文章编号：1004-4353(2014)01-0045-04

基于帧符号化的语音相似性度量方法

刘双君， 金小峰， 崔荣一*
（延边大学工学院 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002）

摘要：提出了将语音帧符号化后度量语音相似性的方法. 首先，去除语音段中的静音部分，并提取每帧语音的 MFCC 参数；其次，将 MFCC 参数进行 k 均值聚类和 KNN 分类，并根据分类结果对语音信号进行符号化；最后，采用编辑距离计算语音段之间的相似性. 实验表明，将语音符号化后，音频之间的可区分性更加明显，识别率也有了明显提高.

关键词：MFCC；k 均值聚类；KNN 分类；符号化；编辑距离
中图分类号：TP391.41 **文献标识码**：A

Research on speech similarity based on frame symbolization

LIU Shuangjun， JIN Xiaofeng， CUI Rongyi*
（*Intelligent Information Processing Lab.*, *Dept. of Computer Science & Technology*, *College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: We presented a method to measure similarity of speech by using frame symbolization. Firstly, removing silence parts from speech segments, MFCC coefficients were extracted from each frame. Secondly, MFCC coefficients were classified by KNN-classification algorithm in terms of k-means clustering results, and speech signals to do symbolization processing according to the classification. Finally, speech similarity was computed by using Levenshtein distance. Experiment results show that frame symbolization makes distinction between different speeches are more obvious, and recognition rate has improved significantly.

Key words: MFCC; k-means clustering; KNN-classification; symbolization; Levenshtein distance

原始的音频数据是一个非语义符号表示的无结构化的数据流，缺乏内容语义的描述和结构化的组织，因而音频分析受到极大的限制[1]. 相似性度量是基于内容的多媒体检索技术的关键步骤之一[2]. 对声音相似性度量方法而言，其面临的主要困难之一是呈现手段的匮乏，这与声音的主观性特点密切相关；同时，以帧为单位提取的单个特征向量也不能完全反映声音片段之间的相似性关系[3]. Subramanya 等人[4]直接针对音频样本的二值图像进行了分割处理，这种方法虽然简单直观，但由于采样率、量化位和时长等因素，并不具备实用性. Foote[5]利用可视化方法对音频的时间结构进行了分析，以寻找音乐中的自相似特点，如提取鼓点的节奏，发现旋律重复的特点. 研究[6]表明，音乐检索中基频作为语音旋律的一个重要特点，将其按照上升、平稳、下降的变化将语音转化为一个三元化的音符序列，也有较好的表现.

本文首先将语音分帧后提取 MFCC 参数[7]，并将 n 段语音的 MFCC 参数进行聚类，然后将其中每一段语音的每一帧进行分类，并将其映射成一个相应的字符；n 段语音相应转化为 n 个字符串后，计算每两段字符串的编辑距离，即每两段语

收稿日期：2013-12-24
* **通信作者**：崔荣一(1962—)，男，博士，教授，研究方向为模式识别、智能计算.

音之间的相异性,从而得出其相似性.

1 语音帧符号化

不同的人说的同一段话其信息是一样的,因此可以假设:相同内容的语音信号应该归于同一类中,不同内容的语音信号归于不同的类中.进一步可延伸为,相同内容的语音信号映射为同一个赋予特定含义的抽象字符串,不同内容的语音信号映射为不同的字符串,这样每段语音信号就可以用一个字符串表示.

1.1 语音信号聚类

如果要将语音信号映射成一个字符串,首先应该将语音信号分为 m 类,即映射后的字符串由 m 个基本字符构成;因此,本文将提取的 MFCC 参数(符合人耳的听觉特性)采用 k 均值聚类[8]的方法进行聚类,得到 m 个互不重叠的类空间.聚类算法如下:

1) 指定簇数目 m,以及簇中心的初值和结束条件.簇中心的初值为样本空间的前 m 个数据,结束条件为迭代 $N(N=1\,000)$ 次内两次迭代的簇中心的差值不超过阈值 $T(T=0.01)$ 或者迭代次数达到 N;

2) 采用欧氏距离计算相似性,计算样本空间中各样本与簇中心的距离,距离最小的样本划归同类;

3) 重新计算每个类的簇中心,得到 m 个新的簇中心;

4) 判断是否满足步骤 1) 中的结束条件,若满足条件则结束,得到 m 个簇中心,否则执行步骤 2).

1.2 语音信号分类

得到 m 个簇中心后,采用 KNN 算法[9]对语音进行分类,即将一段语音映射为一个字符串.首先,计算样本空间中每个样本与每个簇中心的距离(本文采用欧式距离);然后,找出样本空间中每个样本与 m 个簇中心距离的最小值,再将此样本与此簇中心归为同一类,并映射为同一个字符;最后,得到每一段语音所对应的字符串.

2 音频相似性分析

用上述方法将每段语音映射成一个字符串后,语音文件之间的相似性度量就转换为字符串之间的相似性度量.编辑距离则是字符串相似性度量的一个经典算法,为了说明本文方法的有效性,将其与直接采用 MFCC 进行语音相似性度量的方法进行比较.

2.1 编辑距离

编辑距离(Levenshtein distance)[10]由 Levenshtein 于 1966 年提出,是指由字符串 S 变化到目标字符串 T 所需要的最小编辑操作的次数.这里所指的编辑操作是指对字符串的某一个位置的字符进行删除、插入、替换的操作,如字符串“kitten”与“sitting”的编辑距离为 3,计算过程中发生了 2 次替换和 1 次插入操作.为了便于对比多对字符串之间的相似程度,本文将一对字符串的编辑距离与该字符串对中最长的字符串长度相除后的距离作为本文的编辑距离,如“kitten”与“sitting”之间的编辑距离为 3/7.

图 1 为 2 个人的 8 段语音信号,其中(a)与(e)、(b)与(f)、(c)与(g)、(d)与(h)的语音内容相同,(a)、(b)、(c)、(d)是第 1 个人的语音,(e)、(f)、(g)、(h)是第 2 个人的语音.将上述 8 段语音采用本文提出的帧语音符号化后,计算编辑距离得到的结果见表 1.表 1 表明,相同内容的语音段间的编辑距离较小,因此验证了本文所提出方法的可行性.

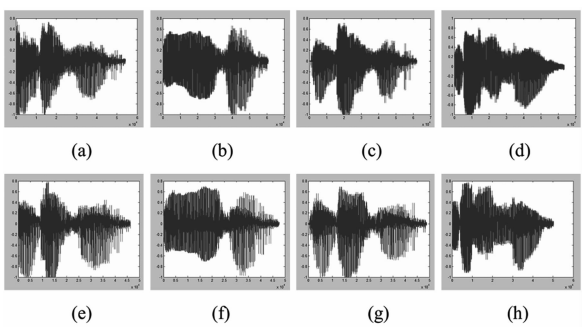

图 1 不同的语音信号

表 1 语音间的编辑距离

	a	b	c	d
e	0.641 9	0.945 0	0.913 9	0.894 7
f	0.728 3	0.747 2	0.924 7	0.884 2
g	0.679 0	0.923 0	0.752 6	0.873 6
h	0.864 1	0.978 0	0.946 2	0.736 8

2.2 DTW 距离

经典的语音相似性度量方法是直接采用符合人耳听觉特性的 MFCC 参数，但由于语音段长度的不同会导致提取到的 MFCC 参数的维度不同，因此需要采用 DTW 算法[11]度量它们之间的相似性. 表 2 是图 1 中各语音段间的 MFCC 参数之间的 DTW 距离. 由表 2 可知，相同内容语音间的差异较小，不同内容语音间的差异较大.

表 2　语音间的 DTW 距离

	a	b	c	d
e	5 297	12 997	6 510	10 790
f	11 683	4 712	15 208	13 761
g	10 054	14 580	5 335	14 082
h	10 855	16 090	12 668	6 330

2.3 本文方法的流程图

首先将语音中的静音部分(此时无人说话)去除，分帧后提取 MFCC 参数，并对 n 段语音的 MFCC 参数进行聚类；然后将其中每一段语音的每一帧数据进行分类，并将其映射成一个相应的字符，n 段语音相应转化为 n 个字符串；最后计算每两个字符串的编辑距离，可得出每两段语音之间的相似性. 处理过程如图 2 所示.

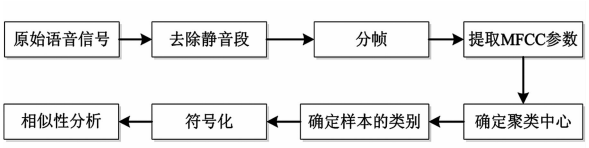

图 2　语音相似性分析流程图

3　实验结果及分析

为验证本文方法的有效性，设计了 3 组实验，实验的语音数据由 SONY 公司生产的PCM-D50 线性录音棒录制，采样频率均为 44.1 kHz.

3.1 聚类个数的对比实验

在语音帧符号化前的聚类阶段，初始的簇数目将对后续的分类结果产生重要的影响，因为它决定了一组音频数据符号化后由几个基本字符构成. 簇数目的确定还没有可靠的理论依据，目前只能通过实验来确定最佳的簇数目.

表 3 表示在簇数目为 16、15、14 时，分别进行 19 次实验所获得的错误识别率. 第 1 至第 10 次实验的数据是不同的人说相同内容时所获得的，第 11 至第 19 次实验的数据是相同的人分两次说相同内容时获得的. 从表中可以看出，在聚类数目为 15 时，其错误率较小，因此，在本文实验中确定聚类数目 $m=15$；第 11 至第 19 次的实验结果要好于第 1 至第 10 次的实验结果，这说明由于不同的人可能来自不同的地方，其特有的地方口音会对实验结果产生影响.

3.2 本文方法与 DTW 方法的对比

为了验证本文方法的有效性，将上述采集的 19 组语音数据分别与 DTW 方法进行对比实验. 表 4 为使用 DTW 方法和本文方法产生的错误率. 由表 4 可知，本文方法的平均错误率为 14.47%，DTW 方法的平均错误率为 21.05%，由此表明本文方法效果要优于传统方法.

表 3　不同聚类数目下各次实验的错误率

m 值	各次实验的错误率																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
16	0.25	0.5	0.5	0.5	0	0.25	0.25	0.5	0.5	0.25	0	0	0	0	0	0	0.25	0	0
15	0.25	0.25	0.25	0.25	0.25	0.25	0	0.25	0.5	0.25	0	0	0	0	0	0	0	0	0
14	0.25	0.75	0.25	0.25	0.25	0.25	0	0.25	0.5	0.5	0	0	0	0	0	0	0	0.25	0

表 4　本文方法与 DTW 方法的错误率

	各次实验的错误率																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
本文方法	0.25	0.25	0.25	0.25	0.5	0.25	0	0.25	0.5	0.25	0	0	0	0	0	0	0	0	0
DTW 方法	0.5	0.5	0.25	0.5	0.5	0	0.25	0.75	0	0.25	0.25	0	0.25	0	0	0	0	0	0

3.3 可区分性比较

为了进一步说明本文方法的优越性,本文引入可区分性度量函数 $H(x)$. $H(x)$ 是对一组数据中的某一个数据与此组数据相似性的一个评价,其公式为

$$H(x)=\mathrm{abs}\left(\frac{x-\bar{X}}{\mathrm{std}(X)}\right), \tag{1}$$

其中 X 为任一组数据,x 为此组数据的任一元素,$H(x)$ 越大说明此元素在此组中的可区分性越好. 图3是本文方法和DTW方法的可区分性比较示意图,其中横轴为实验的次数,纵轴表示每次实验后两种方法最差的可区分性度量. 由图3可看出,本文方法的可区分性较好.

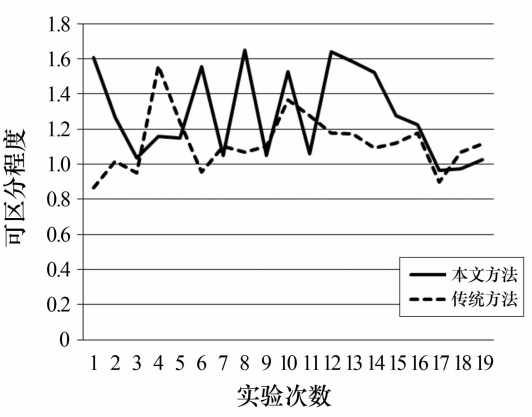

图3 可区分性比较

4 结论

针对音频的相似性度量,本文提出了一种将音频符号化后再计算其相似性的方法. 通过将语音符号化后,使许多单纯的数值表示的语音信号抽象化为由一系列基本字符表示的字符串,简化了相似性的度量运算,并且符号化后的语音信号之间的相似性度量方法的准确率也高于传统的DTW方法. 如何结合其他的语音特征提高算法的鲁棒性,降低不同口音带来的影响,提高识别率是本文今后进一步的研究工作.

参考文献:

[1] 张自强. 基于内容的音频匹配研究[D]. 上海:华东师范大学,2012.

[2] 李丙洋. 基于音频内容的多媒体文件相似性快速比对研究[D]. 哈尔滨:哈尔滨工业大学,2013.

[3] 李超,熊璋,朱成军. 基于距离相关图的音频相似性度量方法[J]. 北京航空航天大学学报,2006,32(2):224-227.

[4] Subramanya S, Abdou Y. Segmentation of audio data based on the binary images of the audio samples[C]//In: Proc of Inter Conference on Intelligent Systems. Denver: ISCA, 1999:137-141.

[5] Foote J. Automatic audio segmentation using a measure of audio novelty[C]//In: Proc of ICME 2000. NY: IEEE, 2000:452-455.

[6] 曹文晓. 哼唱检索中基于分段信息的匹配算法研究[D]. 北京:清华大学,2010.

[7] Skowronski M D, Harris J G. Increased MFCC filter bandwidth for noise-robust phoneme recognition [C]//In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Florida: IEEE, 2002:801-804.

[8] 蔡碧野,吴一帆,谢中科,等. 数据挖掘中聚类的研究[J]. 计算机工程与应用,2003,17(2):39-42.

[9] 孙岩,吕世聘,王秀坤,等. 基于结构学习的KNN分类算法[J]. 计算机科学,2007,34(12):184-187.

[10] Levenshtein V L. Binary codes capable of correcting deletions, insertions and reversals[J]. Doklady Akademii Nauk SSSR, 1966,163(4):707-710.

[11] Itakura F. Minimum prediction residual principle applied to speech recognition[C]//In: IEEE Trans Acoustics, Speech, and Signal Proc. IEEE: 1975, 23(1):67-72.