还需要注意的是，采用建模单元相关 blank 进行建模的 CTC 模型在解码时也需要进行相应的处理。本文使用的解码器是基于文献 8 中的方法构建的，该文中提出一种基于加权有限状态机 (weighted finite-state transducers, WFST) 的 CTC 模型解码方法。在该方法中，解码所用的 WFST 网络由 3 个独立的 WSFT 网络 G, L 和 T 复合而成，如式 (6) 所示：

$$\mathrm{S}=\mathrm{T}\circ\min(\det(\mathrm{L}\circ\mathrm{G})), \tag{6}$$

其中，S 是解码网络，G 和 L 分别表示语言模型 WFST 和词典 WFST, “$\circ$” 代表复合运算，min 和 det 分别表示对网络进行最小化和确定化操作 [21], T 则表示 CTC 标注到词典单元 (如音素) 的映射网络。对于一个词典单元 A, 它的 T 中包含所有可能映射到 A 的帧级标注序列。采用共享 blank 建模时，T 所做的映射可用 1.1 节中式 (1) 所述的 F 变换来表达。例如，在处理了 5 帧之后，模型可能会生成 3 个标注序列 “AAAAA”, “$\oslash$AAAA”, “$\oslash$AA $\oslash$ $\oslash$”, T 将这 3 个序列都映射到词典单元 “A”。而采用非共享 blank 建模时，T 映射改用 1.2 节中的式 (5) 来表示，建模单元相关 blank 只会映射到其相关的建模单元，如 “AAA$\oslash_{\mathrm{A}}\oslash_{\mathrm{A}}$”, “AAAA$\oslash_{\mathrm{A}}$” 和 “A$\oslash_{\mathrm{A}}$ $\oslash_{\mathrm{A}}$ $\oslash_{\mathrm{A}}\oslash_{\mathrm{A}}$” 这些序列会被映射到 “A”, 为了便于区分，我们用 $\overline{\mathrm{T}}$ 来表示该 WFST。由于在本文中我们还使用了上下文相关的三音素进行建模，因此解码网络中还需要包含三音素到音素的 WFST 网络 C[21], 最终的解码网络 S 的构建可由式 (7) 表示：

$$\mathrm{S}=\overline{\mathrm{T}}\circ\min(\det(\mathrm{C}\circ\min(\det(\mathrm{L}\circ\mathrm{G})))). \tag{7}$$

1.3 时延神经网络

CTC 模型允许在任意时刻输出有效标注符号，如果能充分考虑历史和未来信息，就可以在有足够判断依据的情况下对当前状态给出更加准确的预测。因此 CTC 准则通常结合具有长时记忆能力的 LSTM 网络进行建模 [3−11]。但由于 LSTM 网络中递归层的存在，相比于前馈神经网络其结构更加复杂，并行训练难度也更大。

引入文献 14 中提出的一种采样结构的 TDNN 网络结合 CTC 进行建模。TDNN 网络中每一层的

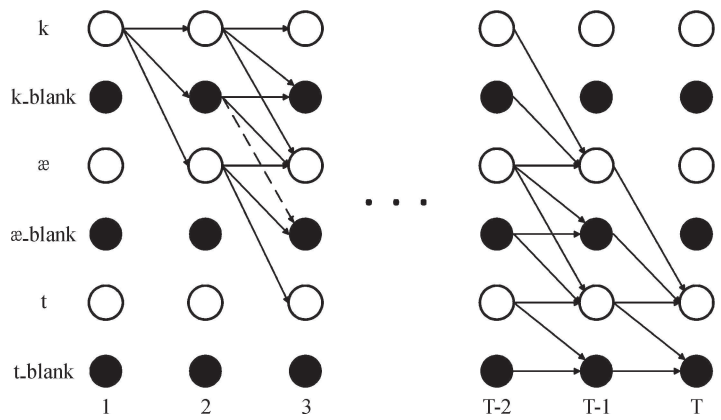

图 3　基于非共享 blank 的 CTC 前后向计算示意图

输入来自前一层若干时刻输出的拼接，通过这种方式深层的网络看到更长的上下文信息，从而具有长时建模能力。由于 TDNN 是一种前馈型网络，网络结构中不存在递归环，且神经元节点的结构较为简单，没有复杂的门控制信息的流动，因此其训练复杂度要远低于 LSTM 网络。

2 实验与结果

使用 Kaldi[17] 在英文电话语音数据集 Switchboard 上展开实验，该数据集共包含约 300 小时语音，其中 Switchboard-1 作为声学模型的训练集，Hub5'00 作为测试集。实验中我们使用了两个 N-gram 语言模型，其中小语言模型是由 Switchboard 数据集 (约含 3M 单词) 训练的三元模型，用于一遍解码，大语言模型是一个四元模型，由 Switchboard 和 Fisher 数据集 (共约 24M 单词) 训练得到，用于重打分。在实验中我们使用文献 18 中的方法对所有训练语音数据进行了数据增强处理，对原始语音数据在语速和声音幅度两方面加了 0.9, 1.0 和 1.1 三种比例的扰动。数据扰动的目的是增加数据的丰富性，提高模型的鲁棒性和泛化性。

2.1 声学特征和建模单元实验

为研究 CTC 声学模型的特性和构建基线系统，我们首先对比了不同声学特征和建模单元对 CTC 模型性能的影响。选取了神经网络声学模型常用的两种特征：fbank 特征[8,11] 和 MFCC[23] 特征进行实验。对于 fbank 特征，我们采取文献 11 的方法来构造网络的输入：提取 80 维的 fbank，每一帧与其前四帧后三帧进行拼接，得到 640 维输入特征；对于 MFCC 特征，我们提取 40 维特征并对每一帧与其前后各两帧进行拼接，共 200 维作为网络的输入。每种特征训练两个 CTC 模型：独立于上下文信息的单音素 (context-independent phone, CI-phone) 模型和包含上下文信息的三音素 (context-dependent phone, CD-phone) 模型。其中，CI-phone 模型的网络输出节点数为 47，包含单因子音素，噪声和 blank; CD-phone 模型的网络输出节点数为 3541，包含 3540 个决策树聚类后得到的三音素和一个 blank。实验中的网络结构统一为 3 层的单向 LSTM 网络，每个 LSTM 层包含 1024 个记忆单元，表 1 给出了实验结果。

表 1 中的第 3, 4 列分别代表一遍解码和重打分的识别词错误率结果。从实验结果中可看出，采用 CD-phone 建模给 CTC 声学模型带来识别词错误率的大幅下降，这与文献 11 中的结论一致，同时也跟传统的 GMM 和 CE 模型所观察到的现象一致。CD-phone 包含的信息更加丰富，建模也更加精细，从而带来模型性能的提升。同时，可以看出 CTC 模型在 fbank 和 MFCC 两种不同特征上的识别结果基本一致，由于 fbank 特征的输入维数更大，计算量也更大，因此接下来实验中我们采用 MFCC 特征作为输入，将结合 LSTM 网络结构和 CD-phone 进行建模的 CTC 声学模型作为基线系统。

表 1 声学特征和建模单元实验结果

特征	建模单元	词错误率 (WER%)	
		3-gram	4-gram
Fbank	CI-phone	23.5	21.9
	CD-phone	21.6	20.3
MFCC	CI-phone	23.5	22.1
	CD-phone	21.5	20.3

2.2 非共享 blank

本实验对 1.2 节中描述的非共享 blank 方法的有效性进行验证，实验的其它配置与基线系统一致，网络的输出节点数为 7080，其中包括 3540 个经过决策树聚类后的 CD-phone 和 3540 个独立的 blank。表 2 给出了与基线系统的对比结果。

表 2 采用非共享 blank 的 CTC 模型与基线对比实验结果

模型	词错误率 (WER%)	
	3-gram	4-gram
基线	21.5	20.3
非共享 blank	20.8	19.6

表 2 的结果可以看出在语言模型为 3-gram 和 4-gram 的条件下，采用非共享 blank 的 CTC 模型相对于基线模型，在识别词错误率上均有绝对 0.7% 的下降。性能提升来自于两方面：第一是非共享 blank 是音素相关的，表达意义更加清晰，使 CTC 模型的建模粒度更加精细；第二是非共享 blank 提高了模型对训练数据的利用率，同时防止大量输出唯一的 blank 导致网络参数过于倾向输出 blank 而产生偏离。但是非共享 blank 会大量增加输出节点数，从而增加计算量，在本实验中，相比于基线系统，非共享 blank 模型的训练速度下降 20%。

2.3 TDNN-CTC 模型

实验中采用 TDNN 网络取代 LSTM 网络进行 CTC 建模。所用 TDNN 为 7 层网络，每层 576 个节点，网络每层的输入配置分别为：$\{-1,0,1\}$, $\{-1,0,1,2\}$, $\{-3,0,3\}$, $\{-3,0,3\}$, $\{-3,0,3\}$, $\{-6,-3,0\}$,

{0}, 其中第 1 层的 $\{-1,0,1\}$ 配置代表将输入特征的 $t-1, t, t+1$ 三帧拼接作为网络的输入， t 代表当前帧。表 3 中给出了 TDNN-CTC 模型与 LSTM-CTC 模型的性能对比结果。表中模型均采用非共享 blank 建模。

表 3 采用 TDNN/LSTM-CTC 模型对比实验结果

建模单元	词错误率 (WER%)		kfps
	3-gram	4-gram	
LSTM	20.8	19.6	1.9
TDNN	20.8	19.5	6.3

表 3 中的最后一列表示了模型训练时的速度，单位是 1000 帧 /s。可以看出 TDNN-CTC 模型在词错误率上表现和 LSTM-CTC 模型基本一致，这是由于通过逐层增加帧扩展， TDNN 网络的深层结构可以接收足够的上下文信息，在本实验的配置中，第 7 层 TDNN 共可接收 30 帧扩展信息，包括 17 帧历史信息和 12 帧未来信息。因此 TDNN 网络在进行预测时所依据的信息量不输于 LSTM 网络通过记忆所获得的信息量。但 TDNN 网络的训练速度是 LSTM 网络的 3.3 倍，速度的差异来自于两方面：首先，本文使用的 LSTM 网络包含 21M 参数，而 TDNN 网络的参数量为 9.8M, LSTM 网络参数量是 TDNN 网络的 2.1 倍；其次，实验中采用文献 19 中提出的方法训练 LSTM 网络，将一句话切分成固定长度的数据块采用 minibatch 的方式进行并行训练，为弥补语句切分造成的上下文信息丢失，相邻的数据块之间有部分重叠，导致 LSTM 网络在训练过程中产生大量的冗余计算。

2.4 模型初始化实验

之前的实验中, 神经网络的参数都是随机初始化的，在文献 20 中提出一种神经网络初始化方法，可以将建模单元之间的相互关联信息融合到神经网络之中, 从而提高模型的性能。本实验中我们在 2.3 节中的 TDNN-CTC 模型基础上引入该初始化方法并针对 CTC 模型中的 blank 符号对该方法进行调整。

首先找出每个 CI-phone 与 CD-phone 之间的映射关系; 然后从最后一层隐层中找出若干代表神经元节点, 每个节点代表一个 CI-phone, 本实验为 46 个。再根据映射关系将代表神经元与对应的网络输出节点之间的权重初始化为强连接，网络的其他权值仍进行随机初始化。对于本文提出的非共享 blank 符号，我们尝试了两种处理方法：第一种将每个 blank 与它相关的 CD-phone 映射到同一个代表神经元节点；第二种是在最后一层隐层中再选出 46 个神经元节点代表 CI-blank, 与相应的 CD-blank 节点之间映射为强连接。我们对不同的强连接初始值进行了实验，但在表 4 中只给出了最佳 (强连接初始值为 9) 实验结果。

表 4 初始化方法对比实验结果

初始化方法	词错误率 (WER%)	
	3-gram	4-gram
随机	20.8	19.5
方法 1	20.7	19.3
方法 2	20.4	19.1

实验中的 3 个模型，除强连接权值以外的其它权值初始值保持一致, 以减少随机因素的干扰。可以看出, 该初始化方法进一步提高了 TDNN-CTC 模型的识别性能，且将 blank 符号视作与音素不同的独立符号进行处理带来的性能提升更大。不过相对于文献 20 中在 DNN 和 CNN 模型上取得平均 5% 的相对词错误率下降，TDNN-CTC 模型的识别词错误率下降只有相对 2%, 原因可能是本实验中 TDNN 网络的参数量是文献中的 1.3 倍, 过多的权重数量削弱了强连接权重的影响。

2.5 CTC 与 CE 准则对比实验

最后，本实验给出了 CTC 和 CE 两个准则的对比实验结果，实验中的 CTC 模型为 2.4 节中的最优 CTC 模型，CE 模型采用与 CTC 模型相同的 TDNN 结构， CD-phone 建模，并通过文献 21 的方法进行模型初始化，结果如表 5 所示。

表 5 CTC 和 CE 准则对比实验结果

准则	词错误率 (WER%)	
	3-gram	4-gram
CE	18.4	16.9
CTC	20.4	19.1

从表 5 中可以看出，在 Switchboard 数据集上 CTC 模型相比于 CE 模型，在识别词错误率上绝对高了 2.1%, 差距仍然比较明显。目前在 CTC 模型的相关工作中, 通常在较大规模的数据量 (数千甚至数万小时) 下才能取得较好的效果 [4,10−11,22] 。

3 结论

本文从声学特征，声学建模单元，blank 符号，神经网络结构以及模型初始化 5 个方面对 CTC 声学模型进行了较为全面的研究和分析。首先构建了

合理的基线系统，然后提出了非共享 blank 方法对 CTC 模型性能进行改进，实验表明该方法可以带来识别精度绝对 0.7% 的提高。接下来我们结合 TDNN 网络进行 CTC 建模与 LSTM-CTC 模型进行对比，发现在取得相同识别精度下的情况下，TDNN-CTC 模型由于参数量更少，模型结构更加简单，可以大幅缩短训练周期，模型训练速度约为 LSTM-CTC 的 3.3 倍。进一步，我们通过在初始化时将建模单元间的关联信息融合到神经网络之中，使得 TDNN-CTC 模型的识别精度绝对提升 0.4%，最终在语言模型为 3-gram 时，TDNNCTC 模型的识别精度达到 20.4%，相对于基线 LSTM-CTC 系统的 21.5%，性能绝对提高了 1.1%。但是在本文所用数据集下，CTC 模型在与 CE 模型的对比中在识别精度上还存在不小的差距，需要进一步的研究。

参 考 文 献

1 Hinton G, Deng L, Yu D *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 2012; **29**(6): 82—97

2 施成龙, 师芳芳, 张碧星. 利用深度神经网络和小波包变换进行缺陷类型分析. 声学学报， 2016; **41**(4): 499—506

3 Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. ICML, 2014: 1764—1772

4 Hannun A, Case C, Casper J *et al*. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv: 1412.5567, 2014

5 Chorowski J, Bahdanau D, Cho K *et al*. End-to-end continuous speech recognition using attention-based recurrent NN: first results. arXiv preprint arXiv:1412.1602, 2014

6 Bahdanau D, Chorowski J, Serdyuk D *et al*. End-to-end attention-based large vocabulary speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016: 4945—4949

7 Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013: 6645—6649

8 Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015: 167—174

9 Miao Y, Gowayyed M, Na X *et al*. An empirical exploration of CTC acoustic models. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016: 2623—2627

10 Sak H, Senior A, Rao K *et al*. Learning acoustic frame labeling for speech recognition with recurrent neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015: 4280—4284

11 Sak H, Senior A, Rao K *et al*. Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv: 1507.06947, 2015

12 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997; **9**(8): 1735—1780

13 Waibel A, Hanazawa T, Hinton G *et al*. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989; **37**(3): 328—339

14 Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. Proceedings of Interspeech, ISCA, 2015: 2440—2444

15 Kurata G, Kingsbury B. Improved neural network initialization by grouping context-dependent targets for acoustic modeling. Interspeech, 2016: 27—31

16 Graves A, Fernández S, Gomez F *et al*. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006: 369—376

17 Povey D, Ghoshal A, Boulianne G *et al*. The Kaldi speech recognition toolkit. IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, 2011 (EPFL-CONF-192584)

18 Ko T, Peddinti V, Povey D *et al*. Audio augmentation for speech recognition. Proceedings of Interspeech, 2015

19 Chen K, Yan Z J, Huo Q. A context-sensitive -chunk BPTT approach to training deep LSTM/BLSTM recurrent neural networks for offline handwriting recognition. Document Analysis and Recognition (ICDAR), 2015: 411—415

20 Kurata G, Kingsbury B. Improved neural network initialization by grouping context-dependent targets for acoustic modeling. Interspeech, 2016: 27—31

21 Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 2002; **16**(1): 69—88

22 Povey D, Peddinti V, Galvez D *et al*. Purely sequence-trained neural networks for ASR based on lattice-free MMI. Interspeech, 2016: 2751—2755

23 张晴晴, 潘接林, 颜永红. 基于发音特征的汉语普通话语音声学建模. 声学学报， 2010; **35**(2): 254—260