**文章编号**：1003-0077(2020)01-0051-07

# 蒙古语长音频语音文本自动对齐的研究

牛米佳，飞 龙，高光来

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

**摘 要**：目前，面向蒙古语的语音识别语音库资源相对稀缺，但存在较多的电视剧、广播等蒙古语音频和对应的文本。该文提出基于语音识别的蒙古语长音频语音文本自动对齐方法，实现蒙古语电视剧语音的自动标注，扩充了蒙古语语音库。在前端处理阶段，使用基于高斯混合模型的语音端点检测技术筛选并删除噪音段；在语音识别阶段，构建基于前向型序列记忆网络的蒙古语声学模型；最后基于向量空间模型，将语音识别得到的假设序列和参考音素序列进行句子级别的动态时间归整算法匹配。实验结果表明，与基于 Needleman-Wunsch 算法的语音对齐比较，该文提出的蒙古语长音频语音文本自动对齐方法的对齐正确率提升了 31.09%。

**关键词**：蒙古语；语音端点检测；语音文本对齐；动态时间归整算法

**中图分类号**：TP391 **文献标识码**：A

## Research on Automatic Speech-Text Alignment for Long Audio of Mongolian

NIU Mijia, FEI Long, GAO Guanglai

(School of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

**Abstract**: At present, the resources of speech database for Mongolian speech recognition are relatively scarce. To automatically annotate the existing Mongolian audios and corresponding texts, such as TV plays and broadcasts, this paper presents an automatic speech-text alignment method for long Mongolian audio so as to expand Mongolian speech database. In the front-end processing stage, noise segments are filtered and deleted by using Voice Activity Detection technology based on Gaussian Mixture Model. In the speech recognition, the Mongolian Acoustic Model based on Feedforward Sequential Memory Networks is constructed. Finally, based on the Vector Space Model, the hypothesis sequence obtained from speech recognition and the reference phone sequence are matched by the sentence-level Dynamic Time Warping algorithm. The experiments show that the automatic speech-text alignment for Mongolian long audio is improved by 31.09% compared with the traditional Needleman-Wunsch algorithm.

**Keywords**: Mongolian language; voice activity detection; speech-text alignment; dynamic time warping algorithm

## 0 引言

据统计，全世界大约有 1 000 多万人在使用蒙古语，其中以蒙古国，中国八省区，俄罗斯部分地区(Buryat, Khalmyc)使用人数占比最大[1]。然而，相对于中文和英文来讲，蒙古语语音识别和蒙古语语音合成仍有待进一步研究。

蒙古语是一种低资源语言，现有的蒙古语语音语料库规模小、覆盖面窄。为了后续的蒙古语语音识别研究工作的进行，扩充蒙古语语音语料库也成为了首要任务。

语音文本自动对齐技术(简称文语对齐)是语音信号处理中的一个重要的研究方向，其主要任务是将语音和相应的参考文本进行对齐，获得语音与参考文本之间的时间对应关系。如今网络上有越来越多的蒙古语视听资源供研究者使用，利用这些互联网语音资源构建蒙古语语音语料库，可弥补蒙古语语音语料稀少的缺点。同时蒙古语文语对齐技术又能够替代人工标注的工作，自动将语音和文本在时间序列上进行匹配，这也减轻了构建语音库的人力和资金负担。

**收稿日期**：2019-08-15 **定稿日期**：2019-09-29

**基金项目**：国家自然科学基金（61563040，61773224）；内蒙古自然科学基金(2018MS06006，2016ZD06)

随着深度神经网络技术和语音识别技术的深入发展,英语、法语、西班牙语等语音对齐技术也被广泛研究。目前蒙古语语音和文本自动对齐的方法还鲜有研究。

Michael 等[2]采用传统的维特比算法[3]将英语语音和文本对齐。Michael 实验得到的结果较理想,但是他的实验测试集是 Librispeech 等[4]英语语料库,该语料库的每条语音的持续时间都很短(小于 3 分钟)。然而传统的维特比对齐算法是基于隐马尔科夫模型(HMM)的强制对齐算法,当实验数据是长音频时,此算法会产生大量的搜索树,对齐效果不稳定。

Fabrice 等[5]和 Stan 等[6]提出了使用基于语音合成方法将法语的语音和文本对齐。首先将文本合成为语音,再分别提取合成语音和原始语音的声学特征(mfcc、pncc 等),最后将原始长语音和合成语音的 mfcc 特征对齐,进而获得语音和文本的对齐。这个方法的缺点是,一旦实验数据是质量较差的语音和文本,mfcc 特征无法准确地代表语音的内容,对齐的效果会急速下降。

本文选择了更容易获得的蒙古语电视剧音频及其对应的剧本文本作为蒙古语文语对齐研究的输入。但是电视剧的音频和其剧本文本不可避免地存在比例不一致的误差,即语音和文本的内容并不是一一对应的。更为困难的是,电视剧语音中还包括噪声、蒙古语方言、蒙古语口头语,甚至几个人同时发言。这些问题对蒙古语语音对齐任务来说也是个巨大的挑战。如何对齐质量差的语音和文本也是本文的重点研究内容。

本文深入研究了蒙古语长音频的端点检测技术及去噪技术。同时根据蒙古语语音和蒙古文文本特点,提出蒙古语语音文本自动对齐方法,并搭建了一个性能优良的蒙古语长音频的文语对齐系统。第 1 节主要介绍蒙古语长音频文语对齐方法的基本框架;第 2 节研究了蒙古语长音频去噪方法及原理;第 3 节对语音识别声学模型的结构和相关理论进行描述;第 4 节介绍改进的基于 VSM 蒙古文句子序列的对齐算法;在第 5 节中搭建蒙古语长音频语音文本自动对齐系统,并分析了实验结果。在最后一节中,对本论文做了一个总结,同时也陈述了对未来研究的展望。

## 1 蒙古语长音频文语对齐系统框架

蒙古语长音频文语对齐系统的架构如图 1 所示。

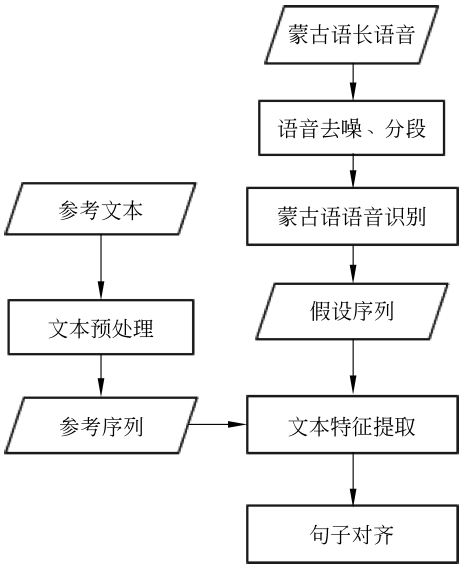


图 1 蒙古语长音频文语对齐系统框架图

首先,根据语音和噪声的频率谱密度特征的差异,采用基于高斯模型的语音端点检测技术将长音频中的噪声段和音频段分开,并标记语音段端点。其次,基于 FSMN 声学模型将语音段逐一识别为蒙古文的拉丁转写形式,得到语音识别的结果序列及该序列对应的时间,此时将结果序列称作假设序列。同时,对参考文本进行拉丁校正处理,得到参考文本序列。最后,基于向量空间模型(VSM)将结果及参考文本的句子提取文本特征,计算假设序列和参考文本序列的相似度,由动态时间归整算法(DTW)得到对齐片段集合。

## 2 基于高斯混合模型的语音端点检测及去噪

由于需要对齐的连续语音序列较长,而目前语音识别引擎对语音时间长度有一定的限制,所以在文语对齐前,必须先将输入的连续语音流划分为较小的片段。同时实验的音频数据中包含了大量的白噪声,而无关的非语音段又影响着语音对齐的效果,所以需要用语音端点检测技术筛选出噪声段并将其删除。

语音端点检测[7](voice activity detection,VAD)是指在嘈杂背景环境下的语音信号流中分离出语音信号和非语音信号,并确定语音信号的起始端点和终止端点。蒙古族在日常生活中讲话语速较快,连读时语句之间停顿并不明显。在有噪声的情

况下，语音段的端点就更难以检测出来。这是研究蒙古语端点检测时遇到的难题之一。本文根据蒙古语语音特点和含噪情况，分析噪声和语音二者的频率谱密度差异，采用了基于二维高斯模型（gaussian mixture model，GMM）的语音端点检测技术分离噪声和语音。即对输入的每帧语音信号进行高斯建模后，分别计算其是语音和噪声的概率，然后再计算对数似然比统计量（LR）来判断是否有语音。

首先将原音频的语音信号降低采样至 8kHz，并利用傅里叶变换将频谱划分成六个子带。根据奈奎斯特频率定理[8]计算得出有信息价值的语音频谱在 4kHz 以下，故划定六个子带的能量范围 80～250Hz，250～500Hz，500Hz～1kHz，1～2kHz，2～3KHz，3～4kHz，同时使用分频方法计算出每个频带能量的特征向量序列。对每一个子带能量特征向量进行建模，每个模型混合了语音和噪声两个高斯分布，如式（1）所示。

$$f(x \mid Z,r)=\frac{1}{\sqrt{2\times\pi\times\theta^2}}\times e^{\frac{-(x-u)2}{2\times\theta2}} \tag{1}$$

其中，$x$ 是选取的六个子带能量的特征向量序列；$r$ 代表模型参数 $u$ 和 $\theta$ 的集合；$u$ 是输入信号的均值，$\theta$ 是输入信号的方差，这两个参数决定了每帧语音高斯分布的概率值。式（1）中，若参数 $Z$ 为 0，则代表计算噪声概率；$Z$ 为 1 则代表计算语音概率。

对子带的每个特征求对数似然比，再求得全局加权似然比之和，如式（2）所示。其中 $K_i$ 是似然比的加权系数。设局部阈值 $T_\tau$ 和全局阈值 $T_a$，若六个子带特征中有一个似然比超过 $T_\tau$，则认为有语音；若六个子带加权似然比之和超过了 $T_a$，则也认为有语音，如式（3）所示。

$$\begin{aligned}L(x(n))&=\sum K_i L(x(n),i)\\&=\sum K_i\log\left(\frac{f_s(x(n,i))}{f_n(x(n),i)}\right)\end{aligned} \tag{2}$$

$$F_{\mathrm{vad}(n)}\begin{cases}1 & L(x)>T_a \parallel L_i>T_\tau\\ 0 & \text{else}\end{cases} \tag{3}$$

最后自适应更新高斯混合模型的参数，使用极大似然估计根据已经分类完的数据重新计算语音和噪声的均值和方差。对下一帧语音重新循环整个过程，标记出语音端点和噪声端点，并删除噪声点之间的片段。

## 3　基于前向序列记忆神经网络模型的蒙古语语音识别

本文将前向序列记忆神经网络结构（FSMN）[9-10]用于构建蒙古语语音识别的声学模型。

前向序列记忆神经网络（FSMN）是一种非递归的前馈型神经网络。相较于传统的神经网络模型，FSMN 同样也包含了多个隐藏层，但其不同之处在于它每个隐藏层旁增加了一个“记忆块”。这些记忆块类似于语音信号处理中的 FIR 滤波器，因为其保存着隐藏层中与当前语音帧相关的历史帧信息和未来帧信息，从而使得模型能够学习到语音信号的长时相关性信息。图 2 为隐藏层中包含了两个记忆块的 FSMN 神经网络模型的结构。

对于语音帧的特征序列 $X=\{x_1;x_2;\cdots;x_t\}$，记忆块被定义为式（4）。

$$\tilde{h}_t^l=\sum_{i=0}^{N1}a_{t,i}^l\odot h_{t-i}^l+\sum_{j=1}^{N2}a_{t,M1-1+j}^l\odot h_{t+j}^l \tag{4}$$

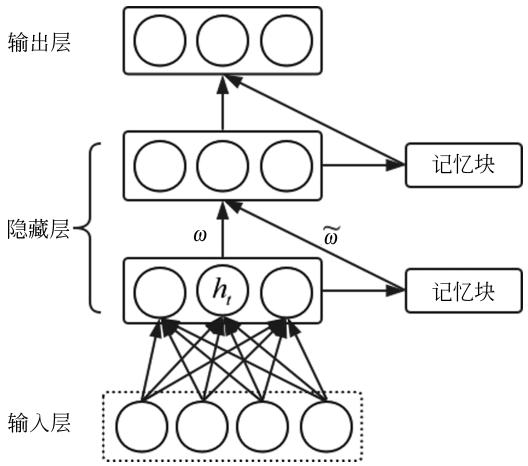


图 2　FSMN 网络结构

其中，$h_t^l$ 为当前的输出；$N_1$ 为回溯的历史语音帧的数量，$N_2$ 为向前的未来语音帧数量；$A^l=\{a_0^l,a_1^l,\cdots,a_N^l\}$ 为时变的记忆块系数，代表 FSMN 的类型为编码系数可各自独立学习的矢量 FSMN（vFSMN）。

## 4　基于动态时间规整算法的文本句子对齐

由语音端点标记的片段经过蒙古语语音识别得到假设序列，参考文本由文本处理得到参考文本序列。由于蒙古文存在同音异形的现象，音素序列比单词序列能够更准确的代表两个句子的相似程度，所以将参考序列和假设序列转化成音素序列。假设序列和参考文本序列通过动态时间归整算法（DTW）进行校准，得到正确的对齐片段集合。

语音对齐通常是以单词或者句子为单位进行的。单词级别的对齐较简单，时间较快，但对齐准确

率和语料完整程度息息相关。根据文本实验目标、数据集含错等特点，本文采用以句子为单位进行对齐。同时利用向量空间模型(vector-space model，VSM)将语音识别的结果序列和参考文本序列都表示为句向量(sentence vector)。即每个句子包含若干音素 $T_1,T_2,\cdots,T_n$，它可以表示为向量 $D(w_1,w_2,\cdots,w_k)$。

对齐过程以假设序列为标准，顺序搜索参考文本，找到与假设序列最为相似的部分。本文采用余弦相似度衡量参考文本和结果文本中两两句子序列之间的相似度，如式(5)所示。

$$\mathrm{sim}(D_i,D_j)=\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{\sum_{k=1}^{n}w_{ik}^2}\cdot\sqrt{\sum_{k=1}^{n}w_{jk}^2}} \tag{5}$$

其中，$w_k$ 是音素 $T_k$ 在每个句子中的权重，可由词频—逆文档频率(TF-IDF)计算得出，如式(6)所示。$\mathrm{TF}_k$ 为音素 $T_K$ 在句子中出现的次数，$N$ 为参考文本的句子总数，$\mathrm{DF}_K$ 表示参考文本中出现过 $T_K$ 的总个数。

$$w_k=\mathrm{TF}_k\times\log\left(\frac{N}{\mathrm{DF}_K}\right) \tag{6}$$

通过式(5)计算假设序列和参考文本序列之间距离矩阵 $\boldsymbol{S}$(cost matrix)。

动态时间归整算法是一种局部优化算法，它能够找到两个时间相关序列之间的最佳对齐方式。假设序列和参考文本序列在时间轴上进行扭曲变换，使得两者的相似度达到最大值。则动态时间归整算法的任务就是在矩阵 $\boldsymbol{S}$ 上找到一条从(0,0)到$(m,n)$的路径 $L$，使得累加路径的总距离最小。这条路径称为最优对齐路径(optimum warp path)，定义为式(7)，其中 $\mathrm{sim}(D_i,D_j)$为参考文本第 $i$ 句和假设文本第 $j$ 句的相似度。

$$L=\mathrm{argmin}\sum_{(i,j)\in L}\mathrm{sim}(D_i,D_j) \tag{7}$$

动态时间归整算法利用了子问题的独立性，采用分治思想将大问题分割成若干简单的小问题。一般来说，为提高搜索效率，DTW 的路径上数据点范围限制到一个平行四边形内，路径的斜率范围在 0.5 和 2 之间[11]。据此，当前点的前格点仅有三种情况，最优对齐路径可以扩展为式(8)：

$$\gamma(i,j)=\min\begin{cases}\gamma(i-1,j-1)+\eta\mathrm{sim}(D_i,D_j)\\ \gamma(i-2,j-1)+\mathrm{sim}(D_i,D_j)\\ \gamma(i-1,j-2)+\mathrm{sim}(D_i,D_j)\end{cases} \tag{8}$$

其中，$\gamma(i,j)$是到第$(i,j)$点的累计距离长度。参数 $\eta(\eta>1)$为从$(i-1,j-1)$点到$(i,j)$点的距离权重，其目的是控制对角线点到距离成本大于 $D$ 中其他点到$(i,j)$距离成本。实验将得到的 DTW 算法的最优对齐路径如图 3 所示。

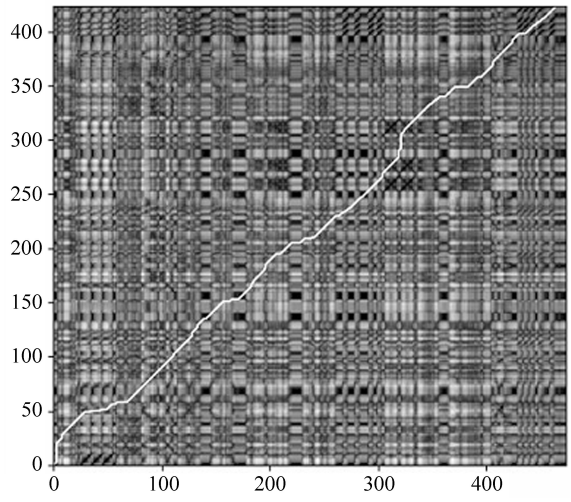


图 3 DTW 最优路径

经过 DTW 算法匹配之后，得到最优路径。在求出距离矩阵中所有元素之后，用逆向搜索法得到最优路径。

## 5 实验

本文以人工对齐的结果作为评判实验性能的标准。实验中采用准确率(accuracy)来评价实验结果，如式(9)所示。

$$\mathrm{accuracy}=\frac{R}{T} \tag{9}$$

其中，$R$ 代表参考文本中被正确对齐的句子数，$T$ 代表参考文本中所有句子的个数。

### 5.1 实验语料

实验的语音数据来源于蒙古语电视剧音频，例如《阿娜尔罕》(蒙古语)、《闯关东》(蒙古语)等 7 部电视剧。其中每个音频都有对应的剧本，将剧本进行预处理，生成参考文本库。

与实验室语音不同，本实验的每个原始音频包含若干说话人，每个说话人的语音自然度较高。并且每个音频中都包含带有丰富情感的非语音片段，像笑、哭、喊叫、叹气等。每个音频时长的范围在 37min 到 46min 之间，总数据集时长约为 5 452min。每个音频的声道均为单声道，采样率为 48 000kHz，每个采样点

进行 16 比特位量化。

经统计，每个电视剧音频相应的蒙古文参考文本大约有 600 句话。参考文本内容和音频的内容差别较大，经人工校正比对得知，参考文本有将近 35.2%以上的文本和音频的语音内容不符。

本实验从电视剧音频中随机抽取出 6 集作为测试集，总计约有 13 542s，并且将测试集的语音和其相应的文本进行人工对齐。

### 5.2　蒙古语长音频文语自动对齐系统的构建

5.2.1　文本处理模块

由于输入的文本语料是蒙古语电视剧的剧本文本，所以实验需要对文本语料进行预处理。

首先，将剧本中每段话的说话人标注删除，并转换文本中的特殊字符。

其次，将文本的蒙古文的编码格式由蒙科立蒙古文编码统一转换为国家标准蒙古文编码格式。

蒙古文文字是一种存在同形异义现象的拼音文字。这导致了人工录制的蒙古文文本的实际意义和文本本身意义存在偏差。若不纠正这些偏差，在后续的对齐过程中，索引空间将扩大，甚至搜索精度也会降低。故在文本处理模块的最后，还需要将蒙古文进行校正，并生成具有唯一编码的蒙古文拉丁形式。

5.2.2　语音端点检测模块

使用基于高斯模型的语音端点检测技术确定噪声片段的分割点。经过实验得知，将局部阈值和全局阈值分别设置为 84 和 960 能够最大限度地删除非语音片段。VAD 切分得到的结果如表 1 所示。

**表 1　VAD 算法的准确率、精确率和召回率**

| | 总时长/s | 准确率/% | 精确率/% | 召回率/% |
|---|---|---|---|---|
| VAD 切分 | 20 580 | 88.6 | 78.1 | 93.9 |

以“人工切分得到的总语音时长”为基准，表 1 准确率显示，VAD 能删除大部分的沉默、叹气、低啜等白噪声片段。表 1 中，精确率代表 VAD 未能筛选出的噪声所占的比例。这部分一般是频谱较高的非语音段，同时它们又和语音段连接得十分紧密。召回率达到 93%以上，表示仅有小部分的语音被误判成噪声，这部分语音一般是频谱较低或者时长较短（小于 0.8s）的语音段。

5.2.3　语音识别模块

构建蒙古语语音识别模块，并将语音端点之间片段分别送入其中。利用 Librosa 库[12]提取 39 维的梅尔倒谱数特征（mfcc），并基于 mfcc 语音特征进行说话人自适应训练。构建基于 Kaldi[13]语音识别框架的 FSMN 的蒙古语声学模型。本文训练了一个含有 6 个隐藏层的 vFSMN 蒙古语声学模型。该模型的每个隐藏层包含 2 048 个神经元节点，激活函数为 ReLu。其中前三个隐藏层带有记忆块，后三个不带记忆块。每个记忆块保存当前帧的前面 5 帧信息和后面 5 帧信息。用 FSMN 声学模型和由参考文本训练得到的 3-ngram 语言模型对语音的特征向量解码，得到假设序列。

5.2.4　文本对齐模块

最后，在 Python 环境下提取文本特征并构建 DTW 算法。本文实验使用了 Praat 工具[14]生成 TextGrid 网格文本，记录时间信息和文本序列。

### 5.3　实验结果及分析

5.3.1　基线系统

本文采用了传统的基于 Needleman-Wunsch 算法语音强制对齐[15]作为本文构建的蒙古语长音频文语对齐系统的对比实验。使用基于隐马尔科夫模型（HMM）的解码器将音频转换成一个音素集合，问题转变成求解最长公共子序列。将参考文本转化并整合成一个音素集合。采用 Needleman-Wunsch 算法[16]将这两个音素集进行强制对齐，并得到对齐结果。

5.3.2　实验结果

蒙古语长音频的语音文本自动对齐系统对齐的结果示例如图 4 所示。

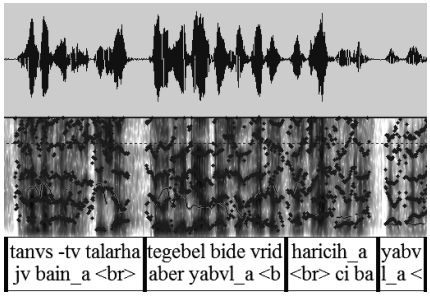


图 4　蒙古语长音频文语对齐系统的对齐结果

图 4 为语音的频谱、共振峰及其对齐结果，其中＜br＞代表参考文本中一个句子的结束。图 4 中四句拉丁文本及其蒙古文分别为：

（1）tanvs -tv talarhajv bain_a ＜br＞

蒙古文为：

（2）tegebel bide vridaber yabvl_a ＜br＞

蒙古文为：

(3) haricih_a <br> ci basa hiciye dE <br> yabvy_a <br>

蒙古文为：

(4) yabvl_a <br>

蒙古文为：

图4中第二段语音“tegebel bide vridaber yabvl _a <br>”还未结束时，第三段语音“haricih_a <br>”就已经开始，这两段语音发生了重叠，导致语音识别出错。第三段中两短句“ci basa hiciye dE <br>”和“yabvy_a <br>”，此处由于后两句语音的停顿不明显，连接较紧密，VAD切分失败。但这两句话的参考文本指向一处语音，对齐仍然成功。图4中其余语音段对齐正确。由于本文实验的目的是扩充蒙古语语音语料库的数据量，所以不必再将蒙古文的拉丁形式转换成传统蒙古文形式。

同时，本文将设置如下对比实验：

**实验一** 基线实验(HMM模型+Needleman-Wunsch)

**实验二** FSMN模型+Needleman-Wunsch

**实验三** VAD+FSMN模型+Needleman-Wunsch

**实验四** VAD+FSMN模型+VSM-DTW

表2中显示4种对齐方法的实验结果。

**表2 基于Needleman-Wunsh对齐算法和蒙古语长音频文语对齐系统结果比较**

| | 实验一 | 实验二 | 实验三 | 实验四 |
|---|---|---|---|---|
| 句子总数 | 3 172 | 3 172 | 3 172 | 3 172 |
| accuracy/% | 55.61 | 63.49 | 73.17 | 86.7 |

从表2中实验一～实验三可以看出，Needleman-Wunsch算法并不适合于错误较多的语音和文本，并且Needleman-Wunsch算法实验花费的时间较长。实验二和实验三证明了VAD去非语音段的必要性。实验三和实验四证明了本文提出的改进的基于VSM的句子DTW对齐方法优于Needleman-Wunsch算法，具有一定的鲁棒性。这是因为基于Needleman-Wunsch语音对齐算法是对整个音频和整个文本进行逐一的音素级别的对齐，它并没有区分单词和句子。该算法的特性决定了一旦某个区域中插入若干单词后，后面的音素对齐也会受到很大的影响。而本文提出的改进的基于VSM的句子DTW方法具有一定的容错性，它保留了音素的特性，并以句子为单位进行DTW，避免了因单词操作(删除、增加、替换)而导致对齐失败的情况。

由表2可知，在本文提出的改进的基于VSM的DTW方法中，有13.30%文本发生强制对齐错误。这是因为切分后的音频仍包含了频谱较高的非语音片段，这部分句子无法被正确识别；亦或由于原文本中某些句子完全错误，导致假设序列和参考文本序列相似度极低，对齐失败。

## 6 总结

本文基于蒙古语语音特点和蒙古文文字特点，提出了蒙古语长音频语音文本自动对齐的方法。并且研究了蒙古语长音频的语音端点检测技术及去噪技术，提出基于VSM序列串动态时间规划的对齐方法。该方法将会有效地扩充蒙古语语音语料库的数据量。同时本文还基于Needleman-Wunsch强制对齐算法设计了4个对比实验。实验结果显示，本文构建的系统具有一定的稳定性，对质量不好的语音和文本也能有86.7%的句子匹配成功。

未来的工作将重点研究基于深度神经网络的蒙古语语音端点检测技术及去噪技术，从而提高对齐的正确率。并且将说话人识别加入到蒙古语长音频文语对齐算法中，根据文本中的说话人标签进行辅助匹配。进一步研究基于深度学习的音素级别的对齐，让计算机替代在构建语料库上所付出的人为劳动。

## 参考文献

[1] Nnesbeitt S L. Ethnologue: Languages of the world [J]. Electronic Resources Review, 1999, 3(11): 129-131.

[2] Michael M, Michaela S, Sarah M, et al. Montreal forced aligner: Trainable text-speech alignment using Kaldi[C]//Proceedings of the Interspeech, 2017: 498-502.

[3] Jr G D F. The Viterbi algorithm[J]. Proceedings of the IEEE, 1973, 61(3): 268-278.

[4] Panayotov V, Chen G, Povey D, et al. Librispeech: An ASR corpus based on public domain audio books [C]//Proceedings of the ICASSP, 2015, 5206-5210.

[5] Fabrice M, Thierry D. High-quality speech synthesis for phonetic speech segmentation[C]//Proceedings of the 5th European Conference on Speech Communication and Technology, EURO SPEECH, 1997: 22-25.

[6] Stan A, Bell P, King S. A grapheme-based method for automatic alignment of speech and text data[C]//Proceedings of the Spoken Language Technology Workshop (SLT). 2012: 286-290.

[7] 韩立华，王博，段淑凤，等. 语音端点检测技术研究进展[J]. 计算机应用研究，2010，27(4)：1220-1226.

[8] 李冬梅，李小静，梁圣法，等. 一种以低于奈奎斯特频率的采样频率进行信号采集方法：中国，CN201310220522.8[P].2013-10-09.

[9] Zhang S, Ming L, Yan Z, et al. Deep-FSMN for large vocabulary continuous speech recognition[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5869-5873.

[10] Wang Y, Bao F, Zhang H, et al. Research on Mongolian speech recognition based on FSMN[C]//Proceedings of the NLPCC 2017: Natural Language Processing and Chinese Computing, 2017: 243-254.

[11] Chen L, Research on DTW algorithm improvement technology based on speech recognition system[J]. Microcomputer Information, 2006, 22(2): 267-269.

[12] Brian M, Colin R, Dawen L, et al. Librosa: Audio and music signal analysis in Python[C]//Proceedings of the Python in Science Conference, 2015: 18-24.

[13] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]//Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011: 1-4.

[14] Boersma P D. Praat: Doing phonetics by computer [J]. Ear & Hearing, 2011, 32(2): 266.

[15] Souza G, Neto N. An automatic phonetic aligner for Brazilian portuguese with a Praat interface[C]//Proceedings of the PROPOR 2016: Computational Processing of the Portuguese Language, 2016: 374-384.

[16] Rose J, Eisenmenger F. A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm[J]. Journal of Molecular Evolution, 1991, 32(4): 340-54.

牛米佳(1996—)，硕士研究生，主要研究领域为语音对齐、蒙古文信息处理。
E-mail: nmj5500@163.com

高光来(1964—)，硕士，教授，博士生导师，主要研究领域为人工智能、模式识别、自然语言处理。
E-mail: csggl@imu.edu.cn

飞龙(1985—)，通信作者，博士，副教授，硕士生导师，主要研究领域为蒙古文信息处理、语音合成、语音检索、语义理解、信息检索、机器翻译。
E-mail: csfeilong@imu.edu.cn