ISSN 1000-0054 清华大学学报（自然科学版）2013 年 第 53 卷 第 6 期 36/40
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2013, Vol. 53, No. 6 903-907

基于分段动态时间规整和后验特征的中文语音模式发现

杨 鹏， 谢 磊， 陈虹洁

（西北工业大学 计算机学院，陕西省语音与图像信息处理重点实验室，西安 710129）

摘 要：语音模式发现是从语音流中检测出重复出现的音节、词或短语等语音单元的任务。该文基于分段动态时间规整（segmental dynamic time warping，SDTW）算法，尝试直接在中文语料上进行语音模式发现。Mel 频率倒谱系数（Mel frequency cepstral coefficient，MFCC）特征在衡量两个语音片段声学相似度上不够鲁棒，特别是针对多说话人语料，语音模式发现的效果大打折扣。该文尝试了基于音素后验概率（posteriorgram）的特征表示方法。实验表明：在多说话人和单说话人的语料上，音素后验特征均可以得到比 MFCC 更好的效果。该文尝试了用词边界确定分段进行语音模式发现，这种设置可以看作基于 SDTW 进行模式发现的效果上限。实验表明：在预知词边界的情况下，效率和正确率都得到了明显提升。

关键词：语音模式发现；后验特征；动态时间规整；分段动态时间规整

中图分类号：TP 391 **文献标志码**：A

文章编号：1000-0054(2013)06-0903-05

Mandarin speech pattern discovery using segmental dynamic time warping and posteriorgram features

YANG Peng, XIE Lei, CHEN Hongjie

(Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Speech pattern discovery aims to identify repeated patterns (e.g., word-like units) from speech. This study analyzes speech patterns in a Mandarin speech corpus using segmental dynamic time warping (SDTW). Mel frequency cepstral coefficients (MFCCs) have not been effective for pattern discovery in multi-speaker conditions. The phoneme posteriorgram features are used here in a template-based method. Tests show that phoneme posteriorgram is significantly better than MFCCs for both single- and multi-speaker conditions. The performance upper-bound of SDTW is also investigated when boundary information is available with the segments divided by word boundaries. The results show that the boundaries significantly improve the pattern discovery in terms of both accuracy and efficiency.

Key words: speech pattern discovery; posteriorgram; dynamic time warping (DTW); segmental dynamic time warping (SDTW)

语音模式（speech pattern）是指语音流中重复出现的音节、词或短语。这些重复出现的单元聚集成一个个“语音模式”。心理学家的研究表明，辨识重复出现的语音模式对于婴儿的词汇习得至关重要[1]。检测词条（term）的重复出现也是主题分析、语音文档检索、语音关键词检出、语音概要生成等众多语音与语言处理任务中的一项基础工作。以语音文档检索为例，建立索引就是统计每个词条（词或子词单元）出现频率的过程。而在主题分割任务中，一些关键词（如人名、地名专有名词）的重复出现是主题粘合（cohesion）的重要线索。通常，检测词条重复的任务是在语音识别抄本或网格（lattice）上完成的。然而，建立大词汇量连续语音识别器需要大量的标注数据、工具与资源。词典未登陆词汇（OOV）问题时常出现，识别错误无法避免。以语音文档检索为例，一些和主题密切相关的实体词往往是 OOV 词汇，这些词汇的错误识别严重影响检索效果。基于子词或识别网格的检索方法虽然能够一定程度上缓解这一问题，但仍然需要搭建繁杂的语音识别系统。对于缺乏语料资源的小语种，训练一个性能优良的大词汇量连续语音识别器几乎是不可能的。近年来，基于少资源的、非监督的语音模式发现方法，开始引起广泛的研究兴趣[2-8]。此类方法无需建立繁杂的语音识别系统，直接在语音信号上进行模式

收稿日期：2013-04-27

基金项目：国家自然科学基金项目（61175018）；
陕西省自然科学研究计划（2011JM8009）；
霍英东基金项目（131059）

作者简介：杨鹏（1989—），男（汉），陕西。

通信作者：谢磊，教授，E-mail：lxie@nwpu.edu.cn

匹配,检测出词条重复,进而聚类成"语音模式",用于文档分类、主题分割、关键词检出等任务。

当前,在无监督语音模式发现任务中,词条重复检测大多采用动态时间规整算法(dynamic time warping, DTW)的变种。例如,美国麻省理工大学(MIT)的 Glass 等使用分段动态时间规整算法(segmental DTW, SDTW)[2-3]在英文语料上检测词条重复,进而利用图结构,将重复词条聚类成"语音模式"簇。Dredze 等使用了一种基于图像处理的 DTW 算法[7],检测出词条重复后,使用类似的聚类算法,生成"伪词条"单元,用于文档聚类和分类等任务。Anguera 等提出了一种 Unbounded DTW 算法[6],在词条重复检测上,能够达到与 SDTW 相当的效果。这类 DTW 变种算法的核心思想是采用模板匹配方法在两个连续的语音流中发现发音相似的片段。

本文基于 SDTW 算法,尝试在中文语料上进行语音模式发现,以验证该算法在中文语音模式发现上的有效性。本文的工作主要体现在两个方面:1) Mel 频率倒谱系数(Mel frequency cepstral coefficient, MFCC)特征在衡量两个语音片段声学相似度上不够鲁棒,特别是针对多说话人语料,语音模式发现的效果大打折扣。为此,从近期基于模板匹配的语音识别[9]和基于 DTW 的关键词检出[10]工作中得到启发,本文尝试在语音模式发现任务上使用音素后验概率特征。实验表明,音素后验特征表现出较好的鲁棒性,在多说话人和单说话人的语料上,该特征都可以得到比 MFCC 更好的效果。2) SDTW 算法靠人工设定的滑动窗在连续的语流上进行分段匹配,衡量两个语音片段相似度的方法并非标准的 DTW 算法,因此匹配效果受限,且计算复杂度大。因此,本文尝试了用词边界来分段进行语音模式发现的效果,期望降低计算代价,同时提高正确率。这种设置可以看作基于 DTW 算法进行模式发现的效果上限。实验表明,在预知词边界的情况下,效率和正确率都得到了明显提升。

1 后验特征

后验特征,即给定一个语音特征向量 $\boldsymbol{o}$,该特征向量在预先定义好的 k 个类$\{C_1, C_2, \cdots, C_k\}$上的后验概率分布 $\mathrm{PG}_{\boldsymbol{o}}$为

$$\mathrm{PG}_{\boldsymbol{o}} = (p(C_1 \mid \boldsymbol{o}), \cdots, p(C_k \mid \boldsymbol{o})). \quad (1)$$

其中,$P(C_i \mid \boldsymbol{o})$是特征向量 $\boldsymbol{o}$ 在第 i 个类上的后验概率。这里的类可以定义为任何种类的语音单元,比如音素。本文使用的就是音素级的后验特征。和传统的 MFCC、感知线性预测(perceptual linear prediction, PLP)等特征相比,后验特征在表征声学相似性上更具鲁棒性。基于模板匹配的语音识别研究表明,后验特征明显优于传统的声学特征[10]。

本文使用了 Brno 大学开发的音素识别器 BUT[11]来提取音素后验特征。该识别器基于多层感知器 MLP,使用长时上下文特征 TRAP 和 Viterbi 解码算法。本文使用约 10 h 左右的中文广播语料,训练出中文音素识别器。MLP 的训练参数设定为:隐节点数 1 000,音素个数为 67(不带声调的中文声韵母集合),音素状态数为 3。在这 10 h 数据上进行交叉检验获得的音素识别错误率为 20.3%。利用训练好的识别器,给定一个语音文件,得到以帧为单位的后验特征。

图 1 所示为语音片段"美国总统克林顿"的后验特征向量序列随时间的变化。本文采用 25 ms 帧长,10 ms 帧移的设置。纵轴代表每帧数据在各个音素上的得分,共 67 维,即式(1)。在每维上的取值就是该语音帧在相应音素上的后验概率,越接近 1 越黑,越接近 0 越白。

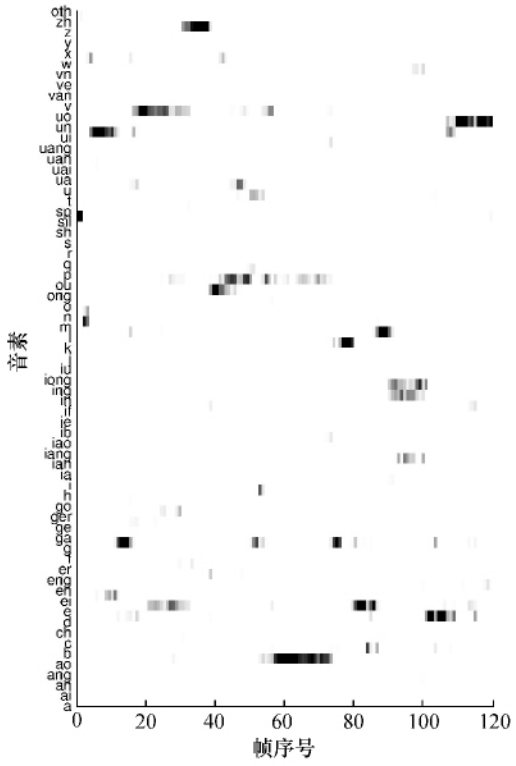

图 1 某语音片段的后验特征向量序列随时间变化图

2　词条重复检测

本文中使用 SDTW 算法[2-3]进行词条重复检测。该算法是 DTW 算法的一种变体。

2.1　DTW 算法

DTW 算法起初被成功用于孤立词语音识别。给定两个连续语音片段的特征向量序列，$\boldsymbol{u}_x=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$，$\boldsymbol{u}_y=\{y_1,\cdots,y_m\}$，$n$ 和 m 分别表示两个语音片段的特征向量的帧数。通过定义语音帧的特征向量之间的距离，建立一个距离矩阵 $\boldsymbol{D}$。用 ϕ 表示 $\boldsymbol{u}_x$ 与 $\boldsymbol{u}_y$ 之间的一种可能的对应关系，$\phi(k)=(i_k,j_k)$，$k=1,\cdots,T$。该算法就是要在距离矩阵 $\boldsymbol{D}$ 中找出一个最优对应序列 ϕ'，从而最小化积累失真值 $\mathrm{Dist}_\phi(u_x,u_y)$，

$$\mathrm{Dist}_\phi(u_x,u_y)=\sum_{k=1}^{T}D(i_k,j_k). \tag{2}$$

其中，$D(i_k,j_k)$表示特征向量 $\boldsymbol{x}_{i_k}$，$\boldsymbol{y}_{j_k}$ 之间的距离。如果语音帧用 MFCC 特征，通常使用欧式距离度量，

$$D(i_k,j_k)=\|\boldsymbol{x}_{i_k}-\boldsymbol{y}_{j_k}\|. \tag{3}$$

如果语音帧用后验特征表示，则使用负对数内积度量[8]，

$$D(i_k,j_k)=-\lg(x_{i_k}\cdot y_{j_k}). \tag{4}$$

2.2　SDTW 算法

DTW 算法一般只适用于比较的两个语音片段是孤立词的情形。词条重复检测任务是从两个较长语音片段中发现其中重复的“子片段”，例如词或短语。因此，直接用该算法得出的打分显然没有意义。此任务可由 SDTW 算法完成。

SDTW 算法的思想是，将距离矩阵 $\boldsymbol{D}$ 划分成子带，在子带中用传统的 DTW 搜索最优路径。首先将距离矩阵 $\boldsymbol{D}$ 划分为若干等宽的斜带状区域。在实际应用中，考虑到两个语音短语中发声相似部分对应的区域可能刚好在某一条带状区域的边界上，因此设定相邻带状区域之间有 50％的重叠，如图 2 所示。其中，s_1 和 s_2 分别是第 1 个带状区域和第 2 个带状区域的起始点。设从 s_1 到 s_2 的位移为 R，斜带状区域宽度则为 $2R+1$，而对于一个 $n\times m$ 的矩阵，其包含的带状区域的数量就是$\lfloor(n-1)/R+(m-1)/R\rfloor$。

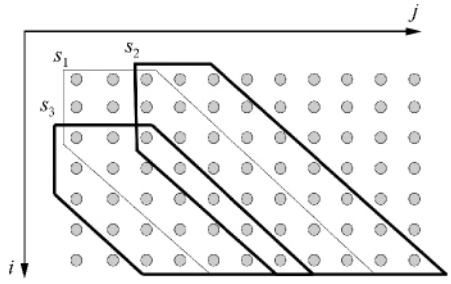

图 2　距离矩阵 $\boldsymbol{D}$ 中的带状区域示意图(Zhang 等[3])

然后，在每个斜带状区域里使用 DTW 算法找出一条最优路径。而在每一条最优路径中，对应两个连续语音片段声学相似部分的往往只有路径上的一小段。因此，需要将最优路径上特定的子路径切割出来，这些子路径满足的条件是：1）子路径包含的点的数量，即子路径长度，长于 L；2）子路径包含的所有点的平均值，即子路径平均值小于 θ。

给定一条包含 N 个点的路径以及子路径长度限制 $L(L<N)$、子路径平均值限制 θ，利用路径切分算法切割出这条路径上的 LCMA(length-constrained minimum average)子路径[12]，

$$f=\min_{1\leqslant s\leqslant t\leqslant N}\frac{1}{t-s+1}\sum_{k=s}^{t}D(i_k,j_k),\quad t-s+1\geqslant L. \tag{5}$$

如式(5)所示，LCMA 子路径就是这条路径上长度大于 L，并且有最小平均值 f 的子路径。而根据上面的两个限定，如果 LCMA 子路径的 f 值大于 θ，则被丢弃，若小于 θ，则被保留，即该路径对应了一对词重复。然后，在该路径上剔除这个 LCMA 子路径，继续寻找 LCMA 子路径，直到找出这条路径上所有的长度大于 L 且平均值小于 θ 的子路径。

图 3 是一个使用 SDTW 算法查找 LCMA 子路径的示例。在这个例子中，作比较的两个连续语音片段来自同一个说话人，语音帧用 MFCC 特征向量表示，算法中的参数设置为 $R=10$，$L=50$，$\theta=7.6$。可以看到，作比较的两个语音短语有两处明显的词重复，分别是“总统”和“克林顿”。本文使用的算法成功找到了这两处词重复，即用粗线标注的片段，片段旁的数字为该片段的平均值。例如，有最小平均值 7.4 的片段对应着词重复“总统”，同样对应“总统”的还有平均值为 7.5 的两个片段。

另外，可以从图 3 中发现，对应两个词重复的片段数目分别为 2 个和 3 个，这是因为设置了 50％的带状区域重叠，使得每一个词重复对应的区域都被至少 2 个带状区域覆盖。也就是说，如果两个语音短语之间存在词重复，那么词重复所对应区域内应该有至少 2 个 LCMA 片段能被找到。因此，在使用 SDTW 算法找到所有的 LCMA 片段以后，舍弃那些孤立存在于某一个区域的 LCMA 片段。

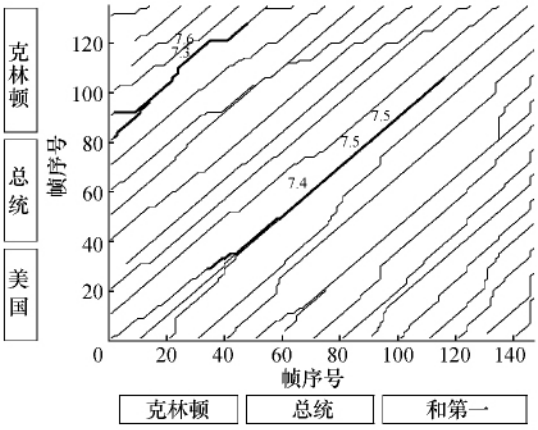

图 3 利用 SDTW 查找 LCMA 子路径及其对应的 f 值

2.3 SDTW 算法在后验特征上的效果分析

为了直观地观察 MFCC 特征和音素后验特征表示的数据在词重复检测上的不同之处,本文对图 2 中的两个语音片段,分别使用 MFCC 特征和后验特征表示,利用 SDTW 算法,进行了 LCMA 子路径的寻找。不同的是,SDTW 的算法参数设置是 $R=10$, $L=50$, $\theta=\infty$, 这里将 θ 设为无穷大的目的是为了让所有 LCMA 子路径都保留下来,以便于观察。

从图 4 可以看出,使用后验特征的 Dot-Plot 图中,词重复区域有很明显的黑色斜带,而非词重复区域则没有。相比之下,用 MFCC 特征的 Dot-Plot 图,词重复区域则没有明显的黑色斜带,词重复区域和非词重复区域没有明显的差别。这表明后验特征能够比 MFCC 更好地表征语音重复片段声学上的相似性。此外,从 LCMA 子路径取值的波动范围也可以看出这一点。基于 MFCC 特征得到的 LCMA 路径的 f 值变动范围是 7.3～8.8,而基于后验特征得到的 LCMA 路径的 f 值变动范围是 1～5.9,词重复区域的 LCMA 路径的 f 值是 1.7、1.1 和 1.8,非词重复区域的 LCMA 路径 f 值大都大于 4,即从数值上区分了“像”与“不像”。

3 实验与分析

本文在 TDT2 中文广播语料库上选取了一个说话人 500 min 的音频数据作为单说话人的实验数据;选取了 10 个人共计 50 min(每人 5 min)的数据作为多说话人实验数据。在这两组数据上分别进行了实验。另外,本文也验证了用词边界为 SDTW 分段的情况下语音模式发现的效果。此时,只需采用 DTW 对两个词在距离矩阵中找出一条最优路径,然后使用式(5)算出这两个词的 f,此时该公式中

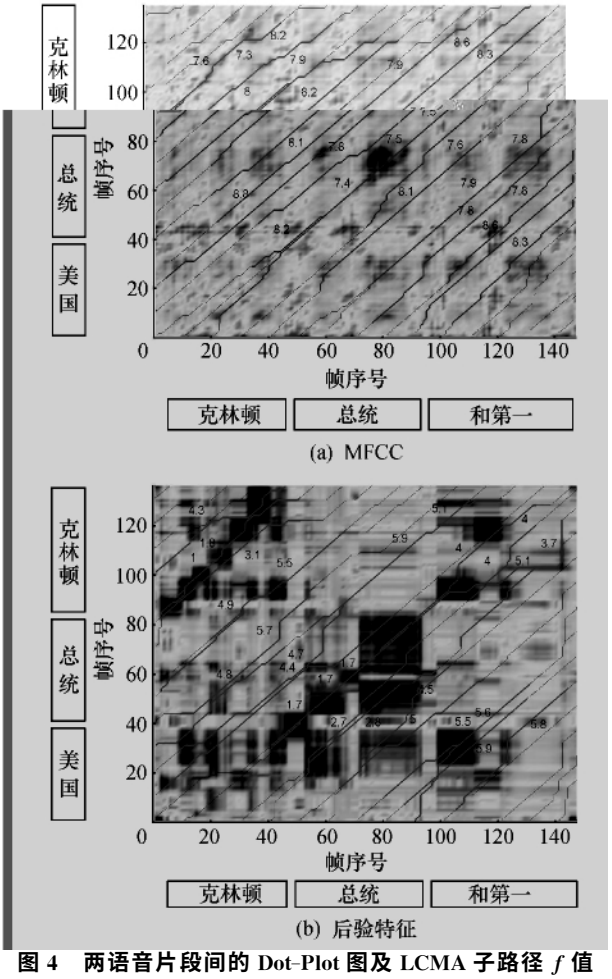

图 4 两语音片段间的 Dot-Plot 图及 LCMA 子路径 f 值

的 $t=N$, $s=1$。然后,设定一定的阈值 θ 对“词对”进行取舍,若 $f>\theta$,则丢弃,否则保留。词边界的时间信息从语料库提供的语音识别抄本上获得。这种有边界信息的实验可看作模式发现的效果上限。实验结果采用查准率、召回率和它们的调和平均即 FMeasure 进行衡量。实验在一台 P4 双核 CPU、4 GB 内存的计算机上完成。通过调整参数 (L,R,θ) 可以获取不同参数情况下的实验结果指标。表 1 记录了在 FMeasure 最优的情况下的实验结果。

表 1 实验结果比较

算法	说话人	特征	查准率	召回率	FMeasure	计算时间/s
SDTW	单	MFCC	0.345	0.277	0.307	13 415
		后验	0.381	0.359	0.370	14 425
	多	MFCC	0.175	0.136	0.153	13 200
		后验	0.297	0.268	0.282	14 160
SDTW+词边界	单	MFCC	0.658	0.363	0.468	329
		后验	0.735	0.620	0.673	1 784
	多	MFCC	0.333	0.195	0.246	318
		后验	0.624	0.549	0.584	1 741

从表1可以看到：1）无论是有无词边界信息的情况下，音素后验特征的效果总是明显优于MFCC特征，特别是在多说话人的语料上，后验特征对FMeasure的提升更大。2）在词边界信息的帮助下，语音模式发现的效果提升明显，不仅FMeasure大大提高，在时间消耗上也要省得多。原因是显而易见的：1）已知词边界后，比对的语音片段直接就是两个词或短语，而不会出现“跨词”的现象，即语音片段对应两个词的各一部分的现象，这样比SDTW算法通过LCMA算法找子路径鲁棒很多，效果自然会得到提升。2）由于省去了SDTW算法划分斜带状区域，并且在每一个斜带中进行DTW算法，而后再通过LCMA寻求符合要求的子路径的步骤，基于词边界信息的方法只需进行一次DTW运算，计算时间自然也就大幅减少。然而，时间消耗的减少是基于已知词边界信息的前提下的，因此寻求一种直接在声学信息上找出词或短语边界的方法，才能真正使这种优势体现出来。

4　总结与展望

本文在中文语料上进行了语音模式发现的研究工作。实验发现：基于音素的后验特征明显优于MFCC特征；词边界信息可以帮助提升模式发现的效果，加快计算效率。本文的工作将对今后中文模式发现任务具有借鉴与推广意义。对于后续的工作，拟从以下两个方面着手：1）解决SDTW效率低的问题。一种思路是循着本文的启发，寻找一种非监督的词边界检测方法；另一种思路是利用Zhang等[13]的思路，使用DTW下界实现加速。2）完成语音模式聚类工作。借鉴文[2]的思路，采用聚类方法，对检测到的词重复进行聚簇，以获得“语音模式”类。

参考文献　(References)

[1] Saffran J R, Aslin R N, Newport E L. Statistical learning by 8-month old infants [J]. *Science*, 1996, **74**: 1926-1928.

[2] Park A, Glass J. Unsupervised pattern discovery in speech [J]. *IEEE Transaction on Acoustic, Speech and Language Processing*, 2008, **6**(1): 1558-1569.

[3] ZHANG Yaodong, Glass J. Towards multi-speaker unsupervised speech pattern discovery [C]// IEEE International Conference on Acoustic, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE Press, 2010: 4366-4369.

[4] Jansen A, Church K, Hermansky H. Towards spoken term discovery at scale with zero resources [C]// Interspeech. Grenoble, France: ISCA, 2010: 1676-1679.

[5] Muscariello A, Gravier G, Bimbot F. Audio keyword extraction by unsupervised word discovery [C]// Interspeech. Grenoble, France: ISCA, 2009: 2843-2846.

[6] Anguera X, Macrae R, Oliver N. Partial sequence matching using an unbounded dynamic time warping algorithm [C]// IEEE International Conference on Acoustic, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE Press, 2010: 3582-3585.

[7] Dredze M, Jansen A, Coppersmith G, et al. NLP on spoken documents without ASR [C]// EMNLP. Boston, MA, USA: MIT Press, 2010: 460-470.

[8] ZHENG Lilei, Leung C, XIE Lei, et al. Acoustic texttiling for story segmentation of spoken documents [C]// IEEE International Conference on Acoustic, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE Press, 2012: 5121-5124.

[9] Aradilla G, Vepa J, Bourlard H. Using posterior-based features in template matching for speech recognition [C]// Interspeech. Grenoble, France: ISCA, 2006: 2570-2573.

[10] WANG Haipeng, Leung C, Lee T, et al. An acoustic segment modeling approach to query-by-example spoken term detection [C]// IEEE International Conference on Acoustic, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE Press, 2012: 5157-5160.

[11] BUT Speech@FIT. Phoneme Recognizer Based on Long Termporal Context [R/OL]. [2013/04/08]. http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[12] Lin Y, Jiang T, Chao K. Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis [J]. *Journal of Computer and System Sciences*, 2002, **65**: 570-586.

[13] ZHANG Yaodong, Glass J. An inner-product lower-bound estimate for dynamic time warping [C]// IEEE International Conference on Acoustic, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE Press, 2011: 5660-5663.