关键词检出的双向跨词解码算法

刘雨辰[1]，徐明星[1]
1.普适计算教育部重点实验室
清华信息科学与技术国家实验室(筹)
清华计算机科学与技术系，北京，100084

文 摘： 基于子词声学建模的关键词检出系统具有实时性好、词表可灵活配置的优点。在解码过程中，展开的词内搜索网络在词的首尾未能充分利用上下文相关的子词模型，而跨词的搜索网络具有难展开、展开规模大、复杂性高的缺点。本文在词内搜索网络上进行跨词搜索，通过反向搜索网络，逆向输入语音，与正向搜索交汇的双向搜索方法来降低剪枝风险。实验结果表明，跨词搜索算法相较于词内搜索有显著的性能提升，在两个测试集上分别相对提高27.8%、18.4%, 双向搜索策略对有剪枝的跨词搜索有一定的性能提升，分别相对提高 1.9%、1.7%。

关键词：关键词检出；解码算法；跨词搜索；双向搜索；语音识别
中图分类号： TP391.4；TN912.34

关键词检出是语音识别的一个分支，其目标是在一段连续语音中检测出预先设定好的关键词表中的词，在命令控制、语音监控、意图理解、音频信息检索等领域均有重要应用。

目前，关键词检出的主流方法可分为两大类[1]：1.基于声学建模的关键词检出：该方法对关键词和非关键词分别建模，通过并联非关键词的模型(Filler)和关键词模型形成搜索网络，在搜索网络中搜索解码从而检出关键词[2][3][4][5][6][7]; 2.基于大词表连续语音识别（LVCSR）的关键词检出（常被称作 Spoken Term Detection）：该方法先进行一遍 LVCSR，从得到的 N-Best 或 Lattice 结果中寻找待查关键词[8][9]。这种方法需要语言模型，搜索空间庞大，检出速度慢，无法处理词表外词(OOV),一般用于对语音数据进行离线预处理以便将来检索，不适用于在线识别系统。

基于声学建模的关键词检出的建模单元可以选取整词或子词。整词建模对每个关键词建立其 HMM 声学模型，对非关键词语音，建立补白模型(Filler)。整词建模在词表更换后需要重新训练模型，灵活性差，不能适应实际应用对快速更换词表的需求。子词建模对声学基元（如音素、三音素）进行建模，构建识别网络时，串联关键词各音素对应的声学基元模型得到关键词模型，并联所有拼音作为词的模型得到补白模型。这种方法的补白模型和关键词模型之间存在重叠，一般用惩罚分作为操作点控制检出率和误警。

在建立搜索网络时，针对词首和词尾处的子词选择，若不考虑词间的音素搭配，即选择的子词单元为双音素(biphone)，这样得到的是词内搜索网络，系统性能不易保证；若考虑跨词的音素搭配（如 triphone），则可以得到更为精确但也更加庞大的搜索网络[10]，大大增加了网络构造的复杂性。针对上述不足，本文提出了一种基于词内搜索网络的跨词搜索解码算法，大幅度提高了系统性能。

经典的帧同步解码不能完整遍历假设空间，路径合并时只保留最优路径，常用的 Beam 剪枝会随着语音流增长损失更多可能的假设。为此，本文提出了双向搜索解码算法，降低了解码的剪枝风险，从而提高了系统性能。

本文的组织结构如下：第一部分介绍关键词检出的系统架构和搜索网络的组织；第二部分介绍解码算法，详细介绍了本文提出的跨词搜索算法和双向搜索策略，第三部分给出了实验结果和分析，最后是结论和展望。

1 关键词检出系统

基于 HMM 的关键词检出系统如图 1 所示，首先对于语音输入提取特征，然后根据关键词列表、发音词典和 HMM/GMM 构建搜索网络，最后将语音特征在搜索网络上根据 HMM/GMM 的打分进行解码，最终获得检出的关键词。

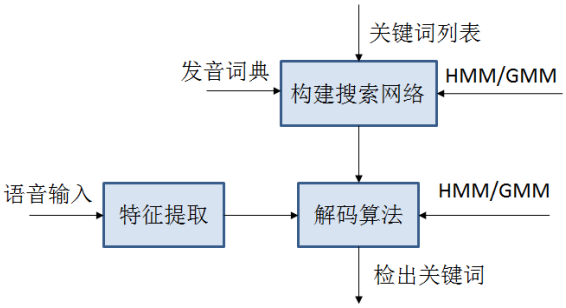

图 1　关键词检出系统

在子词建模中，为了和关键词区分，一般将所有音节和静音都作为补白模型，搜索网络是由关键词、音节和静音等模型并联构成的。关键词和音节的模型是子词 HMM 模型的串联，它们的末尾均设置词尾节点。此外，搜索网络的结束点还有一条反向边连接至起始点。图 2 展示的是搜索网络的拓扑结构。

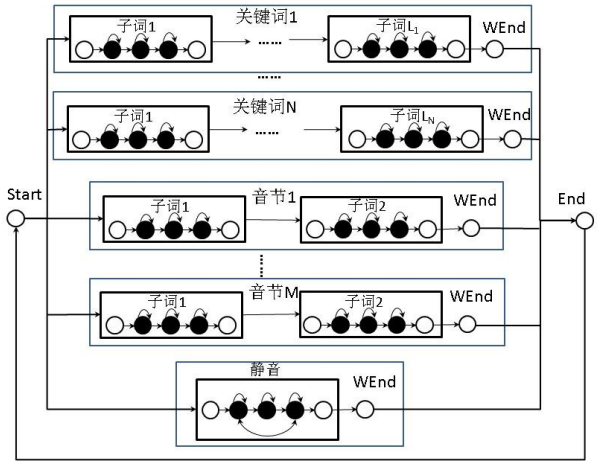

图 2　搜索网络

2　解码算法

2.1　词内搜索

基于 Viterbi 的解码算法是一种基于动态规划原理逐帧推进的帧同步算法。设语音的特征向量序列为$\{O_1,O_2,...,O_T\}$，搜索网络的输出节点为 s，则节点 s 上的推进的方程为：

$$V_t^s = \max_{(s',s)\in E}\left\{V_{t-1}^{s'} + w(s',s) + \log P(O_t \mid s)\right\} \qquad (2)$$

其中 t 为当前时间帧，E 为搜索网络的转移边的集合，这些转移边里既有 HMM 的转移边，也有网络串并联的骨架边。$w(s',s)$为转移边的对数似然分，其后一项为当前特征在该状态 GMM 上的对数似然分。用该方程得到的是搜索路径停留在状态 s 上处理完第 t 帧时的最大似然分。

2.2　跨词搜索

词内搜索在词边界处使用的是 biphone 模型，不是更精细的 triphone 模型，使得声学模型的性能没有被充分利用。虽然生成完整的跨词搜索网络可以解决这个问题，但会增大搜索网络的规模和构造复杂性。本文考虑在不改变词内搜索网络的条件下，在其上使用跨词搜索的策略，即在搜索进行过程中，根据跨词 biphone 所能搭配的不同音素，在词尾和词首产生多条分支来利用相应的 triphone 模型。这是一种无损的跨词搜索策略，具体算法如下。

对于搜索网络上的节点 s，搜索推进过程中根据 s 所处的位置（词首、词中或词尾）产生不同的分支，每条分支代表一个完整的 triphone。在基于词内搜索的搜索网络中，词首和词尾均是 biphone，补上不同音素后形成不同的 triphone，将经过词首和词尾的所有搜索路径经过所有可能的 triphone 模型，则会产生多条路径分支。图 3 给出了基于跨词搜索策略的分支传递流程的示意图。

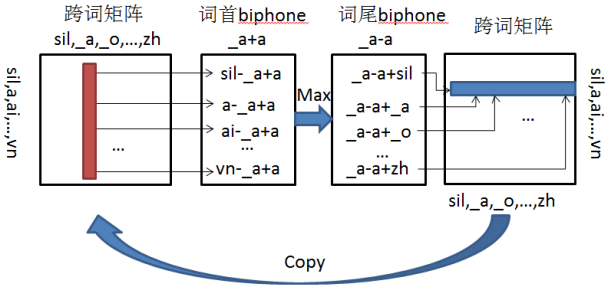

图 3　跨词搜索传递流程

在词首子词节点处产生的新分支，通过补齐节点 biphone 缺失的左音素(即韵母)得到，该音素与搜索路径中前驱词的词尾音素是对应的。类似的，在词尾子词节点处产生的分支，通过补齐节点缺失的右音素(即声母)得到，该音素与搜索路径中的后继词的词首音素相对应。词中的节点都具有完整的 triphone，只产生一条分支。

子词内部各状态中的分支（搜索路径）代表的 triphone 是相同的，故搜索过程中的状态传递与分支是一一对应的。对于相邻子词间的分支传递（路径扩展），先从前驱子词传出的分支中选出一个最优（似然分最大），然后将该分支（搜索路径）传递（扩展）给后继子词的各个分支。

以上分支传递（路径扩展）在搜索网络中从起始节点开始向后推进，所有分支到搜索网络的结束点处收束到一起，然后再传递回搜索网络的开头，进行下一轮推进。为了在收束分支时无损地保留跨词信息，在词边界（即搜索网络的结束

点）处建立一个以音素为下标的跨词矩阵 M，其行下标为到达词尾的跨词分支的词尾音素 p_e，列下标为跨词分支对应 triphone 所搭配的下一词首音素 p_s，元素 $M(p_e, p_s)$指向路径似然分最大的分支。当一帧推进传递到词尾时，所有分支根据词尾音素 p_e和假设的后继词首音素 p_s把路径信息保存到矩阵 M 的相应元素处，然后在搜索网络的起点处，将 M 中相应元素所指路径扩展给词首分支，为下一帧推进作好准备。

跨词搜索的收束和分发过程如下式所示：

$$M_{t+1}(p_e, p_s) = \max\left\{V_t^b \mid s(b) \in W_e(p_e), next(b) = p_s\right\}$$
$$V_t^b = \max\{M_t(p_e, p_s) \mid s(b) \in W_s(p_s), prev(b) = p_e\} \quad (3)$$

其中，$M(p_e,p_s)$表示跨词矩阵的第 p_e 行第 p_s 列元素，b 为一条分支，$s(b)$ 表示分支 b 所在的搜索网络状态，$W_e(p_e)$表示以音素 p_e 结尾的词尾状态集，$W_s(p_s)$表示以 p_s 开头的词首状态集，$next(b)$ 表示分支所搭配的后继词首音素，$prev(b)$ 表示分支所搭配的前驱词尾音素。

2.3 双向搜索

在经典的搜索算法中路径分支是从前往后推进扩展的，根据动态规划原理只保存每个状态上的最优路径。当路径到达词尾时，为结束旧词开始下一个词，通过终点回到起点的路径会和之前的路径相互竞争，剪去了局部较弱的一支路径，带来剪枝风险。而且，在跨词搜索中，由于分支路径数量较大，常需要进行 beam 剪枝，也会带来较大的剪枝风险。

针对上述问题，本文在搜索过程中，除了从前往后的正向解码外，还将搜索网络整体反向，逆向输入语音特征解码，最终在正向路径和反向路径的交汇处求最优路径。这种双向搜索的策略使路径扩展长度减少了一半，从而降低了解码的剪枝风险；此外，当在跨词搜索中使用 beam 剪枝时，一些错误保留的正向路径和错误保留的反向路径也会因最终无法交汇而被剪枝，从而排除了一部分错误搜索路径（结果），进一步降低了剪枝风险。

逆向搜索的 Viterbi 方程类似公式(2),但方向相反：

$$\overline{V}_t^s = \max_{(s',s)\in E}\left\{\overline{V}_{t+1}^{s'} + w(s',s) + \log P(O_t \mid s)\right\} \quad (4)$$

若语音共 T 帧，双向搜索将在 $T/2$ 帧处交汇，求得最佳路径交汇状态 s^*如公式(5)：

$$s^* = \arg\max_s\left(V_{T/2}^s + \overline{V}_{T/2+1}^s\right) \quad (5)$$

3 实验结果与分析

声学特征采用 MFCC 及其一阶、二阶差分共 42 维。训练语音来自 863 数据库，测试语音随机采自 SITEC 数据库，均为朗读语音 16KHz 采样。使用不同关键词表（规模均为 50 个词），共随机选取两组测试集，句子数分别为 544、486。使用上下文相关的 triphone 作为声学基元，每个声学基元含三个输出状态。

通过在搜索网络的词尾加惩罚分方法，可以控制关键词和补白模型的优势差距，从而调整检出率和误警率。图 4 为词内搜索和跨词搜索的 ROC 曲线比较，虚线为跨词搜索的 ROC，实线为词内搜索的 ROC，红色为测试集 1，蓝色为测试集 2。可以看出，在两组测试集上，跨词搜索曲线均远高于词内搜索，更充分利用模型使性能得到很大的提升。

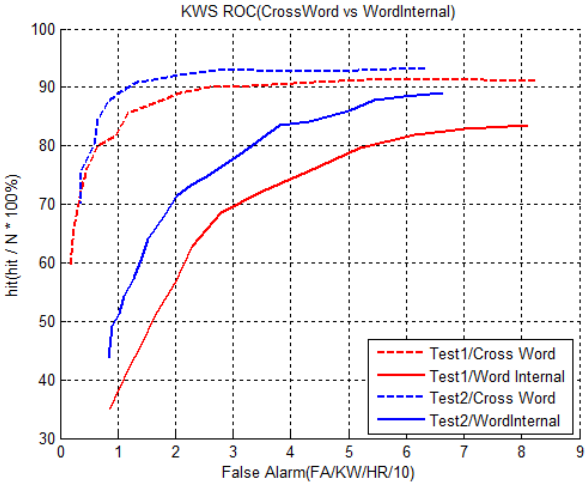

图 4　跨词搜索与词内搜索

图 5 为词内搜索的正向搜索和双向搜索的比较，实线为正向搜索的 ROC，虚线为双向搜索，测试集 1 的曲线为红色，测试集 2 为蓝色。可以看出，在不剪枝的情况下，双向搜索并没有十分明显优势。

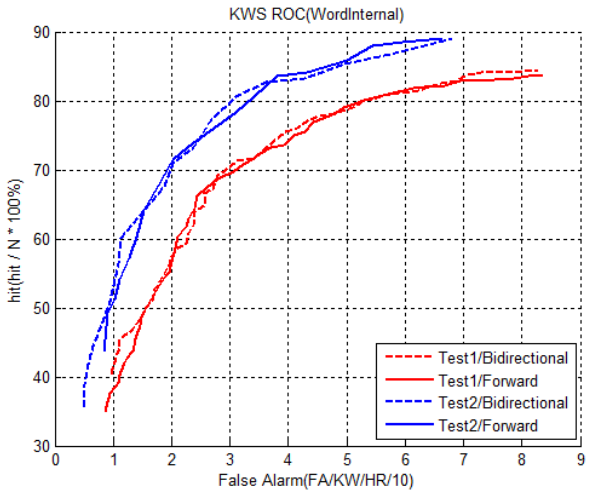

图 5　词内搜索的正向和双向

图 6 为跨词搜索的正向和双向的比较，实线为正向搜索，虚线为双向搜索，红色为测试集 1，蓝色为测试集 2。可以看出，在有剪枝的跨词搜索中，双向搜索相对于正向搜索有明显的性能提升，整体 ROC 曲线位于正向搜索之上，操作点也更靠左上，即拥有更低误警率，更高检出率。

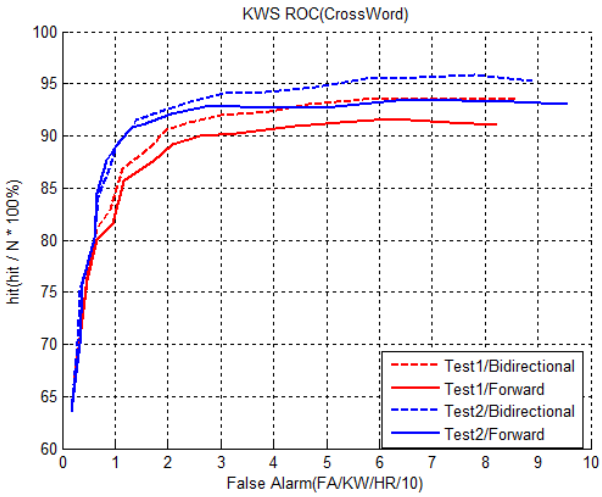

图 6　跨词搜索的正向和双向

本文比较了词内搜索和跨词的正双向共四种算法 ROC 曲线的 FOM，在两组测试集上得到结果见表 1。在两个测试集上，跨词搜索的 FOM 在正向搜索时比词内搜索分别相对提高 26.4%和 17.4%，双向时在两个测试集上分别相对提高 27.8%和 18.4%。词内搜索时，双向比正向在两个测试集上分别相对提高 0.7%和 0.9%，跨词搜索时,双向比正向在两个测试集上分别相对提高 1.9%和 1.7%。

表 1　测试集的 FOM

测试集	词内搜索(正向/双向)	跨词搜索(正向/双向)
1	69.47% / 69.99%	87.78% / 89.45%
2	77.23% / 77.89%	90.67% / 92.20%

4　结论与展望

本文针对词内搜索网络未能充分利用 tri-phone 模型，以及跨词搜索网络过于复杂等问题，提出了一种在词内搜索网络上进行跨词搜索的算法。在此基础上，提出通过双向搜索的办法降低剪枝的风险。

测试集上的对比实验表明，跨词搜索算法相较于词内搜索性能提升明显，说明这种搜索方法充分利用了 tri-phone 模型。双向搜索策略应用于有剪枝的跨词搜索时性能有一定的提升，而对于无剪枝的词内搜索则性能提升有限，可见双向搜索策略能在一定程度上降低剪枝的风险，从而提升系统的性能。

本文在解码的搜索算法和策略上进行研究，今后将尝试在搜索网络的结构以及惩罚分，后端的置信度确认进行一些研究工作。

致谢

本项目工作受到国家自然基金面上项目（No. 61171116）、国家 973 计划（No. 2012C316401）的支持。

参 考 文 献

[1] P. Motlicek, F. Valente, I. Szoke. Improving acoustic based keyword spotting using LVCSR lattices [A]. Proceedings of ICASSP 2012 [C]. Kyoto, Japan: IEEE Press, 2012. 4413-4416.

[2] Wilpon,J.G., Lee,C.H., Rabiner,L .R. Application of Hidden Markov Models for Recognition of a Limited Set of Words in Unconstrained Speech. [A]. Proceedings of ICASSP 1989 [C]. Glasgow, UK:IEEE Press, 1989. 254-257.

[3] R.C Rose and D.B Paul. A hidden Markov model based keyword recognition system [A]. Proceedings of ICASSP 1990 [C]. Albuquerque, NM:IEEE Press, 1990.129-132.

[4] I. Szoke, P. Schwarz, L. Burget,M. Karafiat, P. Matejka, J. Cernocky. Phoneme Based Acoustics Keyword Spotting in Informal Continuous Speech. [A]. in Proc. of TSD 2005 [C]. Berlin : LNCS/LNAI series, Springer-Verlag, 2005.

[5] JE Liang, M Meng, XR Wang, P Ding, B Xu. An Improved Mandarin Keyword Spotting System Using MCE Training and Context-Enhanced Verification. [A]. Proceedings of ICASSP 2006 [C]. Toulouse : IEEE Press, 2006.

[6] Pengyuan Zhang, Jian Shao, Jiang Han, Zhaojie Liu, Yonghong Yan. Keyword Spotting Based on Phoneme Confusion Matrix. [A]. International Symposium on Chinese Spoken Language Processing(ISCSLP 2006) [C]. Kent Ridge, Singapore : 2006.

[7] AUC Optimization Based Confidence Measure for Keyword Spotting. H Li, J Han, T Zheng. [A]. Interspeech 2012 [C]. Portland, USA: 2012.

[8] M. Weintraub. LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting. [A]. Proceedings of ICASSP 1995 [C]. Detroit : IEEE Press, 1995. 297-300

[9] Peng Yu, Seide, F, Jia Liu. A study of lattice-based spoken term detection for Chinese spontaneous speech. [A]. ASRU 2007 [C]. Kyoto, Japan : IEEE Press, 2007. 635-640

[10] Demuynck, K., Duchateau, J., Van Compernolle, D. A static lexicon network representation for cross-word context dependent phones. [A]. In: Proceedings EUROSPEECH [C]. Rhodes, Greece : 1997

Bidirectional Cross-Word Decoder for Keyword Spotting

Yuchen Liu[1], Mingxing Xu[1]

1. Key Laboratory of Pervasive Computing, Ministry of Education

Tsinghua National Laboratory for Information Science and Technology(TNList)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract: Higher real-time and the flexibility of word list are the advantages of the keyword spotting based on sub-word acoustic model. In decoding process, the use of context dependency sub-word acoustic model is not enough on the expanded word-internal search network, and the cross-word search network is hard to expand and will produce tremendous scale and large complexities after expanding. This article perform a cross-word decoding on the word-internal search-net, and try to decrease the beam-pruning risk by a bidirectional search, which search forward and backward simultaneously to meet in center. We do some experiments on a reading speech test set. The experiments show that the cross-word decoding algorithm has large advantages than word-internal decoder that get 27.8% and 18.4% relative improvement on two sets of test data respectively. The bidirectional search makes some improvements on the cross-word decoder with beam-pruning that get 1.9% and 1.7% relative improvement respectively.

Key words: Keyword Spotting; Decoder Algorithm; Cross-Word Search; Bidirectional Search; Speech Recognition