语音合成系统中韵律参数的聚类研究

王玮　　蔡莲红
（ 清华大学计算机科学与技术系　　北京　　100084 ）
联系邮件：cliff_wang@263.net

摘　要

语音合成系统中韵律模块是由音节的多个韵律特征参数组成，如音节的音高、音长和幅度等，这些韵律参数有的是以单值形式表示如音节的音长，有的是则是序列形式如音节的音高，这里的序列指得音节本身所具有连续频率特征。一般认为采用音节最大值和最小值的平均得到基频中值可以描述音高特征，但是这是依赖于语言学家的定性描述。本文从大量的音节基频序列中抽取数据，组成基频向量，通过对音节基频序列的进行聚类分析，说明采用基频中值作为音高特征的描述信息的合理性。

1. 引言

随着语音学和计算机技术的发展，语音合成系统的研究已经取得了重大进展，并且成功地应用于许多不同的场合，但是合成系统的输出语音带有浓重的机器味，与人类自然流畅的发音相比还存在一定的差距。其中重要的原因是受制于人们对韵律特征和规律的认识。

语音作为人的发声器官发出来的一种声波具有声音的物理特性，每一种音都具有一定的音色、音调、音强和音长。音色也叫音质，是一种声音区别于其它声音的基本特征；音调是指声音的高低，在汉语语音学中称为音高，音调取决于声波的频率；声音的强弱叫做音强，它是由声波的振动幅度决定的；声音的长短叫做音长，它体现了发音持续时间的长短。

语音的韵律参数也称为超音段参数，一般是指音节的音高、时长和幅度等参数，通常以声音的基频表示音高，由于汉语是一种声调语言，因此其音高值是声调中最重要的特性。声调的音高值是声带基本振动频率的表现，不同声调的基频值各不相同，从敏感的声学仪器上得出的声调频率值是千变万化的。在语言学上，表述一个声调的音高值时，通常采用相对的物理量进行表示，即从一个人或从几个人的声调频率中求出高低曲直的相对的平均关系值来表示。然而声调的音高值不是一个单一的频率值而是一个持续性的波段，而且在声调持续的过程中，从起点至终点往往又有频率的变化，语言学家认为通过对音节最大值和最小值的平均得到基频中值能够较好地描述一个声调的变化规律。本文对这一问题进行了分析研究，采用动态数据聚类的方法分析基频数据的特征，并进行了实验分析，说明了这种表示方法的合理性。

2. 常用的聚类方法

聚类方法是数据分析的常用方法之一，我们可以将聚类算法分成层次式和非层次式两种，非层次式聚类算法按照一定的标准将数据划分成 K 个聚类（K 是算法要求输入的期望聚类数），其中最常用的是平方差标准，其目的在于找到使得平方差最小的 K 个聚类，即每个聚类内部数据点间的距离尽可能小，聚类间数据点的距离尽可能大。

数据库中的聚类对象是例子，每个例子由不同的属性构成，这些属性主要分成为两类：数值属性（Numeric Attributes,可以比较大小）和符号属性（Categorical Attributes，不能比较大小）。在数据挖掘领域中，由于要处理非常大而复杂的数据集，所以对传统的聚类方法提出两个需要尽量满足的要求：1. 能同时处理数值属性和符号属性；2. 算法的效率要满足大的数据集的大数量、高复杂性、增量的要求。在现存的聚类方法中，如果能同时处理数值属性和符号属性，那么一般来说，效果很低；而对那些效率高的算

法而言，他们大都只能处理数值属性。

2.1 神经网络方法

神经网络方法中用于聚类的方法主要是 SOM(Self-Organizing Feature Map)神经网络，它由输入层和竞争层组成，输入层由 N 个输入神经元组成，竞争层由 m×m=M 个输出神经元组成，且形成一个二维平面阵列。输入层各神经元与竞争层各神经元之间实现全互连接。该网络根据其学习规则，通过对输入模式的反复学习，捕捉住各个输入模式中所含有的模式特征，并对其进行自组织，在竞争层将聚类结果表现出来，进行自动聚类。竞争层的任何一个神经元都可以代表聚类结果。

SOM 方法是一种两阶段（制定聚类中心、聚类中心的修改）基于欧式距离的反复循环过程。显然这种方法只能针对于数值属性。

SOM 网络的最大局限性是，当学习模式较少时，网络的聚类效果取决于输入模式的先后顺序，而且网络连接权向量的初始状态对网络的收敛性能有很大的影响。

2.2 矢量量化方法

矢量量化方法 VQ 中 LBG 方法用来进行聚类，通常的做法是将所有要识别矢量的集合分成若干子集，各个自己中的矢量具有相似特征，因而能用一个具有代表性的矢量来表示。该具有代表性的矢量成为码字，全体码字的集合称为码本。

为了使这种方法的迭代运算不至于无限循环下去，设置了 δ 和 L 两个阈值参数。δ 的值设置得远小于 1，当 $\delta^{(m)} < \delta$ 时，表明再进行迭代运算畸变的减小是极有限的，只是可以停止运算。L 是限制最大迭代次数的，防止 δ 设置得较低时迭代次数过多。

这种方法也是一种两阶段（指定聚类、聚类中心的修改）基于欧式距离的反复驯化过程，针对数值属性。系统的总畸变是它的 M 个码字决定的状态空间点的函数。在大多数实际情况中，该函数并非凸函数，既有全局最小点，又有多个局部最小点。所以算法一般取得的是一个局部最优解。

2.3 C 均值方法

C 均值算法是一种分割的而非分层的聚类方法，在数据分析中得到广泛的运用。其可以描述成给定的一个例子的集合 X ，集合 X 中每个属性均为数值属性，和一个整数 $k(k \le n)$ ，算法将 X 分割为 k 个聚类并使得在每个聚类中所有值和该聚类中心距离的总和最小，每个聚类的聚类中心是每个聚类的均值。

C 均值算法具有如下特点：1. 能有效的处理大数据集；2. 经常终止于一个局部的最优解；3. 由于欧式距离的局限性，仅能处理数值属性；4. 聚类结果具有凸的外形；5. 算法的执行结果和例子的顺序有关。

鉴于 C 均值算法的特点，决定采用这种方式处理语音基频序列数据。

C 均值算法描述如下:

step1：选择把 N 个样本分成 C 个聚类的初始划分，计算每个聚类的均值 m_1, m_2, Λ, m_c 和 J_e。这里 N_i 是第 i 聚类 Γ_i 中的样本数目， m_i 是这些样本的均值，即

$$m_i = \frac{1}{N_i} \sum_{y \in \Gamma_i} y \qquad (1)$$

把 Γ_i 中的各样本 y 与均值 m_i 间的误差平方和对所有类相加后为：

$$J_e = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \| y - m_i \|^2 \qquad (2)$$

J_e 是误差平方和聚类准则，它是样本集和类别集的函数。 J_e 度量了用 C 个聚类中心 m_1, m_2, Λ, m_c 代表 C 个样本子集 $\Gamma_1, \Gamma_2, \Lambda, \Gamma_c$ 时产生的总的误差平方。

Step2：选择一个备选样本 y ，设 y 现

在在Γ_i中。

Step3：若$N_i=1$，则转 Step2，否则继续。

Step4：计算

$$\rho_j=\begin{cases}\dfrac{N_j}{N_j+1}\left\|y-m_j\right\|^2 & j\neq i\\ \dfrac{N_j}{N_j-1}\left\|y-m_j\right\|^2 & j=i\end{cases} \tag{3}$$

Step5：对于所有的j，若$\rho_k\le\rho_j$，则把y从Γ_i移到Γ_k中去。

Step6：重新计算m_i和m_k的值，并修改J_e。

Step7：若连续迭代N次J_e不变，则算法结束，否则转到 Step2。

3. 基频数据的预处理

语音合成系统中的基本单位是音节。通过对音节波形的过零率的统计可以计算出音节对应的基频序列。

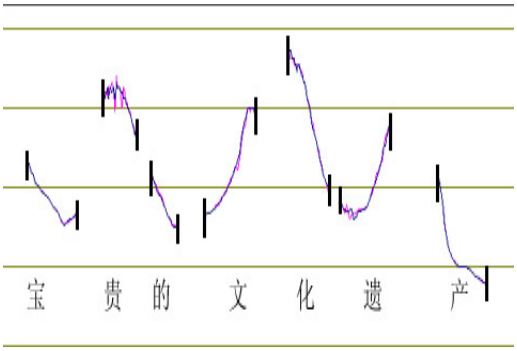

图 1 音节的基频序列

基频序列包含了语音特有的信息，它和音节的声调调值有密切的关系，但又不完全是调值的机械映射。它还和音节在句子或短语中的位置以及整个句子或断语的语调有着密切的关系。这在声学的定性研究中已经得到了肯定。这里，希望通过对实际数据的分析来指导语音合成，所以，要抽取基频序列的具体信息。

如图 1 所示，音节“宝（bao）”的基频序列值为：{235,…,225,…,156,…163}。这一系列的基频值描绘了一条起伏的基频曲线。

作为研究的对象，基频序列需要进行长度的归一化处理，而且没有必要将整个基频序列的值作为归类的对象。这里采用的方法是将整个基频曲线等分成三部分，取每部分序列的均值。最终，每个音节对应一个三维的矢量。由此，数据量减少了，但是保留了基频曲线的变化趋势。

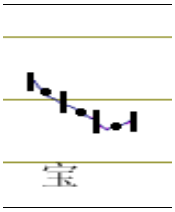

图 2 抽取特征向量

抽取基频的特征向量如图 2 所示，这里音节“宝(bao)”的特征矢量为：{219，179，156}。

4. 聚类分析

使用C均值聚类方法对经过预处理的基频数据进行动态聚类。在具体的数据聚类中采用以下三种方式进行比较说明。

第一种方式：直接对产生的基频矢量集合进行聚类操作。保留了能量信息。得到如图 3 所示的结果.

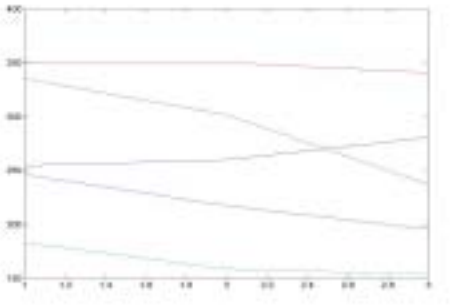

图 3 直接聚类的中心矢量

图 3 中每一条曲线都代表了一个聚类中心矢量。从图 3 中我们可以看出直接聚类的

结果是反映了调型信息。

第二种方式：音节的基频矢量减去每个句子（短语）的平均基频值。得到的聚类中心如图 4 所示，然后再进行聚类操作，得到如图 5 所示的结果。聚类的结果是反映了调型信息。

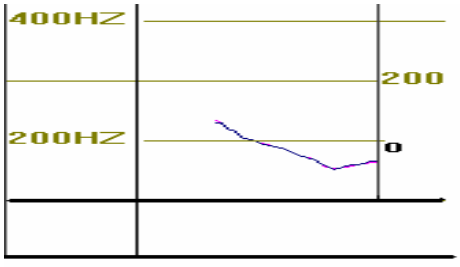

图 4 减去短语的平均基频

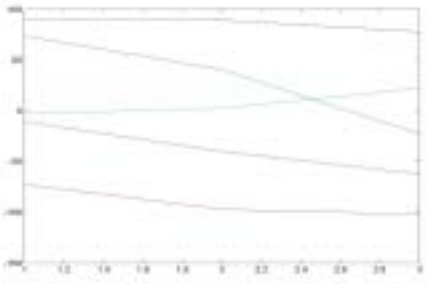

图 5 减去短语平均基频的聚类中心矢量

第三种方式：每个基频矢量减去自身三个分量的均值，归一化到坐标原点的附近，只留下调型信息，其结果如图 7 所示。

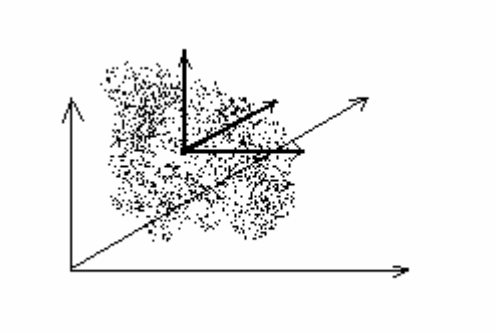

图 6 矢量位置的平移

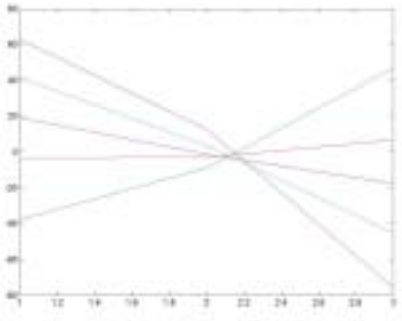

图 7 减去自身分量均值的聚类中心矢量

从图 3、图 5 以及图 7 的比较可以发现，直接对基频矢量进行聚类和将基频矢量减去短语的均值后产生的聚类中心基本一致，只出现了少量的坐标平移。说明在分布上，各个短语的均值之间的相差并不大，然而各音节本身的均值差异却比较大，因此说明采用基频中值描述韵律参数的音高特征是切实可行的。

在消除了音节本身的均值大小差异后，产生的聚类中心向量可以看出语调的雏形。当然，这几个中心向量并非完全按照语调来分布，原因是基频序列所反映的语调受到整个短语语气的影响，并非完全按照语调的调型。

5. 结束语

本文采用了聚类算法对语音合成系统中的音高特征的描述方法进行了研究，实验结果表明直接对基频矢量进行聚类和将基频矢量减去短语的均值后产生的聚类中心基本一致，各音节本身的均值差异却比较大，说明采用基频中值代替音频序列描述音节的音高特征是合理的。

6.参考文献

1. 郭锦桴，汉语声调语调阐要与探索，北京语言学院出版社，1993
2. 姚天任，数字语音处理，华中理工大学出版社，1992
3. 杨行峻，迟惠生，语音信号数字处理，电子工业出版社，1995
4. 边肇祺，模式识别，清华大学出版社，1995