语音乐律研究报告

Status Report of Phonetic and Music Research
2013

北京大学中文系语言学实验室

Linguistic Lab
Department of Chinese Language and Literature
Peking University

目录

中文篇

英文篇

压轴篇

语言文化数字化传承的理论与方法*

孔江平

（北京大学中文系，中国语言学研究中心）

摘要： 语言文化传承是人类文明最重要的组成部分，传承的方法也随着社会和科技的进步在不断发展。当今全球一体化的进展使得世界有声语言和口传文化的多样性濒临消亡的境地。论文从传统文字记录语言文化的性质和缺陷以及现今音频、视频记录语言文化的不精确和不完整性，讨论了有声语言和口传文化数字化传承的基本方法，并提出了一种基于语音多模态的数字化传承方法。最后，论文还从有声语言和口传文化认知的角度讨论了全面精确传承语言文化的理论基础。

关键词： 有声语言；口传文化；数字化方法；文化传承

一、引言

人类的文明和进步基于人类优秀文化的积累和传承，而文化的传承又基于语言文化的记录方式和载体，目前世界文化传承主的要形式是语言文字和口传文化。人类语言产生于人类漫长的进化过程，自从有了语言，人们就开始创造文化，特别是口传文化。时至今日，世界上大部分的语言都没有自己的原创文字，其语言文化主要是以口传的方式传承，例如基于各种民族有声语言的史诗、口传经文、原生态民歌等。因此只有部分民族的有声语言和口传文化被记录下来，然而，随着部分有声语言的消亡，基于这些有声语言和口传文化的知识体系也随之消亡。

中国是一个多民族的国家，有丰富的语言及口传文化资源（谭克让，麦克康奈尔主编，1995；《中国戏曲音乐集成》编辑委员会，2002）。但随着经济全球化趋势的加剧，大众传媒的高度发达，使用不同语言的人群接触频繁，致使一些使用人数较少的语言交际功能日渐衰退，消亡速度日益加快。事实上，语言不仅是交际的工具，同时也是文化的载体。每一种语言都承载着一种独特的文化，凝聚着一套知识体系，是人类文明最重要的组成部分。语言，尤其是占绝大多数的无文字语言一旦消亡，大量人类的知识体系也将随之消失，这无疑将给文化的多样性和人类文明的传承带来不可挽回的损失。任何一种民族的口传文化都是在百万年人类进化的长河中逐步形成的，因此，这种“文化物种”的产生和发展的过程并不亚于自然物种，都是人类智慧的宝贵财富。但目前各国列出的语言文化遗产，大都处境艰难，濒临绝灭，这说明全球都面临着如何保护优秀传统文化的问题。应对自然物种的消亡，目前主要采用建立种子库和基因库的方式加以保护。而对有声语言及口传文化这种“文化物种”的保护，世界各国还没有形成共识，主要方法有书面记录、录音和录像，然而，这些方法达不到全面保护和传承的目的。如何有效地对有声语言及口传文化进行抢救和保护是人类面临的一个迫切需要解决的问题。

从理论上看，文化遗产具有多样性。其中，语言或称有声语言和口传文化就是世界文化遗产中最为重要、特殊和具有价值的一种，之所以这样说是因为：1）语言是人类思维的工具，存在于人的大脑中，人类所有非物质文化都基于人类的语言；2）有声语言和口传文化是以生理信号（脑电、声门阻抗信号等）和物理信号（音频、视频等）信号的方式来承载的，瞬时即失；3）活的语言一旦消亡，口传文化就随之消失。这些特殊性质对有声语言和口传文化保护和传承的方法提出了不同的要求。目前从世界语言和口传文化记录与传承上看，还缺少基本的理论框架和有体系的科学方法。

随着科学技术的进步和社会发展，特别是现代通讯技术的兴起和应用，语言研究已经涉及到自然科学的许多领域，如，基于生理和声

*本项研究由国家社会科学基金重大项目“中国有声语言及口传文化保护与传承的数字化方法研究和基础理论研究”资助，项目批准号：10&ZD125。

学的语音数字化研究、基于有声语言的口传文化研究、基于语音信息量演化计算的语言认知和编码研究、基于语言发音生理的人类语言进化研究、基于语言生理和语言情感的歌曲、戏剧和原生态声乐研究等。显然，语言和口传文化研究已经逐步科学化。本文在对有声语言生理、声学和数字化研究以及对具有独特性质的口传文化的生理和声学研究的基础上，对有声语言和口传文化数字化传承的基本理论和方法展开讨论。

二、 记录语言和口传文化的传统方法

人类传统记录语言和口传文化的方法主要是文字和乐谱，由于语言的不同和口传文化的差异，文字系统和记录乐谱的形式也十分不同，这里以中国语言和口传文化为例，简要阐述语言文字系统和口传文化记录的形式、性质和缺陷。

1. 传统语言文字记录

人类自从有了文化，就试图记录下来。从中国各地不同时期的岩画中，除了图画外，可以很容易地发现很多抽象的符号，如广西的华山岩画。显然人们早期使用图画来记录生产和生活事件，然后简化图画逐步符号化。从中华文明的历史轨迹看，除了早期的岩画和符号外，比较系统的图画符号体系是东巴文，东巴文是介于语言文字系统和图画符号系统之间的一种形式和阶段。从语言学的角度看，中国最早的文字体系是殷商的甲骨文。后来经过大篆、小篆、金文、楷书等发展阶段，最终形成了汉语的文字系统。除了汉字系统以外，中国主要的传统文字还有藏文、蒙文、维吾尔文、朝鲜文、彝文等。所有这些传统文字最终构成了中华民族的语言文字记录系统。

从语言文字记录文化的发展历史看，文字系统有许多重要的性质。首先，文字记录是以语言为基础，因此文字系统是语言系统的一种特殊形式。由于语言结构上的差异，文字系统在结构上也有很大的差异，例如汉藏语系语言的单音节性和声调，使得汉字采用表示单音节的方块字形式。其次文字系统具有鲜明的民族性，因此，文字系统中包含了一个民族特有的文化背景，如汉语的诗歌主要是句尾韵母押韵，而蒙语的诗歌主要是句首重音押韵，这和语言的结构有密切关系。第三，文字系统随着语言的演化而改变。众所周知，语言随着时间会产生变化，文字系统也会随语言产生相应的变化。

2. 传统口传文化记录

在口传文化的记录和传承上，不同的民族和语言采用不同的方式，但理论上一般采用语言文字加乐谱的形式，因此更为复杂。中国传统的记谱法主要有十三种,分别是:工尺谱、古琴谱、燕乐半字谱、弦索谱、管色谱、俗字谱、律吕字谱、方格谱、雅乐谱、曲线谱、央移谱、查巴谱和锣鼓经（王耀华，2006 年）。这些记谱方法除了可以记录音乐外，部分可以用来记录口传文化。如律吕字谱和工尺谱，前者通常使用宫（do）、商（re）、角（mi）、清角（和）（fa）、徵（sol）、羽（la）、变宫（变）（si）来表示不同的音高，中国古代律学的理论体系已经十分健全[1]。而工尺谱用合、四、一、上、尺、工、凡、六、五、乙等字表示音高，可相当于 sol、la、si、do、re、mi、fa、sol、la、si。高八度将谱字末笔向上挑，或加偏旁亻，低八度将字的末笔向下撇。另外，用“丶”或 “×、－”等符号作为节拍符号。昆曲是中国一种最重要的口传文化，包含了中国古代音乐和文学的精华，有六百年的历史，昆曲直至现在还在用工尺谱来记录[2]。除了表示音高外，还能记录许多唱腔和唱腔的组合[3]。昆曲的特点是以字行腔，所以在旋律上和汉语的声调有更密切的关系。

口传文化主要依靠口头流传，如古代传说、史诗、歌曲、戏剧、说唱等，它们各有自己的特点。首先口传文化是有声语言的一种特殊形式，通过特殊的声音来表达有声语言无法表达的特殊情感。在形式上主要是靠韵律、旋律、重音和节奏的变换表达特殊情感。其次，记录口传文化除了要用到语言外，还要用特殊

[1] 中国的古代的律学的理论体系已经很完善，王骥德的《曲律》，是中国律学的经典著作。

[2] 中国的昆曲一直沿用工尺谱，有大量的曲本，2009 年有广陵书社出版的《昆剧手抄曲本一百册》收集和整理了昆曲的主要曲目，整体上反映了工尺谱记录昆曲的面貌。

[3] 昆曲的声腔和节奏在俞振飞的《粟庐曲谱》中有详细的记录和论述，这些论述对现今昆曲的打谱和研究都有重要的参考价值。

的记谱方法来记录具有音乐性的旋律和节奏，即用文字记录内容和利用乐谱记录旋律，如记录昆曲除了用文字记录词以外，还要用工尺谱记录旋律和节奏。

3. 语言文字和乐谱记录的缺陷

虽然文字可以很精确地记录语言要表达的内容，但文字在记录时有一个致命的弱点，即文字记录文化和人们认识事物的程度相关，当认识不足时记录的内容只是人们当时认识事物的水平。众所周知，人们认识世界和创造文化都有一个逐步深入的过程，这就是说，古人记载的所有事物和文化现象都是古人当时的认识水平，并不一定全面和正确，有些可能是假象。这就造成了后人很难理解古人的许多文献纪录。例如，我们现在可以看到古代留下来的韵书，比较清楚地了解声韵调的类别，但我们还是很难知道古代汉语确切的读音。我们有《尔雅》和《说文解字》[4]，但我们对古代很多汉字的确切含义仍有质疑。因此中国小学中的音韵学、训诂学等都要研究和考证语言文字记录的不精确性带来的后果，这说明目前文字并不能很精确全面地记录人类的文化。同样对口传文化进行记录也是如此，虽然古人采用了特殊的方式记录口传文化，实际上还是不能精确记录它们。例如，我们今天能看到李白的诗歌，但我们还是无法知道李白吟诵诗歌时的字音和旋律。我们能看到姜夔《白石道人歌曲》中的十七首词曲[5]，但仍然很难知道准确旋律和节奏。

从以上的讨论可以看出，目前人类大多数文化遗产和口传文化主要靠语言文字和音乐符号记录，但语言文字和音乐符号的记录完全基于人们当时的认识水平，并不是完全精确和客观的记录。人们认识事物的错误和不精确也都被记录了下来。因此随着人们对事物和自然界规律认识的深入，我们不得不去考证古人留给我们的文献，去鉴别真伪。

[4] 郭璞的《尔雅》和许慎《说文解字》是中国最早的有系统地解释汉字字义的字典，这两部典籍对古汉的解读和研究至关重要。

[5] 在中国宋词文献中，只有姜夔的《白石道人歌曲》中有 17 首标注有工尺谱，这我们现在研究宋词吟唱旋律和节奏的珍贵资料。

三、语言记录的理论基础

语言是思维的工具，极为复杂，同时语言又具有民族性，并形成了各民族不同的语言文化，在此基础上最终构成了一个民族的知识体系。下面从语言记录的形式、语言的认知范畴来讨论语言记录的理论基础。

1. 语言记录的形式

从形式上看，语言记录可分为记录信号和记录介质两大部分，第一部分可以分为：1）图画类（图画，绘画）；2）图形类（东巴文）和 3）符号类（甲骨文，大篆，小篆，金鼎文，楷书，文字的数字编码）。符号类主要是文字，又可分为：a）象形；b）音符；c）象形音符组合。学界也有将文字按结构类型分为“非字母文字”和“字母文字”两大类的。从理论上讲，文字只有达到了记录语音的程度才能算作真正的文字系统。第二部分可以分为：1）硬介质（骨料，石料，竹简，木简）；2）纸张介质（锦帛，纸张）；3）电子介质（磁盘，光盘，硬盘）。在电子介质中只有达到完全数字化后，才可以无损耗复制。因此，目前人类在这方面也可以认为已经达到了传承方式的极限。

2. 语言的认知基础

现在人类认识自然和人文活动的能力和深度还都十分有限，这就决定了我们只利用语言文字和一些记录口传文化的符号系统不能很完善地记录我们的语言和口传文化，这是不是说现在我们就不能全面精确记录人类的语言和口传文化了呢？答案是否定的。虽然人类认识自然和人文活动的深度还需要很长的时间，但并不是说我们不能用科学的方法来记录我们的语言和口传文化用于将来的研究。那么怎样证明我们记录的语言和口传文化是全面精确和没有丢失信息呢？这就涉及到人类语言的认知能力和极限。如果我们知道人类语言的认知能力、范围和极限，就可以采用更多的科技手段和方法来记录语言和文化，达到全面和精确的程度，以便后人研究。

人类语言认知能力和极限一直都是语言学家、心理学家和脑神经学家研究和探索的目

标，虽然近十年来在这个交叉领域有了巨大的进展，但距离了解人类大脑的语言能力还相去甚远。但不用这些先进的仪器和设备，从语言学和统计的角度人们也可以对人类语言的认知能力有很好的了解。语言学家的研究告诉我们，在人类的上千种语言中，虽然有一千多个音素可以成为音位（phoneme），但对于每一个语言来说，其音位一般只有几十个左右（Ladefoged et al，1996）。在词汇方面，人们在口语中常用的词汇有三千左右，而且能覆盖语言交际和普通文献的百分之九十五以上。在语言词汇的习得方面，郑锦全先生经过多年对不同语言词汇掌握情况的研究提出了“词涯八千”的理论[6]，即一个人在学习一种语言时，能够掌握和应用的词汇不会超过八千个。例如汉语目前已经有三十万左右的词汇，但每个人掌握的词汇从古至今不会超过八千，是有限的。在语法方面，每种语言的句法结构大都在二百种左右，并可以用语言学的方法分析得到（萨丕尔，1985）。以上的讨论告诉我们，虽然从大脑神经的角度人类对语言的认知能力的研究还有很长的路要走，但人类语言的认识能力和极限已经勾画了出来。

3. 语言记录的相关理论

从语言文字的形式和语言认知的角度看，无论是什么文字形式，文字记录语言主要是将符号系统和语言的音位系统联系起来。从语音学和音位学层面看，虽然人类能分辨出上万种声音，也能将上千种声音用于语言的交际，但对于一种语言来说，只有大约几十个能用于语言的音位，这说明人类的大脑对语位的感知范畴是有限的。以汉语普通话为例，声韵调的组合可以构成 1200 个左右的音节，将声母（22 个）、韵母（38 个）和声调（4 个）组成一个三维空间，标出 1200 个坐标点就构成了汉语普通话最基本语音发音和感知空间。这个空间的理论坐标点为 3344 个，但实际只用了 1200 个左右，巨大的冗余度保证了汉语普通话交际的稳定性。如果将声韵调和基本词汇联系起来，其对立关系又体现出每个音位的结构对立负担量[7]。如果再将其放在大规模文本中，又能计算出汉语普通话的功能负担量（functional load）（王士元，1967）和信息熵（Information Entropy）（Shannon et al, 1949; Shannon, 1951）。语言信息和结构的研究表明，无论是什么语言，其信息结构具有共性， 如语言词汇的分布会遵循齐普夫定律（Zipf’s Law）（Zipf, 1935,1949）。以上的分析可以看出，语音学、音位学、文字学、语音感知空间、语言冗余度、语言对立的结构负担、语言功能负担量、信息论、文献统计分布定律等构成了精确记录和传承一种语言系统的理论基础，这就为全面记录和传承语言奠定了基础。

四、口传文化记录的理论基础

口传文化是有声语言的一种特殊形式，从功能上看，口传文化，如歌曲、戏曲等，都是在表达比有声语言更为强烈、独特和难以言表的情感，由于这些特殊的情感超出了语言表达的范围，因此需要用语言之外的表达方法。下面从口传文化的形式和认知范畴来讨论口传文化记录和传承的理论基础。

1. 口传文化记录的形式

在口传文化记录和传承的形式方面，除了需要语言文字记录语言内容外，其形式主要是曲谱，可以分为两类：1）文字谱，用语言文字记录口传文化的音乐旋律，例如碣石调最初是民歌，叫《陇西行》，后来用它唱曹操的诗《碣石篇》，就改叫《碣石调》，显然早期音乐的曲谱是用文字来记录的；2）符号谱，用符号来记录口传文化的音乐旋律，例如，五线谱、简谱、工尺谱等。在中国传统口传文化中影响最大和使用最普遍的符号谱是工尺谱，其符号来自汉字，而中国传统曲谱的记录介质主要是纸张。总之，从形式上看，口传文化的记录除了包含语言的记录形式外，还增加了口传文化音乐的记录形式。

2. 口传文化的认知基础

[6] “詞涯八千”的理论是台湾郑锦全教授经过多年对不同语言不同作者著作的统计，发现的人类语言认知在词汇上的极限，该理论的提出对认识人类的语言习得和认知能力有重要意义。.

[7] 见孔江平论文“藏缅语音位负担量计量研究”，该文将于 2013 年 8 月发表于王士元先生 80 寿辰纪念文集。

人类的语言有其认知的基础，并具有一定的民族性，同样口传文化也具有认知的基础和一定的民族性。在音乐层面，口传文化的性质更具有普遍意义。因此，一种口传文化除了民族性以外，人类对音乐的认知更为趋同。比如，音乐的旋律通常主要由基频决定，根据心理物理学的研究，人们对基频音高的感知是基本相同的。但现代声学的研究告诉我们，人们对音高的感知并不局限于基频这一种物理量，比如，泛音（overtone）同样能被感知为音高，形成音乐的旋律。这时旋律不是依据这个声音的基频，而是由声音某个较强的谐波决定，一个最典型的例子就是口弦琴。口弦琴是一种乐器和语言发音器官结合的口传文化，演奏者可以通过调节声道形状将能量集中在某个谐波上，从而产生一个很高的旋律。另外，共振峰同样能被感知为音高，从而成为声乐的旋律，最典型的例子是蒙古族的呼麦。人们感知到的呼麦旋律非常复杂，通常呼麦会被感知为两个旋律，一个是由声带产生的非常低的旋律，称为喉音唱法（throat singing）（Lindestad, 2001）通常人们的声带不可能产生这样低的振动，但呼麦是通过真声带和假声带同时振动，并在声带的开合上有一个相位差，从而产生一个低于原声带振动一倍的次谐波降低音高的感知，这种发声的类型被称为“喉室发声”（ventricular voice）或“双声带发声”。另一个旋律是由声道调节产生一个很强的共振峰而形成了一个远高于声带振动的旋律，也可以说是泛音的一种形式。从以上的分析可以看出，人们在感知旋律时非常复杂，并不是单一的物理量在起作用，同时我们也可以看到人们对声乐的认知基础和极限，在了解了声乐的认知基础和极限后，我们就能以此建立口传文化记录和传承的理论框架。

3. 口传文化记录的相关理论

在了解了人们对各种口传文化旋律认知的基础上，就可以讨论口传文化记录和传承的理论基础。我们知道旋律的感知基础是音高和节奏，而音高的物理量是基频、泛音以及共振峰的综合体。传统的记谱方法对于大多数口传文化来说是有效的，这使得我们能够传承古代遗留下来的许多文化遗产。但从现代科学的角度看，只用这些记谱方法记录丰富多彩的口传文化，并不是很精确也有很多局限性。人的音高和频率的感知关系可以用数字来精确描述，钢琴键盘体现了基频和音高之间的数学关系。

人对音高的感知也有范围和极限，最低在二十赫兹左右，最高在 20000 赫兹左右；颤音在一定的范围内人是感知不到，比如，声音以平均每秒 6～7 次的周律在半音限度内平稳地上下摆动形成颤音不会被察觉音高的变化，但音质会发生改变。传统的记谱方式很难对微小基频的变化所导致的音色改变去进行精确记录，例如昆曲通常也被称为“水磨腔”，但是古代和现代文献很少涉及水磨腔发音生理和声学的科学解释。从声学分析看，一定频率范围内的颤音是昆曲唱腔幽雅婉转、清丽悠远的主要声学特征，这些声学特征在生理上体现为发声类型的反复变化。显然普通的曲谱很难记录下这些信息。虽然我们传承下来了很多曲谱，但仍然很难想象出古人在吟唱时那优美的旋律和使用的嗓音发声类型。另外，中国的许多原生态民歌中存在大量的装饰音，但目前的许多记录方法主要是以骨干音为主，看看一首歌曲的基频曲线我们就知道丢掉了多少信息，虽然有些装饰音不能被明确感知出来，但它们对音色已经产生了微妙的改变。

目前人们对旋律、音色和声音物理量之间关系的了解还不是很深入，但在知道了声学的基本特性和声乐的感知基础后，就可以从理论上建立声乐记录的基本理论框架。从以上的分析可以看出，心理物理学、音乐声学、心理学等构成了声乐精确记录和传承的理论基础。在这个基础上，可以先利用现代科学技术，记录和传承那些我们目前还无法解释的现象和信号。

五、语言及口传文化记录的基本方法

传统的有声语言和口传文化记录方式主要是文字和乐谱，上个世纪中晚期人类又发明录音机、摄影机和录像机，从而用文字无法记录的信息被记录了下来。信号的存储方式也从纸张发展到模拟电子介质和数字电子介质。数字信号在传承过程中已经达到了一个极限。这样问题就集中在记录信号上，即用什么样的方

法采集什么样的信号才能精确和全面的记录语言和口传文化，而不用考虑承载形式和介质的问题。

1. 现有信号记录的缺陷

从方法上看，国际国内对有声语言和口传文化的保护基本是利用文字、语音和录像三种方法。现代科学的研究表明，简单地使用文字、录音和录像的方法会丢失很多信息。首先，前文讨论了文字记录语言和口传文化的性质，其中最主要的是在人们对事物认识不足的情况下不能精确和全面的记录语言和文化信息。其次，现在有了录音机和录像机是否就能精确记录语言和口传文化了呢？答案同样是否定。录音是将声学信号转变成数字信号记录和保存起来，的确能够记录语言和口传文化的大部分信息，但众所周知，听录音和听现场实际的声音有时会有很大的不同，特别是在听早期的录音时，会明显发现语言的辅音会混淆，辨别率较低。虽然现在录音机从早期的模拟录音机发展到高采样频率的数字录音机，已经到达了人类听觉的极限，解决了录音技术问题，但是录音还有其他的问题，例如，录音有频响的问题，因此听到的现场声音和回放的录音是有差别的，即使是微小的差别也会导致人对声音感知的错误。第三，关于视频信号，虽然现在的录像质量已经能达到高清，但和人现场的视觉感知还是存在差异，例如，一幅实际的场景由于物体的形状和光线不同会产生阴影，人眼能从阴影处提取很好的和有用的视频信息，但是现在高质量的录像机也无法象人眼的视网膜一样录下阴影处有用信息。另外，视频信号的采集还涉及到三维立体的问题，其中三维空间的信息会丢失的更多。总之拿录音机和录像机的性能和人的听觉和视觉认识能力相比，前者的性能还差得很远，因此我们说在现有技术下，文字、录音和录像都会丢失有用信息。另外，从人的认知角度看，我们感受到的世界也未必都是完全真实的。现代科学证明，不同的人感受同样的事物，结果可能不同。我们通常说“眼见为实耳听为虚”，但实际上眼见不一定就为实。认知科学的进步，为语言文化数字化传承带来了更广阔的视野。

2. 语言信息领域的新方法

在语音学、语言学、信息科学和神经科学领域，语言和口传文化的研究在近十年已经有了很大的发展，人们从语音的发音到声音的物理特性，从对语音信息的感知到大脑神经的活动，使用了许多新的研究方法和得到了很多语言信息传递的知识，下面对这些新的方法进行简要地介绍。

传统语音学和语言学的方法主要通过语音和语言学家听音，用国际音标记音记录有声语言和口传文化的语音、研究音位系统和创制文字系统来进行有声语言和口传文化的记录，例如，无文字语言的语音系统、语音和词汇的关系、口传文化的特殊发音方式等。这种方法是传统语音学、音位学、语言学和声乐学的研究方法，记录语言和声乐的田野调查工作是语音学和语言学家、语言人类学家和音乐学家的基本功。

实验语音学的方法主要是利用声学和生理学的一些方法来对有声语言的语音和口传文化的特殊发音方式进行录音，然后对语言的元音、辅音、声调以及韵律进行声学分析，对口传文化的特殊发音方式进行生理学的研究等，最终达到记录、研究、保护和传承的目的。

言语科学的方法主要是利用各种科学仪器对有声语言和口传文化进行各种信号的采集，然后进行信号处理，对提取出的参数建立数据库和进行各种计算。目前，国际和国内根据目的的不同，建立了大量有声语言的声学语料库和数据库，主要集中在语音的合成和识别方面。

模型化研究的保护方法主要根据对语音多模态的研究，提出模型化研究的保护方法。这种方法主要是对某一种有声语言或口传文化进行建模，并在建立相关声学和生理模型的基础上对有声语言和口传文化进行有效的保护和传承。

很显然，随着时代和科技的进步，记录有声语言和口传文化的介质、记录形式和记录方法都在不断地改进，为中国和世界有声语言和口传文化的数字化传承奠定了基础。

3. 言语链和信号的不可逆性

语言的交际过程是一个链条，称之为言语链（邓斯，1963），它从一个人的大脑思维活动、神经驱动言语生理肌肉运动、产生声波到听觉声学神经转换、听话人的大脑感知及思维活动，构成了语言传递和交际的一个链条，在这个链条中都包含有语言的信息。但是语言链条之间的转换并不是完全可逆的，或者说目前我们的科学技术还没有找到完全可逆的方法，这样对我们来说就是未知的，如果我们只记录言语链的一个环节，就有可能丢失许多信息，不能精确和完全记录语言，也就无法有效地传承。如，声道和语音共振峰之间的关系不是一对一的，而是一对多的关系，因此，从共振峰很难精确推出声道的形状，目前从语音中还不能很好地计算出声带的振动方式，如果不记录声带振动的方式，只从声音很难了解一种语言的发声类型。呼吸是语言的动力来源，它同时还能反映出语言交际时的情感和规划，这些信息很难从语音声波中得到。另外，言语者当时的真实情感，单从声音很难得到真实的情感信息。

4. 基于语音多模态的数字化方法

根据本文的讨论和作者在这个领域的长期研究和实践，从信号的角度提出一种适合目前技术水平的有声语言和口传文化的数字化方法。第一是语音信号，采用指向性强的高保真话筒，话筒的频响性能要好。采样频率每秒不小于 22050，量化至少用 16 比特，录音环境的本底噪音要保持在 20 分贝左右；第二是嗓音信号，这种信号要用喉头仪来采集，它对采集声带振动信息非常有用，特别是声带闭合相的振动信息，对于保留语言发声类型的信息十分有用；第三是录像信号，对于有声语言来说，唇形和面部的动作是十分有用的信息，众所周知，言语的交际信息不只是包含在语音信号中，唇形也包含有语言交际的信息，McGurk 效应已充分证明了这一点（McGurk, H. et al，1976）。第四种是电子腭位信号，这种信号通过一个电子腭采集舌和上腭的接触信息，对于研究和记录言语辅音的发音动作信息十分重要。第五是呼吸信号，呼吸信号包括腹呼吸和胸呼吸两种信号，因为目前的研究表明腹呼吸和胸呼吸在言语中常常不同，腹呼吸和言语动力更密切，而胸呼吸和发音更密切，这对研究和记录语言的韵律及风格十分有用；第六种是心率信号，这是一种心理信号，它能反映说话人的心理状况；第七种是指电压信号，这也是一种心里信号，指电压比较敏感，可以反映心情的变化状况。实际上心理信号也可以录制脑电信号（ERP），因为脑电信号的时域精度较高，但相对其它信号的录制比较麻烦，需要有很多通道，因此，心理信号还需要进一步研究。前五种信号对于一种有声语言的信息录制、研究和传承都是必不可少的，因为这些信号可以基本建立一种语言的声学和生理模型，从而达到比较精确和全面记录一种有声语言和口传文化的目的。另外，录制有声语言还有其它一些仪器可以利用，如超声仪、电磁发音仪、磁共振等都可以得到言语的有用信息，而且这些信息都是从语音中无法得到的，但录制设备和技术上比较麻烦，目前不容易大量录制。

六、语言及口传文化记录和传承的理论思考

随着信息科学技术的发展，语言学和信息科学的关系越来越紧密，研究方法也越来越自然科学化。语言学和信息科学新的理论方法促使人们对语言和口传文化的数字化传承产生了许多新的思考，本节就语言和口传文化记录和传承的理论问题提出几点思考。

1. 人类知识体系是自然生物进化的高级形式

从十九世纪达尔文发表《物种起源》开始[8]，人类对自然物种的起源和进化进入了科学的研究轨道，经过一百多年的发展，人们建立起了科学的物种起源和进化理论。在研究方法上，已经从物种的考古、体质分类发展到了基因分析和基因库的层面。然而人类对人类本身语言和知识体系的起源和进化却知之甚少。生

[8] 达尔文的原文为 The Origin of Species，最早于 1859 年发表于英国伦敦。该书的出版对人类和其他物种起源的科学研究起了巨大的推动作用。

物的进化可以通过考古和基因图谱来追寻史前的轨迹，但人类史前知识体系的研究却很难从考古中找到线索和获取信息。虽然古人类学家已经发现了许多古人类的化石，但大多残缺不全，很难找到足够的信息来追寻发音器官和大脑语言起源和进化的轨迹。就更不用说建立史前人类知识体系起源和演化的理论框架了。然而这并不是最重要的，因为，目前世界上还有上千种活的语言，如同活的生物，这种存在于人大脑中的活的知识体系是伴随着人类长期进化发展起来的，人类从走出非洲就开始了民族、语言和文化的多样性分化，但人类的知识体系的多样性并没有得到应有的重视，由于人的原因，自然界的物种在急速消亡，人的语言和知识体系也在急速消亡。目前我们还不太清楚语言和人类知识体系从多样性分化转向快速消亡的时间分界点是从何时开始，但语言文化和人类知识体系的快速消亡已经是不争的事实。因此提高对人类语言和知识体系的科学研究和认识，将其看做是自然生物演化的高级形式是记录、保护和传承人类文明的重要前提。

2. 人类知识体系的理论极限

人类的语言和文化构成了人类的知识体系，从历史上看人类记录和传承知识体系存在许多缺陷，随着信息科学技术的进步，这些缺陷在不断得到弥补，那么从理论上记录和传承人类知识体系的极限在哪里？也就是说从理论上怎样才算完全精确记录和传承了人类的知识体系呢？由于人类的知识存在于语言和口传文化中，所以语言科学家更关心这个问题。从理论上讲，语言的交际和知识传递是通过语言链完成的，在早期的语言链研究中，人们只能认识到语言的发音器官动作和区别语言的最小单位“音位”。随着科学方法的进展，人们对发音器官的生理机制、物理声学性质、音位的听觉感知范畴以及脑电活动有了更新的认识。然而在大脑的语言记忆、感念形成、思维活动等方面，我们还知之甚少。虽然我们已经对人类的语言和认知能力有了新的认识，但我们还不完全了解构成人类知识体系的生物基础。在言语链中，言语者的大脑思维活动为最底层，然后通过一系列神经活动和生理物理活动达到语言的表层“言语声波”，然后，从言语声波又经过一系列的听觉神经活动进入大脑底层的思维活动。从历史上看，人对语言和口传文化的记录和传承都是在语言的表层，即语音的层面。上个世纪中期以后，人们开始记录、研究和传承语言和口传文化的更底层的信息，如，发音器官的活动、声带的振动信号、语音听觉的神经信号等，并开始建立语言的生理、物理和听觉模型。这些模型是精确和全面记数字化传承语言和文化的基础。由于语言某些表层结构和深层结构转化之间存在不可逆性，通过模型记录和传承语言更底层的信号和信息也许是我们到达精确记录语言文化和人类知识体系的未来形式。研究表明，语言表层有其理论极限，这个极限可以通过语音学、语言学、语法学等来勾画出语言大致的表层认知范围和极限。语言的生理、物理和听觉平面可以通过言语生理、声学和听觉研究来找出人类语言较深层的认知范围和极限，这两个层面的极限我们已经有了很多知识。但在最底层，大脑思维活动的极限可能还需要很长的时间去探索。因此，回到语言文化和人类知识体系的数字化传承上，目前我们能做的有：1）文字符号记录语言表层信息（有声语言）；2）物理信号记录语言表层信息（音频和视频信号等）；3）生理信号记录较深层的语言发音信息（X 光、电子腭位、超声、磁共振等）；4）生理信号记录较表层的语言情感信息（心率、指电压等）；5）生理信号记录较深层的听觉信息（ERP 等）； 6）生理信号记录较深层的语言理解信息（眼动、ERP、脑磁、磁共振等）。这些语言不同层面信号的研究，一定能使我们逐步认识到语言和人类知识体系的最终极限，从而找到全面精确数字化传承人类知识的方法。

3. 中国的语言和口传文化保护

中国是一个多民族和多语言的国家，语言有五个语系一百多种语言，经语言学家确认的也有八十几种。众多的语言和丰富多彩口传文化构成了中华民族独有的知识体系，是人类不可多得知识宝库。然而，随着全球经济一体化的进展，语言文化和认知的多样性同自然物种一样，生存受到了极大威胁，面临濒危的境地。一些民族的语言一旦消亡，经过千万年和人类

生理一同进化形成的知识体系也会随之消亡。因此中国在记录和传承语言和口传文化方面，面临更巨大和艰苦工作。

中国是一个注重文化保护和传承的国家，文化典籍的整理和编纂是我们的文化传统。在历史上，我们有明代的“永乐大典”和清代的“四库全书”，这两部大典在保护和传承中华民族文化和知识体系方面起到了巨大的作用。在现代这个民族和语言大融合的时代，我们有责任利用基于现代语言信息技术的数字化传承的理论与方法记录中国的有声语言和口传文化，将其编纂为“中华有声语言和口传文化数字化多媒体大典”，并传承下去。

参考文献

1. 《中国传统音乐乐谱学》，王耀华，2006 年，福建教育出版社，ISBN： 978-7-5334-4340-5

2. 《言语链：说和听的科学》，邓斯和平森，1983，曹剑芬，任宏谟译，中国社会科学出版社。

3. 《世界的书面语·中国卷》，1995，谭克让，格兰特 D. 麦克康奈尔主编，拉瓦尔大学出版社，魁北克。

4. 《中国戏曲音乐集成》，《中国戏曲音乐集成》编辑委员会，中国 ISBN 中心，2002。

5. 《语言论》，1985，爱德华·萨丕尔著 陆卓元译 陆志韦校，商务印书馆。

6. The Sounds of the World’s Languages, Peter Ladefoged and Ian Maddieson, 1996, Wiley-Blackwell.

7. The Measurement of Functional Load, 1967, William S-Y. Wang, Phonetic, Vol.16. 36-54.

8. The mathematical theory of communication, 1949, Shannon, C.E. and Weaver, W.:, University of Illinois Press, Urbana.

9. Prediction and entropy of printed English, 1951, Shannon, C.E.: Bell System Technical Journal, 30: 50—64.

10. The Psychobiology of Language, 1935, George K. Zipf, Houghton-Mifflin.

11. Human Behavior and the Principle of Least Effort, 1949, George K. Zipf, Addison -Wesley.

12. McGurk, H. and J. MacDonald (1976). "Hearing lips and seeing voices." Nature 264: 746-748.

13. Voice Source Characteristics in Mongolian Throat Singing Studied with High-Speed Imaging Technique, Acoustic Spectra, and Inverse Filtering, 2001,Per-Åke Lindestad, Maria Södersten, Björn Merker, and Svante Granqvist, Journal of Voice,Vol. 15, No. 1, pp. 78-85, The Voice Foundation.

On the Theory and Digital Method of Speech and Oral Culture Inheritance

Kong Jiangping

(Department of Chinese Language and Literature, Research Center for Chinese Linguistics, Peking University)

Abstract: Speech and oral culture are the most important parts of human civilization. Their methods of inheritance have been developing along with the development of society and technology. The diversities of speeches and oral cultures in the world are endangered with the advance of globalization. This paper discusses the basic method of speech and oral culture inheritance after the analysis on the inaccuracy and imperfect records of traditional writing, audio and video and then proposes an inheritance method based on multi-speech models. Finally, the theoretical basis of speech and oral culture inheritance is profoundly discussed form the view of speech and oral culture cognitions.

Key Words: speech; oral culture; digital method; cultural inheritance

普通话双音节V1n#C2V2 音节间逆向协同发音[1]

李英浩 [1]，孔江平 [2]

（1. 延边大学 外国语学院，延吉 133002；2. 北京大学 中国语言文学系，北京 100871）

摘要：本文使用动态电子腭位研究普通话V1n#C2V2双音节中后续音节C2和V2对前鼻韵尾/n/以及V1的逆向协同发音影响。实验结果发现：(1) C2舌前音和舌面后塞音决定前鼻尾的发音部位和舌形姿态，一般不存在V2对V1的逆向影响。(2) C2为双唇音和舌尖中音条件下，存在V2对V1n的逆向影响，前者表现为从V1后过渡段开始舌体动作向V2过渡，后者只出现在鼻音时段后部。(3) C2为舌面后擦音和唇齿擦音的时候，前鼻尾发音部位一般不同化或部分同化，这可能与发音人的发音策略有关。上述结果支持普通话辅音的协同发音阻力等级序列，同时能够为建立普通话发音生理模型奠定理论基础。

关键词：动态电子腭位；前鼻韵尾；逆向协同发音

中图分类号：H017

Anticipatory Coarticulation in V1n#C2V2 Sequences in Standard Chinese

LI Yinghao[1], KONG Jiangping[2]

(1 Foreign Languages School, Yanbian University, Yanji 133002, China; 2 Department of Chinese Language and Literature, Peking University, Beijing 100871, China)

Abstract: This paper studies the anticipatory coarticulation of C2 and V2 on alveolar nasal coda /n/ and V1 in V1n#C2V2 sequences in the Standard Chinese. The results show: (1) the articulatory place and tongue gesture of the alveolar nasal coda is determined by the following C2 with the anterior portion of tongue and tongue back being the primary articulator. The transconsonantal vocalic influence is not observed. (2) V2 affects the V1n when C2 is bilabial or alveolar consonants. In the former case, the tongue body gesture undergoes a smooth transition from V1 offglide to V2; and in the latter case, the V2 effect occurs in the offglide of nasal coda. (3) When the following C2 is velar or labiodental fricative, partial place assimilation or complete alveolar closure is found in the majority of tokens, indicative of the articulatory strategy used by the subject. The above results support the scale of coarticulation resistance for consonants and lay theoretic foundations for the articulatory modeling of the Standard Chinese.

Key words: electropalatography; alveolar nasal coda; anticipatory coarticulation

对语流中前鼻韵尾/n/的研究在言语产生和言语工程中都具有重要地位。前鼻尾在声学和听感上一般表现为三类：纯鼻音、元音鼻化以及鼻尾完全丢失。[1][2] 在汉语普通话以及方言中，前鼻尾比后鼻尾/ŋ/更易于脱落。[3] 在发音生理方面，前鼻尾/n/的发音动作包括软腭下垂打开鼻腔通路和舌尖上抬形成齿龈阻。其中，软腭下垂动作相对稳定，但是口腔动作易受语音环境的影响，[4] 其受影响的程度大于同部位的塞音。[5]

语流中前鼻尾受后接音节声母影响易于产生部位同化的现象。对英语/n#k/辅音丛的EPG(Electropalatography)研究结果发现，前鼻尾的口腔动作有三种类型：未同化的鼻音、部分同化的鼻音和完全同化的鼻音。该研究表明部位同化并非是范畴化的音系过程。[6] 使用EPG对普通话V1n#C2V2的研究发现，C2为舌面后音时，前鼻尾发生部分部位同化；C2为舌前音时，前鼻尾的舌形姿态与C2叠加。[7] 基于汉语大语料库的研究还发现，语流中鼻尾脱落与后续音节声母的发音方式有关。[2] 上述研究结果说明前鼻尾发音动作的实现以及声学特征在很大程度上取决于后续C2的类型。然而，对C2影响前鼻尾发音动作的机制以及声学表现的认识仍不全面。

本文使用EPG分析普通话双音节V1n#C2V2 (#为音节边界)中后续音节的C2和V2对前字鼻尾/n/的逆向协同发音影响。本文还考察C2和V2对V1的F2轨迹的逆向协同发音影响。最后，本文从辅音协同发音阻力(Coarticulation Resistance，简称CR)的角度来阐述V1n#C2V2中音节间的逆向协同发音过程和机制。

1 方法

1.1 语料

选用北京大学语音学实验室的普通话动态电子

[1] 本文已发表于《清华大学学报（自然科学版）》。

腭位数据库中男性发音人的V1n#C2V2双音节样本，共585个，双音节语音构成见表1。多数双音节为普通话中的字典词或者短语，少部分为没有意义的双音节组合。

表1 双音节的语音构成

双音节前字音节		双音节后字音节	
C1	V1n	C2	V2
唇音	/an, ən/	唇音(LA): b p m f 舌尖前音(AP): z c s	/i, u, a, (f)ei/ /i1, u, a/
舌面后音	/un/	舌尖中音(AL): d t n l 舌尖后音(RE): zh ch sh r	/i, u, a, (n,l)y/ /i2, u, a/
舌面前音	/in, yn/	舌面前音(PA): j q x 舌面后音(VA): g k h	/i, iu, ia, y/ /ei, u, a/

注：V2列中/i1/为舌尖前元音，/i2/为舌尖后元音。

1.2 信号采集和处理

通过提词器屏幕把语料呈现给男性发音人(27岁，前北京大学电视台播音员)。正式录音前，发音人佩戴特制的62电极假腭进行30分钟左右的语音适应训练。实验开始后发音人用正常语速朗读双音节语料，语料之间停顿1秒左右。

使用WinEPG动态电子腭位仪同步采集发音人的EPG信号(采样频率为100Hz)，声学信号和喉头仪信号(EGG)（采样频率为44.1kHz)。三路信号在EPGAnalyzer(Ver. 1.0)分析平台上进行预处理，使用EPG/语音对齐算法对信号进行时域调整，[8] 然后根据EPG关键帧，EGG信号和语图进行半自动语音标记。

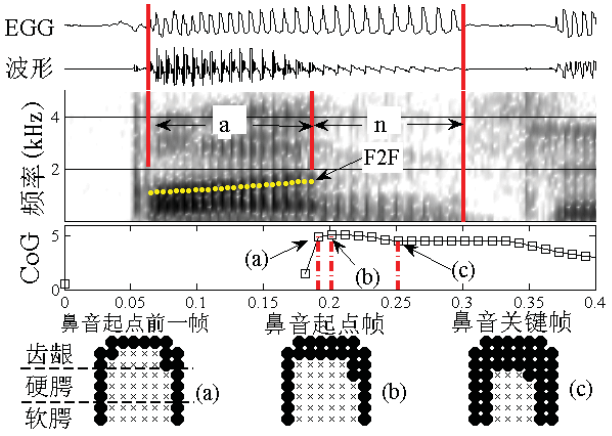


图 1 语音标记、EPG腭位帧和部分参数(图中最下行为 3 帧腭位，黑点表示舌腭有接触，‘X’表示没有接触。最左侧腭位标明本文假腭的生理分区。)

图1为三路信号以及部分参数的示意图。V1的起点根据EGG信号自动标记。V1声学结束点(即前鼻尾声学起点)由语图和/或EPG确定，C2为塞(擦)音和边音的时候，一般位于舌腭完全成阻的第一帧之后或者语图上元音和鼻音断层位置之后；C2为擦音的时候，如果V1和鼻音之间的断层不明显，则基于EPG和其他声学特征进行判断(声学特征参见[1]和[3])。使用LPC协方差方法获得共振峰原始数据，而后对V1的F2轨迹进行手动修正。

假腭分为三个区域，前三行电极为齿龈区，对应齿龈脊到齿龈后区域；中间三行为硬腭区，对应硬腭前和硬腭区域；最后两行为软腭区，对应硬腭后到软腭前区域(参见图1腭位帧(a))。鼻音关键帧定义为前鼻尾声学时段中点附近3帧腭位中接触面积最大的一帧。选择这个位置有两个原因，第一，此刻的舌腭接触要么相对稳定，要么表现为向后音节音段发音动作过渡。第二，V2的隔位逆向影响表现较为明显(参见图1的腭位帧(c))。

1.3 参数定义

我们使用接触重心指数(Center of Gravity，简称CoG)和接触趋中指数(Contact Centrality，简称CC)分别描写舌腭收紧点的部位特征以及舌腭接触是否形成阻塞。两个参数的计算方法见[9]。CoG与舌腭接触收紧点部位有关，收紧点位置越靠前，CoG越大；收紧点纵向接触宽度越大，CoG则降低。CC与舌腭收紧部位的收紧程度有关，如舌前塞音的CC一般比同部位擦音的CC要大。两个参数与发音事件存在一定的对应关系。通过考察585个鼻音关键帧的两个腭位参数的数据分布和与发音生理的对应关系，发现CoG数值与收紧点位置之间的关系为：在0~2之间为软腭区，在2~4.2之间为硬腭区，在4.2~7.5之间为齿龈区；CC数值与舌腭是否形成阻塞的对应关系为：CC在0~0.76之间说明未形成阻塞(未封闭区)，在0.76~1之间说明形成阻塞(封闭区)。需要说明的是，因为EPG不能完整反映软腭塞音的阻塞情况，因此软腭塞音的CC均小于0.76。

使用后腭接触面积(POS)分析V2对鼻音关键帧腭位的影响。POS为假腭后四行电极的接触电极数与后四行电极总数的比率。

声学分析使用F2音轨方程。音轨方程由四个参量构成：斜率(Slope)，截距(Intercept)，确定系数R^2，显著性水平p。斜率可以用来表征辅音发音部位以及受元音协同发音影响的程度。[10] 音轨方程的计算方法为：固定C2，以5个V1的F2轨迹的中间点的F2值作为自变量，以V1的F2轨迹的结束点值(称为F2F)为因变量，进行线性回归分析。

选用F2F分析V2对V1的逆向影响。由普通话V1#C2V2的音节间跨音段逆向协同发音的结果可推知，[11] 如果V2能影响前鼻尾时段的腭位，则F2F也有可能受到V2的逆向影响。

2 结果

2.1 C2 对前鼻尾/n/的逆向影响

以C2的发音部位为分类变量，分别计算五种V1条件下鼻音关键帧两个腭位参数的均值。图2为105种

V1n和C2组合条件下CoG和CC两个参数的分布情况。可以看出，C2为舌前音(包括三组舌尖音和舌面前音)条件下，鼻音关键帧腭位与后接C2的发音部位和舌形姿态具有一致性。后接舌尖前音和舌尖后音的前鼻尾的成阻部位分别处于齿龈区和硬腭区；后接舌尖中音时候的成阻部位在齿龈区；后接舌面前音时候的成阻部位分布在齿龈区和硬腭区。C2为塞(擦)音的时候，前鼻尾的一般形成阻塞；C2为擦音的时候，前鼻尾的一般不形成阻塞。这说明C2不仅影响前鼻尾的成阻部位，也影响它的舌形姿态。

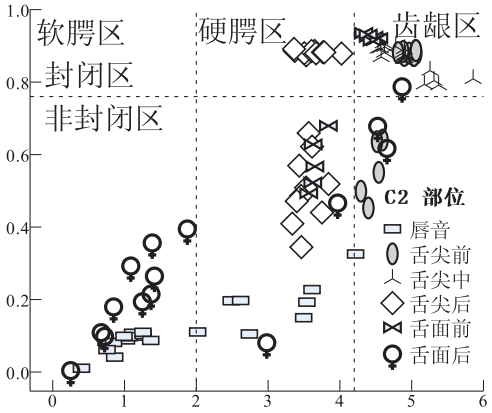


图 2 五种 V1 条件下鼻音关键帧腭位参数均值的分布(横轴表示 CoG，纵轴表示 CC。)

C2为舌面后音和唇音条件下的前鼻尾腭位参数分布范围较大。后接舌面后音的前鼻尾腭位多分布于软腭非封闭区，说明前鼻尾同化为后鼻尾；少数情况下出现在硬腭非封闭区、齿龈封闭区及非封闭区，说明前鼻尾未同化或者部分同化。图3为后接三个舌面后音的鼻音关键帧舌腭接触频率(不考虑V1和V2)。可以看出，C2为舌面后塞音的时候，软腭区电极的接触频率最高，硬腭区的接触频率相对较低，只有一个样本(C2=g)的舌腭接触延伸到齿龈区。这说明在后接舌面后塞音条件下，前鼻尾一般实现为后鼻尾。后续C2为舌面后擦音的情况则不同：37%的样本形成齿龈阻，52%的样本在齿龈区有电极接触，但没有形成阻塞，2个样本的舌腭接触延伸到硬腭区，1个样本没有舌腭接触。

V1n#C2V2(C2=g)

	0	0	3	0	0	0			
0	3	0	0	0	0	3	0		
3	0	0	0	0	0	0	3		
3	0	0	0	0	0	0	3		
18	0	0	0	0	0	0	0	25	
51	14	0	0	0	0	0	7	51	
81	29	0	0	0	0	0	44	88	
77	88	7	0	0	29	96	100		

V1n#C2V2(C2=k)

	0	0	0	0	0	0			
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	14	
53	3	0	0	0	0	0	0	53	
67	21	0	0	0	0	32	71		
71	71	0	0	0	21	71	89		

V1n#C2V2(C2=h)

	59	59	51	44	40	51			
81	66	29	25	40	37	55	77		
81	55	14	3	7	29	59	88		
74	40	7	0	0	11	59	81		
62	11	0	0	0	0	51	77		
85	14	0	0	0	0	18	96		
77	3	0	0	0	0	3	92		
74	7	0	0	0	0	3	77		

图 3 C2 为三个舌面后音时关键帧接触频率(%)

后接唇音的前鼻尾一般不形成齿龈阻，收紧点分布于三个区域。通过对样本的观察发现，腭位参数的分布因C2发音方式以及V2/V1的不同而不同。后接双唇塞音和鼻音的时候，V1或V2为/i/的时候鼻音关键帧在硬腭区和/或软腭区出现舌腭接触；在时域上表现为舌体动作从V1向V2过渡。后接唇齿擦音的时候，86%的样本在齿龈区有电极接触(包括一个样本形成齿龈阻)，两个样本的舌腭接触延伸到硬腭区，只有两个样本没有舌腭接触(V1=V2=/a/)。同舌面后擦音的情况一样，这有可能与发音人的发音策略有关。

2.2 V2 对前鼻尾/n/的逆向影响

为考察V2对前鼻尾腭位的逆向影响，只比较V2为非低元音/i,i1,i2,ei/和低元音/a/(舌面前音时为/ia/)条件下鼻音关键帧的POS是否有显著差异。对四组舌前音以V1和C2发音部位的组合作为分类变量，分别对两种V2条件下的鼻音POS进行独立样本*t*检验。结果发现，除舌尖中音外，其他舌前音条件下鼻音的POS在不同V2条件下没有显著差异。C2为舌尖中音的时候，在V1/u/的情况下发现显著差异；在其他V1条件下，非低元音V2情况下POS的均值全部大于V2低元音的情况。进一步分析相同条件下鼻音结束帧腭位，结果发现V2对该帧腭位的POS有显著影响。这说明后接舌尖中音的时候，V2对鼻音时段的舌体动作产生逆向影响，越接近鼻音结束点，V2的影响就越明显。

对唇音和舌面后音以V1和C2的组合作为分类变量进行独立样本*t*检验。结果发现C2为双唇塞音和双唇鼻音的时候，两个V2条件下鼻尾腭位的POS存在显著差异，这与2.1的结果一致。

2.3 C2和V2对V1的F2轨迹的影响

使用2.2的方法分析V2对F2F的逆向影响。独立样本*t*检验的结果表明，后接舌尖前音、舌尖后音和舌面前音条件下，V2对F2F没有显著影响。C2为舌尖中音的条件下，除了V1为/a/的时候，V2对F2F没有显著影响。结合发音生理的结果，可以看出C2为舌尖中音条件下，V2/i/舌体动作对V1后过渡段的影响相对较弱。C2为双唇音的条件下，除了V1为/u/的时候，V2对F2F均有显著影响。C2为舌面后音的时候，只分析完全部位同化的样本。结果发现，V2对F2F没有显著影响。

图4为21个声母的F2音轨方程的斜率和截距的分布情况(双唇塞音和双唇鼻音按V2/i/和/a/分别计算)。可以发现：(1) 声母的音轨方程的斜率和截距成反比。斜率数值上的排列为：舌面前音、舌尖后音和舌尖前音(0.35~0.5)<舌尖中音(0.5~0.6), 舌面后擦音(0.53)、唇齿擦音(0.61)<舌面后塞音(0.87~0.96)。而截距的大小与斜率的排列正好相反。(2) 舌面后擦音的斜率和截距与同部位的两个塞音明显不同，这与前文发音生理分析结果一致。(3) 唇齿擦音的斜率和截距接近舌尖中音，这与前文的生理分析结果一致。(4) 双唇音条件下V2能够影响V1后过渡段的腭位以及F2轨迹，因此，在图4中可以发现V2/i/和V2/a/

对音轨方程的斜率和截距有明显的影响。V2/i/条件下，V1的舌体动作由V1向V2过渡；而在V2/a/条件下，虽然鼻音时段很少有舌腭接触，但是舌前部分有可能呈现出上抬的动作趋势，[7] 因此斜率和截距接近舌尖中音的情况。

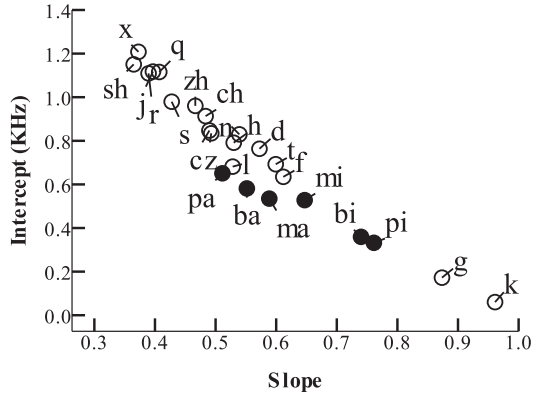


图 4　V1n#C2(V2)双音节条件下C2 音轨方程的斜率和截距分布(R^2 在 0.73 和 0.93 之间，p<0.001。图中黑点为两种V2条件下双唇音的两个参量的分布情况。)

3　讨论和结论

本文从发音生理和声学两个角度分析了普通话V1n#C2V2双音节音节间的逆向协同发音现象。主要研究结果为：(1) C2舌前音和舌面后塞音决定前鼻尾的发音部位和舌形姿态，一般不存在V2对V1的逆向影响。(2) C2为双唇音和舌尖中音条件下，存在V2对V1n的逆向影响，前者表现为从V1后过渡段开始舌体动作向V2过渡，后者只出现在鼻音时段后部。(3) C2为舌面后擦音和唇齿擦音的时候，前鼻尾多未同化或部分同化，这可能与发音人的发音策略有关。

上述结果证明辅音的协同发音阻力(CR)是制约普通话V1n#C2V2中音节间逆向协同发音的重要因素。协同发音阻力与音段对舌体(tongue dorsum)发音动作的控制密切相关，如果音段产生过程中舌体的发音动作受到限制，则该音段的发音动作不易受语音环境的影响，反过来，它会对周边音段的发音动作产生较强的协同发音影响。普通话辅音音段CR的研究表明，舌尖前音、舌尖后音和舌面前音对舌体动作的限制较高，舌形动作受周边元音影响较弱，因而CR较大。[12] 前鼻音/n/对舌体动作的控制较弱，[13] 且在音节尾的位置上其口腔动作更易于弱化。[14] 因此，当前鼻尾后接这三个部位的声母的时候，后续声母发音部位和舌形特征逆向影响前鼻尾，同时，V2对V1的逆向影响被阻断。舌尖中音的CR小于上述三组声母，其主发音器官为舌尖，舌体易于受到周边元音的影响，从文中结果可以看到，V2/i/可以影响鼻尾时段的腭位，但是这种影响对V1后过渡段的影响较弱。双唇音没有舌体动作要求，因而CR最小，在多数情况下鼻音段的发音动作表现为从V1向V2的过渡。舌面后塞音的主发音器官是舌体，其收紧点的位置受周边元音的影响，因此，前鼻尾时段的舌体动作一般表现为从V1向后字音节舌体动作过渡。

参 考 文 献

[1] 许毅. 音节和音联 [C] // 吴宗济，林茂灿. 实验语音学概要. 北京: 高等教育出版社, 1989: 193-220.
XU Yi. Syllable and Juncture [C] // WU Zongji, LIN Maocan. A Prime of Experimental Phonetics, Beijing: Higher Education Press, 1989: 193-220. (in Chinese).

[2] 方强. 连续语流中的鼻尾丢失 [C] // 语音研究报告, 2004: 91-97.
FANG Qiang. The acoustic analysis of nasal codas in connected Chinese speech [C] // Report of Phonetic Research, 2004: 91-97. (in Chinese).

[3] Chen M Y. Acoustic analysis of simple vowels preceding a nasal in Standard Chinese [J]. *J. Phonetics*. 2000, **28**(1): 43-67.

[4] Trigo Ferre R L. The phonological derivation and behavior of nasal glides [D]. Boston, MIT, 1998.

[5] Hardcastle W J. Assimilation of alveolar stops and nasals in connected speech [C] // Levis J W. Studies in General and English Phonetics: Essays in Honor of Professor J D O'Connor. London: Routledge, 1995: 49-67.

[6] Ellis L, Hardcastle W J. Categorical and gradient properties of assimilation in alveolar to velar sequences: evidence from EPG and EMA data [J]. *J. Phonetics*. 2002, **30**(3): 373-396.

[7] 郑玉玲, 刘佳. 普通话N1C2(C#C)协同发音的声学模式 [J]. 南京师范大学文学院学报, 2005, **3**, 150-157.
ZHENG Yuling, LIU Jia. The phonetic mode of concord articulation in N1C2(C#C) in Mandarin Chinese [J]. *Journal of School of Chinese Language and Culture Nanjing Normal University*, 2005, **3**: 150-157. (in Chinese).

[8] Li Yinghao, Pan Xiaosheng. Temporal alignment algorithm for electropalatographic and acoustic signals in long utterances [C] // Proc. 3rd IEEE/IET ICALIP. Washington, DC: IEEE Computer Society, 2012: 957-960.

[9] Fontdevila J, Pallares M D, Recasens D. The contact index method of electropalatographic data reduction [J]. *J. Phonetics*, 1994, **22**(2): 141-154.

[10] Krull D. Consonant-vowel coarticulation in spontaneous speech and reference words [C] // Papers in Linguistics from the University of Stockholm, 1989, **10**: 101-105.

[11] 李英浩, 孔江平. 普通话双音节V1#C2V2音节间的逆向协同发音 [J]. 清华大学学报(自然科学版), 2011, **51**(9): 1220-1225.
LI Yinghao, KONG Jiangping. Anticipatory coarticulation in V1#C2V2 sequences in Standard Chinese [J]. *J.Tsinghua Univ (Sci and Tech)*, 2011, **51**(9): 1220-1225. (in Chinese).

[12] LI Yinghao, ZHANG Jinghua, KONG Jiangping. The coarticulation resistance of consonants in Standard Chinese-An electropalatographic and acoustic study [C] // Proc. 8th ISCSLP. Washington DC: IEEE Computer Society, 2012: 454-458.

[13] Recasens D, Pallares M D, Fontdevila J. A model of lingual coarticulation based on articulation constraint. *J. Acoust. Soc. Am.,* 1997, **102**(1): 544-561.

[14] Krakow R R. Physiological organization of syllables: a review [J]. *J.Phonetics*, 1999, **27**(1): 23-54.

言语产生的胸腹呼吸机制[1]

杨锋[1]，侯兴泉[2]，孔江平[3]

（1. 暨南大学 华文学院，广州 510610；
2. 暨南大学 中文系，广州 510632；3. 北京大学 中国语言文学系，北京 100871）

摘要：本文通过胸呼吸和腹呼吸信号研究言语产生的胸腹呼吸机制，以及呼吸与韵律间的关系。实验结果表明：言语以胸腹联合式呼吸为主，腹呼吸重置时间早于胸呼吸重置时间和语音起始时间，呼气相时长约等于语音时长。吟诵所需气息量比朗读大。胸呼吸主要作用是保证足够的气息量，在发音时胸腔保持扩张状态至发音结束。腹呼吸主要作用是，通过腹肌和膈肌稳健收缩，控制气流持续释放，以获得连续的语音。韵律句起始处对应一个胸腹呼吸重置，韵律短语边界对应胸腹呼吸间断。本项研究对言语呼吸生理机制的认识和理解具有重要意义，为言语产生呼吸生理建模提供研究基础。

关键词：呼吸机制；言语产生；胸腹呼吸

中图分类号：H017

The chest and belly breathing control in speech production

YANG Feng[1], HOU Xingquan[2], KONG Jiangping[3]

(1. College of Chinese Language and Culture, Jinan University, Guangzhou 510610, China
2.Department of Chinese Language and Literature,Jinan University,Guangzhou,510632, China
3. Department of Chinese Language and Literature, Peking University, Beijing, 100871, China)

Abstract: The paper studies the chest and belly breathing control in speech production and its relationship with the speech prosody. The results show that Chest-belly compound breathing is manifested in normal speech, with the belly breathing reset preceding the chest reset and start of speech signal, and the length of expiration phase relatively equaling that of speech signal. Poem chanting requires more air volume consumption than in normal speech. The function for the chest breathing is to provide sufficient air volume for the maintenance of the thoracic cavity expansion till the end of utterance. The function for belly breathing controls the release rate for the chest breathing through the contraction of the abdominal muscle and diaphragm, thus securing the continuous speech flow. Immediately before a prosodic sentence, a reset is found for both chest and belly breathing. Immediately before a prosodic phrase, the belly and Chest breathing signal trough will be found. This research is of great signficance to understand the respiratory mechanism in speech production, and will lay foundation for the articulatory modelling of respiration.

Key words: breathing control ,speech production; Chest and belly breathing

呼吸是人类赖以生存的基础，同时也是言语产生的动力源，是言语产生过程的首要环节。呼吸机制的研究最早始于医学病理研究，后来逐渐应用到语言学研究中。在人类语言的进化过程中，对呼吸器官的灵活控制，使得我们能够发出持续的语音[1-3]。通过调节声门下压实现重音、声调和语调。参与呼吸运动的主要器官是肋间内肌、膈肌和腹肌[4-6]。呼吸重置往往发生在语法边界上[7]。

国内吴宗济等最早利用气流气压信号对汉语进行研究[8]。谭晶晶、张锦玉等分别对汉语普通话不同文体朗读时的呼吸重置特点进行了研究[9, 10]。

本文主要通过同步的胸呼吸、腹呼吸、语音和嗓音信号研究言语产生的胸腹呼吸机制以及呼吸与韵律间的关系。

1 实验说明

1.1 语料和发音人

本文共采集 28 名发音人的相关信号。其中：4 位专业播音员，6 位吟诵老先生，其他 18 位为在校大学生。语料文体包括诗词、新闻、散文等，共计 120 篇。吟诵人从语料中挑选熟悉的篇目吟诵，同时也录制普通朗读和方言朗读。自然呼吸信号在发音人休息间隙录制，采集发音人自然状态下的呼吸信号。

1.2 信号采集

本实验的录音工作在北京大学中文系语言学实验室录音室内进行。录音使用的采集设备是 AD Instruments 公司生产的 PowerLab PL3516 十六通道高速记录仪，使用自带软件 Chart5 同步采录 4 个通

[1] 本文已发表于《清华大学学报（自然科学版）》。

道的信号：第 1 通道为通过麦克风和调音台采集的语音信号，第 2 通道是通过电子声门仪（EGG）采集的嗓音信号，第 3 通道为通过呼吸带传感器采集的胸呼吸信号，第 4 通道为腹呼吸信号。采样频率均为 40kHz。

录音使用的话筒是 Sony ECM-44B，调音台是 Behringer XENYX502。电子声门仪是 KA6103。胸腹两根呼吸带是 AD Instrument MLT1132。

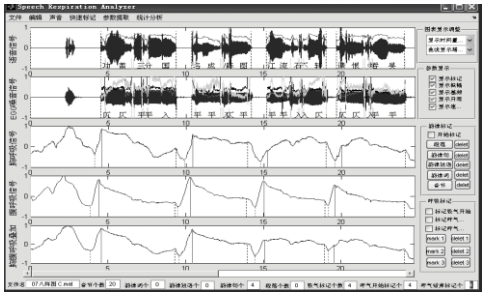


图 1 用 Matlab 编写的数据分析平台界面

2 参数提取

本文使用 Matlab 编写程序建立分析平台，如图 1 所示，对语音、嗓音、胸呼吸、腹呼吸四路信号进行同步分析，提取语音、嗓音和呼吸三类参数。首先对语音信号进行音节标记，根据音节间停顿时长的聚类分析结果划分韵律边界层级，输出各韵律单元时长、韵律边界时长、振幅等参数。然后使用程序标记出胸腹呼吸每个周期的吸气开始、呼气开始和结束时间点，由此计算出胸腹呼吸的重置幅度、时长、斜率、面积等参数。

3 实验结果与分析

吸气过程是肋间外肌提起肋骨，导致胸腔增大，膈肌收缩下移，腹肌扩张，胸腔内压力变小，气体进入肺部。呼气过程是肋间内肌收缩及肺的弹性回缩力使得胸腔变小，膈肌上移，腹肌收缩，声门下压力增大，气流被呼出。气流呼出的体积流速度和气息量，直接决定着声带振动频率的高低和振幅的大小，继而形成不同的语音基频和能量[11-12]。

3.1 平静时的呼吸特征

本文的统计显示，平静和言语时均以胸腹联合式呼吸为主，平静时呼气相时长与吸气相时长接近，而言语呼吸中呼气相远远大于吸气相。平静时的呼吸频率大于言语时，平静时男性每分钟呼吸 15-18 次，女性每分钟呼吸次数比男性快 3 次左右，言语时每分钟呼吸 8-10 次。

如图 2 所示，是一位男性平静状态下的一段胸腹呼吸信号，每个呼吸周期平均为 3.5s，平均每分钟呼吸 17 次。上图是胸呼吸信号，吸气相平均时长为 1.1s，呼气相平均时长为 1.2s，呼吸重置平均幅度为 0.49，呼吸静止段平均时长 1.2s。下图是腹呼吸信号，吸气相平均时长为 1.3s，呼气相平均时长为 1.4s，呼吸重置平均幅度为 0.52，呼吸静止段平均时长 0.8s。上图胸呼吸中虚线吸气开始时间早于下图腹呼吸平均 0.16s，上图胸呼吸中实线呼气开始时间早于下图腹呼吸平均 0.2s。

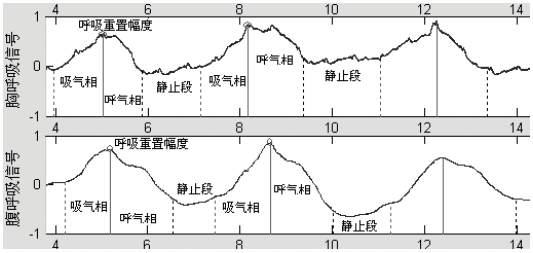


图 2　男性自然胸腹联合式呼吸

3.2 朗读的呼吸特征

言语状态时的呼吸是由中枢神经系统控制，呼吸和发音器官高度协调运动。朗读时的呼吸特征是，以胸腹联合式呼吸为主，呼气相时长远大于吸气相，呼气相时长与语音时长相等。胸呼吸主要作用是保证足够的气息量，在发音时胸腔保持扩张状态至发音结束。腹呼吸主要作用是，通过腹肌和膈肌的稳健收缩，以保证稳定的声门下压，控制气流持续释放，获得连续的语音。腹呼吸重置时间早于胸呼吸和语音起始时间。与平静状态相比，朗读时气息需求量大，呼吸重置幅度增大，即呼吸深度增加，吸气相时间缩短；呼气时间长、呼出气流平缓释放。

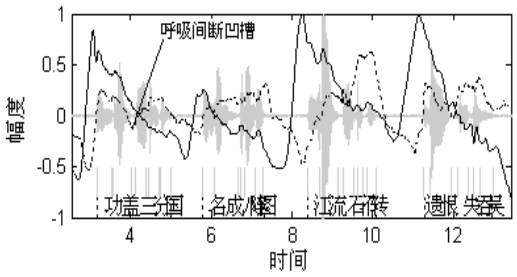


图 3　近体诗朗读中胸呼吸（虚线）、腹呼吸（实线）

如图 3 所示，是五言近体诗《八阵图》朗读时的呼吸信号，实线是腹呼吸信号，虚线是胸呼吸信号，灰色是语音信号。腹呼吸吸气开始时间（波谷）和呼气开始时间（波峰）均早于胸呼吸中。腹呼吸吸气相平均时长 0.45s，平均斜率 2.4，表现为腹肌快速扩张吸气。呼气相平均时长 2.2s，平均斜率 -0.22，表现为膈肌和腹直肌缓慢收缩，控制气流稳健释放，保证持续稳定的声门下压来产生语音。四个虚线胸呼吸重置呈拱型，呼气相与语音段基本重

合，在发音时胸腔保持扩张状态至发音结束。呼气相时长是吸气相时长的 4.8 倍，四句诗句呼气段与语音段重合。“功盖”与“三分国”之间是韵律短语边界，对应胸呼吸曲线出现呼吸间断凹槽，在其他韵律短语边界处有同样的间断出现。

3.3 吟诵的呼吸特征

吟诵是介于“读”和“唱”之间的一种拉长声音的读诗文的方法，有一定的腔调和旋律。吟诵的呼吸特点是气息量比朗读中大，一级呼吸重置增多。图 4 是同一位发音人朗读与吟诵的呼吸对比，图 4 上是朗读的胸腹呼吸信号，实线腹呼吸曲线只有一个一级呼吸重置，幅度 0.52，呼气相区域出现连续的呼吸间断凹槽，与韵律边界位置相同，呼气相面积为 1.6。虚线胸呼吸曲线也只有一个一级呼吸重置。图 4 下是吟诵的胸腹呼吸信号，实线腹呼吸曲线出现两个一级呼吸重置，幅度为 0.76 和 0.68，腹呼吸呼气相面积为 3.7，是朗读中的 2.3 倍；虚线胸呼吸也有两个一级呼吸重置，图 4 下吟诵中气息量远大于上图朗读中的气息量。

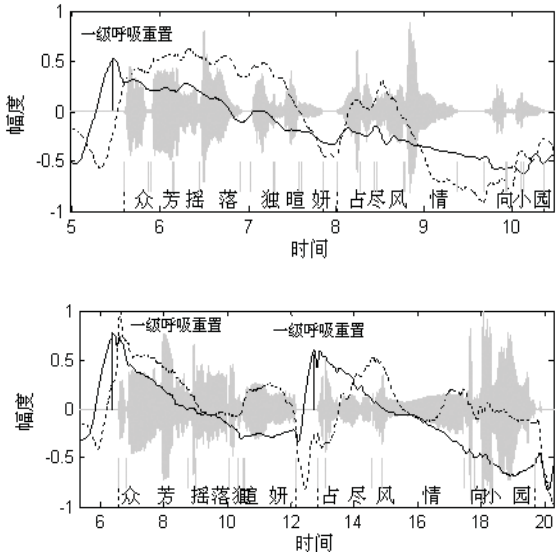


图 4　朗读（上）和吟诵（下）呼吸对比

3.4 呼吸与韵律的关系

韵律句起始处对应一个胸呼吸和腹呼吸重置，当韵律句过长时需增加一个呼吸重置。韵律短语边界对应呼吸间断，多个韵律句或一个段落对应一个呼吸群，一个呼吸群由一个一级呼吸重置和若干个二级、三级呼吸重置构成。

图 5 是一段新闻朗读的胸呼吸、腹呼吸和语音信号，文本共三小句，最后一小句共 28 个音节，在中间“党政机关”之后停顿，添加一个呼吸重置，共四个呼吸重置。四个腹呼吸和胸呼吸重置的呼气段与语音段基本重合，实线腹呼吸呼气段平缓下降，腹肌和膈肌稳健收缩。

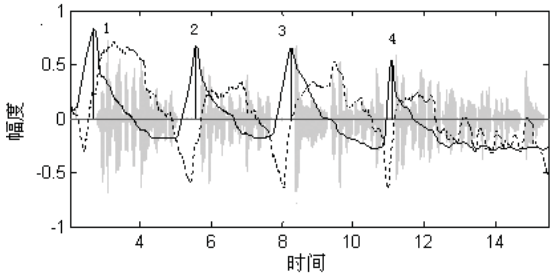


图 5 新闻朗读的胸呼吸（虚线）、腹呼吸（实线）

（文本：新华社合肥 3 月 15 日电：安徽省政府 14 日宣布，2006 年安徽省党政机关将向全国公开考录 1719 名公务员。）

图 6 是朗读《赤壁怀古》上阕，对实线腹呼吸重置幅度进行聚类分析，聚为三类，起始处一个一级腹呼吸重置，幅度为 0.97，两个二级腹呼吸重置，平均幅度为 0.61，三个三级腹呼吸重置，平均幅度为 0.39，形成一个呼吸群。

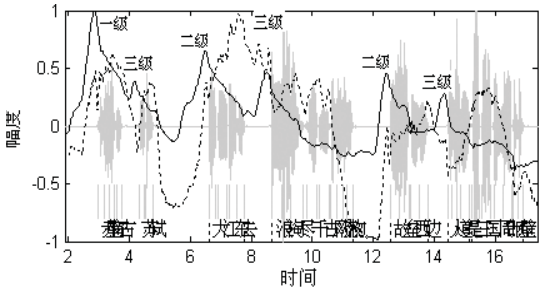


图 6　诗词朗读的胸呼吸、腹呼吸和语音信号

4 结论与展望

本文主要通过胸呼吸和腹呼吸信号研究言语产生的胸腹呼吸机制以及呼吸与韵律间的关系。实验结果表明：（1）平静时和言语时的呼吸模式不同，平静时吸气相和呼气相时长大体相等，言语时呼气相远大于吸气相，平静时的呼吸频率大于言语时。（2）言语以胸腹联合式呼吸为主，腹呼吸重置时间早于胸呼吸重置时间和语音起始时间，呼气相时长与语音时长基本相等。吟诵气息量比朗读大，一级呼吸重置增多。（3）胸呼吸主要作用是保证足够的气息量，在发音时胸腔保持扩张状态至发音结束。腹呼吸主要作用是，通过腹肌和膈肌稳健收缩，控制气流持续释放，以获得连续的语音。（4）韵律句起始处对应一个胸呼吸和腹呼吸重置，当韵律句过长时需增加一个呼吸重置。韵律短语边界对应腹呼吸或胸呼吸间断凹槽，由肋间内肌、膈肌和腹直肌的波动引起。（5）多个韵律句或一个段落对应一个呼吸群，一个呼吸群由一个一级呼吸重置和若干个二级、三级呼吸重置构成。

进一步的研究需要同步录制气流气压信号，这样对于计算气流量、声门下压力、肺的容积变化等问题将更加精确。同时结合螺旋 CT 拍摄的肺部图像，建立肺部的动态三维几何模型，对深入研究言语产生的呼吸机制具有重要作用。

参考文献

[1] Draper M H, Ladefoged P, Whitteridge D. Respiratory muscles in speech[J]. *Journal of Speech, Language and Hearing Research,* 1959, **2**(1): 16.

[2] Ladefoged P, Loeb G. Preliminary studies on respiratory activity in speech[J]. *UCLA Working Papers in Phonetics*, 2002,**101**: 50-60.

[3] MacLarnon A M, Hewitt G P. The evolution of human speech: The role of enhanced breathing control[J]. *American journal of physical anthropology,* 1999, **109**(3): 341-363.

[4] Ohala J J. Respiratory activity in speech[M].*Speech production and speech modelling. Springer Netherlands,* 1990: 23-53.

[5] Hixon T J, Goldman M D, Mead J. Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung[J]. *Journal of Speech, Language and Hearing Research,* 1973, **16**(1): 78.

[6] Stathopoulos E T, Hoit J D, Hixon T J, et al. Respiratory and laryngeal function during whispering[J]. *Journal of Speech, Language and Hearing Research,* 1991, **34**(4): 761.

[7] Slifka J L K. Respiratory constraints on speech production at prosodic boundaries[D]. *Massachusetts Institute of Technology,* 2000.

[8] 吴宗济，林茂灿. 实验语音学概要[M]. 北京：高等教育出版社，1989：33-34.

WU Zongji, LIN Maocan. A Prime of Experimental Phonetics[M]. Beijing: Higher Education Press, 1989: 33-34. (in Chinese).

[9] 谭晶晶，李永宏，孔江平. 汉语普通话不同文体朗读时的呼吸重置特征[J]. 清华大学学报(自然科学版)，2008，4：613-620.

TAN Jingjing, LI Yonghong, KONG Jiangping. Breathing-reset when reading literature in Mandarin[J]. *Journal of Tsinghua University(Science and Technology)*, 2008, **4**: 613-620.（in Chinese）

[10] 张锦玉，石锋，白学军. 讲述与朗读状态下呼吸差异的初步分析[J]. 南开语言学刊. 2012, 01:56-63.

ZHANG Jinyu, SHI Feng, BAI Xuejun. Preliminary Analysis of Respiratory Diversity between the States of Narrating and Reading[J], *Nankai Linguistics*, 2012,**01**:56-63.（in Chinese）

[11] Sakamoto T, Nonaka S, Katada A. Control of respiratory muscles during speech and vocalization[J]. *Neuronal Control of the Respiratory Muscles*, 1996: 249-258.

[12] Ron J. Baken et al.Chest wall movements prior to phonation[J]. *Journal of Speech and Hearing Research*, 1979,**22**:862-872,

收稿日期：2013-04-27

基金项目： 国家社会科学基金重大项目(10&ZD125)，教育部人文社会科学基金青年项目(13YJC740119)，中央高校基本科研业务费专项资金资助(13JNQN028)。

作者简介：杨锋（1978－），男（汉），湖北，讲师。

通信作者：孔江平，教授，E-mail：kongjp@gmail.com。

论声纹鉴定意见的表述形式*

曹洪林[1,2] 李敬阳[3,4] 王英利[5] 孔江平[1]

1 北京大学中文系；2 中国政法大学证据科学研究院；3 公安部物证鉴定中心；4 智能语音技术公安部重点实验室；5 广东省公安司法鉴定中心

摘要 本文针对目前国内外讨论比较热烈的声纹鉴定意见表述问题进行了评述。首先介绍了实践中正在使用的听觉分析法、声谱比对分析法、声学分析法、听觉-声学分析法和说话人自动识别五种鉴定方法，指出了各种方法的优缺点；然后对现存的二元判决、可能性等级、似然比和英国立场声明四种鉴定意见表述形式进行了介绍和评析，通过分析发现，上述四种意见表述形式都存在一定的问题，实践中选择何种形式表述鉴定意见要综合考虑其科学性、逻辑性、现实性和可行性等多种价值选项；最后认为解决该问题的根本方法是各相关领域的专家应加强在鉴定方法上的合作性的基础研究。

关键词 声纹鉴定；鉴定意见；可能性等级；似然比；英国立场声明；贝叶斯原理；价值选择

On Expert Opinion of Forensic Speaker Identification

Honglin CAO[1,2] Jingyang LI[3,4] Yingli WANG[5] Jiangping KONG[1]

1 Department of Chinese Language and Literature, Peking University, Beijing, China. 100871;

2 Institute of Evidence Law and Forensic Science, China University of Political Science and Law, Beijing, China.100088.

3 Institute of Forensic Science, Ministry of Public Security, Beijing, China. 100038

4 Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, Hefei, China. 230061

5 Center of Criminal Technology, Public Security Bureau of Guangdong Province, Guangzhou, China. 510050

Abstract: This paper is concerned with a critique of how to express the expert opinion (conclusion) in forensic speaker identification (comparison) cases, for which considerable discussion has been made. First, five approaches that have been put into practice currently, namely, auditory approach, spectrographic approach, acoustic approach, auditory-acoustic approach and automatic speaker recognition are reviewed, and the merits and demerits are pointed out respectively. Then four frameworks of expert opinion of forensic speaker identification, namely, binary decision, probability scale, likelihood ratio and UK Position Statement are presented and commented. The authors argue that all the four frameworks are problematic somehow when they are put into implementation, and thus we contend that scientificity, logicality, reality, feasibility and other value choices should be balanced before choosing framework for expert opinion. As a conclusion, we believe that basic and cooperative research should be strengthened by experts with related backgrounds on improving the forensic speaker identification approach of analysis, and maybe this is the fundamental method for solving this problem.

Keywords: Forensic Speaker Identification, Expert Opinion, Probability Scale, Likelihood Ratio, UK Position Statement, Bays' Theorem, Value Choice.

*该文已经发表在《证据科学》杂志，2013 年第 5 期，第 605-624 页。

1. 引言

近年来，关于法庭科学证据鉴定意见的表述问题已经成为一个非常热门的话题，很多专家和团体都对法庭科学证据传统意义上的意见表述形式提出了批评[1-3]，从中我们不难看出作者都将 DNA 的分析体系看做其他所有法庭科学证据（包括指纹、足迹、油漆、纤维、笔迹等）的典范，当然声纹鉴定[1]也包括其中。目前，声纹鉴定意见如何表述无论在国外还是国内都处在激烈的争论之中，其中争论的焦点主要是围绕基于量化的似然比（Likelihood Ratio，缩写为 LR）的意见表述形式是否适用于声纹鉴定展开的。国际上以澳大利亚的 Rose 和 Morrison 等人为代表，鲜明地支持和倡导量化的 LR 体系在声纹鉴定中的应用[5-10]，而英国的 French、Nolan、Foulkes、Harrison 和 McDougall 等绝大多数从业人员[11-15]却明确反对量化的 LR 体系在声纹鉴定中的使用，提倡采用包含贝叶斯原理现代思想的英国立场声明形式[11]，荷兰的 Broeders[16, 17]以及瑞典的 Eriksson[18, 19] 等学者也表达了对使用 LR 体系的忧虑。2011 年国际声纹鉴定协会（IAFPA）年会[2]中甚至专门设置了关于“在法庭上展示证据”的讨论专题，遗憾的是支持 LR 体系的学者并未在此会议上就此发表文章。近年来，国内也有专家讨论 LR 体系在声纹鉴定中的应用问题[4, 20-24]，然而观点并不一致。为此，本文将在介绍国内外相关研究成果的基础上，谈谈对该问题的一些看法。由于鉴定意见的表述与鉴定所依据的具体方法是密不可分的，下面先从声纹鉴定的方法谈起。

2. 声纹鉴定的方法

从历史上讲，声纹鉴定的分析方法有听觉分析、声谱比对分析、声学分析、听觉-声学分析、说话人自动识别等几种方法，很多教科书和文章中都有对这几种方法及主要分析参数的详细描述，下面仅作简单介绍。

2.1 听觉分析法

在古代，由于分析技术的局限，说话人鉴定只能通过听觉进行，当时的说话人鉴定也就是今天我们所指的耳闻证据（ear-witness evidence），从技术的角度讲，此类证据只是基于证人对声音的记忆进行的。目前，实施听觉分析的主体多是经过一定专业（语音学、语言学）训练、具备一定经验的语音学家，他/她们常常都具备使用国际音标对语音进行严式或宽式记音的能力[3]。

听觉分析法（auditory approach），也称听觉语音学分析法（auditory-phonetic approach）或听觉-感知分析法（auditory-perceptual approach），该方法可以分为整体（holistic）感知和解析（analytical）感知两个层面，其中后者应用更为普遍。整体感知就是要集中精力着重分析检材语音与样本语音是不是一个人说的，而不是努力去做成分分析，该方法类似于非专业人员进行的说话人鉴定，但相比之下，语音专家比普通人具有更大的优势。与整体感知相反，解析感知需要对语音材料进行成分分析，而这通常可以从音段层面（segmental level）和超音段层面（supra-segmental level）两个方面进行，有时还要涉及到一些超出语音学和音系学的一些非语言学特征（non-linguistic features）。

实践中，常常首先从超音段层面对发声音质（voice quality）进行评估。德国[26, 27]和英国[15]的鉴定专家通常使用 Laver [28]提出的嗓音分析框架对语音的发声类型（如气嗓音、挤喉音、粗糙嗓音和假声等）进行评估和打分，其中英国 J P French Associates 实验室的专家在鉴定中考虑的嗓音参数多达 38 维[15]4。然后还要对整个语料的韵律模式、语速和语调等特点进行检验。比如英语中不标准的重音模式和汉语普通话说话人不标准的声调模式等。音段层面的感知主要分析元音和辅音发音的具体实现情况，鉴定人员常需要根据检材（和样本）语音的音系标准去分析和比较某特殊音段/音素的各种变体，多数时候还需要使用（严式的）国际音标对其进行描述和记录。同时，鉴定人员还常检验语音的一些动态特点，如说话人在相邻音节中是否出现同化、增音或减音的现象。除此之外，说话人在特殊的

---

[1] 在中国大陆，声纹鉴定有广义和狭义之分，狭义上的声纹鉴定，即说话人鉴定或语音同一性鉴定，本文取其狭义，论述的是说话人鉴定意见的表述形式。关于声纹鉴定的称谓，大陆学者之间存在一些争议，作者相信大多数从业人员还是主张继续沿用这个叫法[4]，因此本文采用这种叫法。

[2] 会议信息和论文可以从网站上下载：http://www.kfs.oeaw.ac.at/content/blogcategory/152/533/.

[3] 国外也有普通人（非专业人员）使用听觉分析法进行说话人鉴定的情况，通常分为两种形式，一种是对熟人进行的，另一种是对陌生人进行的，后者常采用辨认的形式进行，称为“语音辨认(voice line-ups/voice parade)”。这种形式的鉴定在国内的研究较少。关于该领域的综述性介绍可以参见[25]P716-719。

[4] Nolan 对于在鉴定实践中分析音质的做法持怀疑态度，一方面因为语音学家很少接受这方面的训练[29]P331-332，另一方面电话信道对音质的感知有重要影响[14]P384 [30]P396。不可否认，针对该特征的感知检验具有较强的主观性。

言语缺陷（如口吃）、方言、外国口音、不常见的发音错误、词汇选择、语法使用、讲话模式、话语标记（如“啊”、“嗯”、“唉”、“是不是”等）使用的频率和分布、语码转换、停顿行为（无声或填声停顿、停顿的时长和位置、停顿时带不带鼻辅音，有没有出现声门化、鼻化等）等特点上的具体表现也是必须要分析的，人们在说话时常常意识不到自己会出现类似的特点，即便是意识到了，很大程度上也控制不了，因此在实际鉴定中，这些因素有时会起到决定性的作用，这种情况并不罕见。最后，还要检验一些非语言学的听觉特征，像呼吸、清嗓、咂舌、笑声的模式等。上述很多特征也是声学分析的内容，但在听觉阶段也应该进行前期的分析和评估，两种方法分析的侧重点不同（详见 2.4 听觉-声学分析法）。

2.2 声谱比对法

spectrographic/voiceprint/voice gram approach 即声谱比对法的产生得益于 20 世纪 40 年代美国贝尔实验室的一项重要发明——声谱仪（Sonagraph），该仪器可以将声音信号转换成可见的语言（visible speech），常常以声谱图（spectrogram，即三维语图——声音的时间、频率和振幅信息）的形式表现出来。最早提出将声谱图用于说话人鉴定的是贝尔实验室的工程师 L• G• Kersta [31]，他将说话人鉴定与指纹鉴定相类比，认为可以依据人们言语中独一无二的特征（unique features）——声谱痕迹（spectrographic impression）进行说话人鉴定。语音与指纹之间存在明显的区别，前者是说话人发音器官结构和发音习惯的间接反映，具有很大的灵活性，稳定性也是相对而言的[5]，后者是手指正面皮肤花纹的总称，具有终身基本不变的基本属性，对于非严重变形的指印（指纹与客体接触留下的印迹）一般采用形象比对的方式进行鉴定。然而，作为工程师的 Kersta 并没有深刻认识到这一点，尽管他声称声谱（纹）比对法的准确率高达 99%，也有坚定的支持者[32, 33]，但是该方法还是受到很多专家的批评（综述性的评论可以参见[5]P107-122、[34] P18-25 及 197-206、[35]P207-230 与[36]P115-134）。文献中关于该方法的详细介绍并不多见，但是其核心是只利用声谱图对语音进行模式匹配（pattern-matching）[35]P212、[36]P122，由于该方法仅对声谱图进行整体性的视觉比对，忽略了语音自身及其他因素引起的变化属性，其有效性和可重复性较差。2007 年 IAFPA 年会中更是通过会议公告的形式给 Tosi[33]提倡的声谱（纹）比对法定性：“该方法对于说话人鉴定来讲是整体性地，换言之非解析性地，比较言语图谱，对于图谱模式与发音动作的声学反映及声道结构之间的关系缺乏理解和解释。本协会考虑到该方法缺乏科学基础，认为它不应该用于鉴定实践。”[37]

值得一提的是，20 世纪 70 年代，美国的声谱比对分析人员注意到听觉分析在说话人鉴定中也有很重要的作用[35]P215、[36]P129，便开始从只注重声谱的视觉比对，发展为“视听”结合的方法（aural-spectrographic approach），即听觉分析与视觉比对两种方法并用（该方法不同于听觉-声学分析方法，详见下文）。不可否认对于声谱比对分析方法来说，这是很大的进步，但是其科学基础依旧没有得到完全认可，1979 年的一份很有影响力的科学报告（也称为“Bolt 报告”）曾指出：“委员会已经注意到，用视-听方法进行嗓音鉴别[6]在实验室条件下，可以达到很高的精度。在可控制的非法庭的情况下，误差率可低到 1-2%……与此同时，委员会已经注意到了科学家们对涉及法庭条件下的嗓音鉴别在精确度的估计方面很不一致。目前有关误差率的实验可用证据是来自数量相对较少的、孤立的、互相之间不合作的那些实验结果。单凭这些结果不能对实践中经常遇到的、各种条件下的误差率进行估计。”[38]P67

问题的关键在于，这里所谓的“视听”结合的方法依旧是对声谱图的模式进行整体性的视觉匹配与整体性的听觉印象之间的结合，并非解析性的方法[5]P118，因此也没有得到学术界的普遍认可，不少人依旧将“视听”方法与声谱比对法视为同种性质。IAFPA 在公告中显然回避了对此问题的看法，只是否定了对图谱进行单纯视觉比对的方法，并非否定“视听”方法[7]。关于该方

[5] 说话人在讲话过程中无时无刻不在发生变化，无论是同一时刻说相同的两个音节，还是在一天中的不同时间段，一年中的不同季节，一生中的不同年龄段，语音都不是完全相同的，有时还会差别很大，同时说话人的健康情况、情感状态等因素也会给语音带来很大影响，人们常将这种现象称之为语音自身的变化性（intra-speaker variation）。另一方面，语音还会受到语言、方言差异、录音设备和传输信道等因素的影响而发生较大变化，当然变化最大的多是来自伪装语音。

[6] 原文中的术语是“voice identification”，作者认为此该术语译为“语音鉴定”更为合适一些，但本文在引用时未做修改。

[7] 对此，Morrison 有不同意见，他的理解是 IAFPA 针对声谱（纹）

法较详细的介绍，可以参见[40-43]。

实际上，美国联邦调查局（FBI）在 1979 年“Bolt 报告”发表后的近三十多年来一直使用“视听”结合的声谱比对分析法，但是只用于侦查目的，并不将其结果作为证据在法庭上使用[39, 44, 45]8，同时美国的部分私人实验室也仍在使用该方法[15, 43]，而且声谱比对证据在美国部分州中仍然是可采的9[50] P231、[48]P423-424、[45]P151，但是由于司法界对该技术方法科学基础的怀疑，声谱比对技术的从业人员从原来的五六十人萎缩到只有大约十几人[48]P425-426。可以预见的是“视听”结合的声谱比对分析法在短期内并不会退出历史舞台（比如在美国的情况参见[51]）。

由于在该方法在名称上与国内“声纹鉴定技术”的名称有一定的关联性，因此有必要澄清的是，目前我国声纹鉴定技术的方法并非单纯、机械的声谱图比对，而是语音学分析法（听觉-声学分析法）。同时不可否认，针对学科名称，学界存在不同的观点，但是按照国内法庭科学中多个学科的命名习惯和特点，我国业内绝大多数人主张不妨延用“声纹鉴定”一词[4, 52]。

2.3 声学分析法

声学分析法（acoustic approach），也称声学语音学分析法（acoustic-phonetic approach），是指借助计算机技术对特定语音单元的声学特性进行定量测量的方法，分析对象常常是由听觉分析挑选出的音素、音节、韵律短语或句子等特定的语音单元，具体可以测量语音单元的频率、时长和/或振幅等信息。实践中常用的声学分析参数有很多，如基频的均值、标准差、中位数、众数、范围、长时平均基频分布情况；共振峰的频率、轨迹的动态特性及长时共振峰的分布情况，共振峰参数可以提供说话人的很多发音和音系模式的细节信息，对于说话人鉴定来讲非常重要。除了上述频率维度的信息之外，一般还要测量嗓音起始时间（VOT）、语音单元（如元音、辅音等）的时长（或时长比例）、整个语料的发音速率等。在分析过程中不仅会涉及到数学计算和数据统计方面的知识，而且常常还会用到很多现代语音信号处理的技术，如比较常用的线性预测技术（LPC）和快速傅里叶变换方法（FFT），由于整个过程都需要大量的人工干预，一般来讲需要花费较长的时间。值得注意的是，声学分析与声谱比对分析不同[10]，前者是解析分析方法，后者则是“格式塔”式的整体分析，同时，声学分析并不排斥利用声谱图特征来分析语音，很多情况下，声谱图对于提取部分声学参数（如共振峰）来说是不可或缺的。

2.4 听觉-声学分析法

顾名思义，auditory-acoustic approach 即听觉-声学分析法[11]是听觉分析与声学分析两种方法的结合，之所以在此将其单独论述，不仅是因为两种方法在实践中密不可分，而且目前这种“结合”使用的方法在世界范围内得到了广泛认可。国内专家提到的“语音学分析法”[54, 55]，其基本内容包括了听觉鉴定、视谱比较和定量比对三个方面，其实质与国外学者提到的听觉-声学分析法是一致的。

尽管有时为了论述方便，将听觉分析与声学分析分开来讲，但实际上两者之间是共生的、互补的关系[13, 14, 29, 56]，两种方法能够捕捉到的语音特征信息各有侧重点，不能互相替代，而且在实际分析过程中往往是交替进行、循环使用的。听觉分析的主观性较强，很多内容还常常涉及到心理感知的问题，声学分析一方面可以为部分听觉分析的结果提供量化的支持，另一方面还可以提供新的特征。实践证明单独依靠任何一种方法都是不可取的，两种方法同等重要，均不可或缺[5]P33，这一点已经成为国内外声纹鉴定从业者的一个基本共识。声学分析不能单独进行，理由很简单，鉴定中有必要先听录音，判断语音质量，理解语音内容，在此基础上选出相同的分析对象

比对分析法的公告，同时意味着彻底否定了“视听”结合的方法，但是从 Morrison 与 IAFPA 主席 French 的个人交流中可以看出，公告的起草人和支持者并非持相同的想法。关于 Morrison 对该问题的看法详见[39]。

8 最近有报道[46,47]显示，目前 FBI 已经不再将“视听”结合的声谱比对方法作为主要的分析手段，转而使用说话人自动识别技术，而且依旧只服务于侦查或谍报工作，而不是用于审判目的；然而，FBI 并非单独使用说话人自动识别技术（第一步），而是与（有资质的受过训练的）言语鉴定专家进行的听觉-感知分析方法（第二步）相结合，而且只有当两种分析手段都得出“匹配（match）”结果时，才能得出最后的匹配结论，如果两种方法得出的意见不一致，则只能得出“无结果（inconclusive）”的结论。

9 关于单纯的图谱视觉比对形式和“视听”结合形式的声谱比对分析方法在美国法院中（依据 Frye 规则、联邦证据规则和 Daubert 规则等）被采信情况的综述性介绍可以参见[48, 49]。该方法的支持者 Maher 提到，目前没有普遍接受的科学研究能够完全量化该方法的错误率，因此法院在做出是否采信由此方法得出的鉴定意见时，必须要具体案件具体分析[43] P92。

10 声谱比对分析法主要是在一些培训班里集中讲授的（课时较短），对象主要是一些没有语音学和语言学背景的人，很明显，单纯的图谱比对并不是由那些受过适当训练的人做出的声学语音学的系统分析[25]。

11 还有学者将此方法称为听觉-仪器分析法（auditory-instrumental method）[53]，但是该叫法并不是很常用。

（例如相同音系环境中相同音节中的相同元音）进行进一步的比对，很明显这是声学分析的前提，如果没有这种听觉分析的选择和控制，鉴定意义上的声学分析便无从谈起[5]P33-34。王英利更是认为“鉴定意义上的相同音节是指调音音质完全相同的音节。具体到鉴定实践中就是人耳无法辨识调音音质有何不同的音节。因为，到目前为止，还没有发现比人耳性能更好的辨音‘仪器’。”[52]这些鉴定意义上的相同音节应该成为声学分析的重点对象。另一方面，在综合评断环节，由于对检材语音和样本语音进行声学分析的结果往往不会完全相同，在对差异点进行分析的时候，听觉分析的结果无疑会提供直接的佐证。

如今在声纹鉴定实践中该方法无疑是占主导地位的，欧洲和其他地区的大多数语音鉴定专家都使用该方法，IAFPA 的绝大多数成员也都在该框架下工作[18]12，国内的情况亦是如此[57]。值得注意的是，该方法既非听觉分析与单纯视觉比对形式的声谱分析法的简单相加，亦非“视听”结合形式的声谱比对分析法，原因在于该方法更加注重从听觉和声学上对语音进行解析式的分析，并未局限在整体性的“语图”特征上。值得一提的是，Jessen 将国外语音比对（forensic voice comparison）[13]中使用的方法分为“解析法”和“整体分析法”两个层面，并认为应该综合使用这两个层面的方法，而非仅仅使用“解析”层面的方法[25]。的确，我们不应该因为声谱比对分析给人留下的“格式塔式”的整体分析的错误印象，就一味地排斥所有的整体性分析方法，毕竟综合运用各种方法是明智的选择。

2.5 说话人自动识别

尽管广义的说话人自动识别技术（automatic speaker recognition）可以追溯到 20 世纪 60 年代，但是该技术在声纹鉴定领域的应用却只有 10 年左右的时间。该方法不同于传统意义上由人工实施的声学分析法，多是自动完成的，其基本原理是，由语音信号处理工程师设计一定的程序和算法，对语音中说话人的特征参数分离提取，然后针对特征建立语音模型，进行距离计算，最后确定与其最为接近的一个已知说话人的语音模型。目前，最常用的声学特征参数是美尔倒谱系数（MFCC）和线性预测倒谱系数（LPCC），最常用的语音模型是高斯混合模型（GMM）。说话人自动识别方法的优点是自动化程度高，需要较少的人工干预，一旦建立比较省时省力；缺点是具体分析参数在语音学上很难得到解释，而且更为重要的是，尽管近年来该方法已经取得了很大进步，不少国家的侦查部门已经（或准备）用来帮助侦查活动，但是该技术针对实际案件的正确识别率还不是很高，远没有达到可以为法庭提供可靠证据的程度，至少目前是这样。或许，说话人自动识别的最终发展趋势还是需要与上述几种专家方法相结合，形成半自动综合识别的方法。

2.6 各种方法在鉴定中的应用

最近，Gold 与 French 通过对 5 大洲 13 个国家的 36 位声纹鉴定专家进行了一项有关说话人鉴定的国际调查[14]，调查发现，仅就鉴定方法而言，有 2 人单独使用听觉分析法（约占 6%），1 人单独使用声学分析法（约占 3%），25 人使用听觉-声学分析法（约占 71%），7 人使用说话人自动识别并加人工分析的方法（占 20%），没有人单独使用说话人自动识别的方法（占 0%）[15][58]。很明显可以看出，目前听觉-声学分析法在世界范围是使用最广泛的方法。尽管 Gold 与 French 在调查中并没有将声谱比对分析法作为调查的选项，但是并不意味着该方法在实践中已经消失了，只不过该方法早已不是鉴定方法的主流，哪怕是在其诞生地美国。据笔者所知，我国大陆地区开展声纹鉴定业务的鉴定机构主要使用听觉-

[12] 另据 Foulkes 和 French 介绍：英国、澳大利亚、奥地利、芬兰、德国、荷兰和瑞典的声纹鉴定专家都主要使用该方法（作者将其称之为“语言学-声学方法”(linguistic-acoustic approach)）；在美国，该方法同样受到最著名的语言学家 Hollien、Labov 和 Ladefoged 等人的支持和拥护；另外，至少还有 20 个国家和地区都使用这一方法，其中包括加纳、香港、印度、巴基斯坦、俄罗斯、南非和前南斯拉夫等；包括国际调查委员会（International Commission of Enquiry）和联合国国际战犯法庭（the UN international War Crimes Tribunal）在内的各个层次的法院也都接受该方法[15]。

[13] 近年来该术语(或称说话人比对 forensic speaker comparison)在国外比较流行，用以替代之前的语音/说话人鉴定/识别（forensic voice/speaker identification/recognition），之所以改用“比对（comparison）”一词的原因在于从业人员意识到他/她们从事的工作只是对检材语音和样本语音进行客观的比对，给出的结果只是一种比对的意见，并非“鉴定（identification）”得出的结论，同时避免给事实裁判者（法官或陪审团）提供一种“最后定论”的印象。然而，目前“说话人鉴定”的概念在国内还是非常普及的。

[14] 这 13 个国家包括澳大利亚、奥地利、巴西、中国、德国、意大利、荷兰、韩国、西班牙、瑞典、土耳其、英国和美国，36 位专家有 18 人在大学和科研机构，13 人在政府实验室/机构、9 人在私人实验室、7 人是个体从业者，其中有的专家在多个机构工作，这些专家都是欧洲法庭科学研究网工作组（ENFSI）、美国国家标准与技术研究院（NIST）或国际声纹鉴定协会（IAFPA）三个组织的成员。

[15] 估计此处一共有 35 人反馈了分析方法一项的调查，因为原文作者提到允许被调查人做选择性回答。

声学分析法，拥有自动识别系统的鉴定机构多将其识别结果作为参考使用，极少数鉴定机构也有使用说话人自动识别技术作为主要鉴定方法的[16]，目前缺乏类似的准确统计。

3. 现有声纹鉴定意见的表述方式

Gold 与 French[58]在文中也对不同国家/专家的声纹鉴定意见的具体表述方式进行了调查，结果发现不同国家/专家的表述情况存在很大差别，相关科学团体就此问题的观点也很不一致。调查发现13个国家的36位专家使用的意见表述形式主要有以下 5 种：二元判决形式（binary decision）、经典的可能性等级形式(classical probability scale)、数字似然比形式（numerical LR)、文字表述似然比形式(verbal LR）、英国立场声明形式（UK Position Statement)。上述5种意见表述形式分别有2(巴西、中国[17])、9（澳大利亚、奥地利、巴西、德国、荷兰、韩国、瑞典、英国、美国)、4（澳大利亚、德国、意大利、西班牙)、2(荷兰、美国）和5（德国、西班牙、土耳其、英国、美国）个国家[18]的从业人员所使用。尽管意见表述方式与具体的鉴定方法密切相关，但也并非一一对应，比如在使用听觉-声学分析方法的 25 位专家中，就有 10 人使用经典的可能性等级表述、10 人使用英国立场声明形式、2 人使用二元判决形式、2 人使用文字表述 LR 形式、1 人使用数字 LR 形式。同时使用经典可能性等级表述的 14 位专家中，有 1 人使用听觉分析法、10 人使用听觉-声学分析法、3 人使用自动识别加人工分析的方法。下面对各种意见表述形式进行介绍。

3.1 二元判决形式

二元判决，即在表述意见时只有两种选择，要么绝对认定，要么绝对否定，没有中间选项。

3.2 经典的可能性等级表述形式

鉴定专家依据听觉-声学分析方法（如上文所述，也有少部分专家单独使用听觉分析法和自动识别加人工分析的方法）进行检验，通过比较检材和样本的语音特征，分别从听觉上和声学上对这些语音特征之间的相似性/差异性进行判断，最后对所分析到的语音特征组合进行综合评估，在充分考虑特征质量和数量限制的基础上，得出鉴定专家对检材语音与样本语音是否是同一人所说的确信程度。通常这种确信程度由不同等级的可能性来表述，尽管不同国家/专家划分等级的数量可能不同（从 5 级到 11 级不等），具体叫法也有差异，但是基本原则都是一样的[14] P388。如瑞典警察就推荐使用 9 个等级的分类体系[18, 60]。

+4 结果支持假设接近确定（support the hypothesis with near certainty）

+3 结果很强地支持假设（strongly support the hypothesis）

+2 结果支持假设（support the hypothesis）

+1 结果在某种程度上支持假设（support the hypothesis to some degree）

0 无结果（inconclusive）

－1 结果在某种程度上不支持假设（contradict the hypothesis to some degree）

－2 结果不支持假设（contradict the hypothesis）

－3 结果很强地不支持假设（strongly contradict the hypothesis）

-4 结果不支持假设接近确定（contradict the hypothesis with near certainty）

这里的“假设”是指检材语音与样本语音是同一人所说。

德国的政府机构和大部分或所有的私人专家[53, 61]、芬兰的鉴定专家[18]也采用相同的 9 级分类体系；法国则采用 7 级分类体系[62]P215；美国以“视听”结合方法进行鉴定的专家有的将意见分为 7 级[41, 42]，有的则采用 5 级分类体系[38, 43, 63]；中国大陆则多采用 5 级的分类体系[54, 55 64-66]19；其他大多数国家也应该是这样的。

3.3 似然比（LR）形式

在声纹鉴定领域中，LR 反映了对认定说话人的两个竞争性假设—起诉假设(检材语音和样

[16] 通过2010年司法部司法鉴定科学技术研究所组织的能力验证(语音同一性鉴定项目）的反馈情况看，有鉴定机构通过一些自动识别软件来进行鉴定，尽管结果正确，但由于其鉴定文书存在多方面的问题，其反馈结果为“不满意”，详见[59]。

[17] 我们无法得知被调查人是谁，但是目前在中国大陆开展声纹鉴定业务并经过实验室认可的鉴定机构，都没有采用这种形式，就笔者所知，现在极少有人采用这种形式表述鉴定意见，这种表述方式至少在形式上不是中国大陆声纹鉴定从业人员的主流选择。

[18] 与我国实行统一管理的鉴定体制不同，在许多英美法系国家中，从业人员（专家证人）在选择何种形式的意见表述方式上有很大的自由度，因此，同一国家的不同专家可能会选择不同的意见表述方式。

[19] 稍有不同的是，中国大陆的 5 级分类中包含了“明确的”认定和否定的形式（比较脚注 20)；美国的部分专家也采用类似明确的形式，但国外的分类表述多是模糊性的，不含“绝对的”肯定或否定表述，同时也多是对称性的，但也有非对称的情况，如在[67]的一项调查反馈中发现，有专家在排除说话人时也采用“明确的否定”形式，但在认定说话人方面却不是。

本语音同源，即是同一人所说）和辩护假设（检材语音和样本语音不同源，即是不同人所说）——之间的关系，在数值上等于同源和不同源两种可能性的比值。另外也经常用语音特征之间的相似性（similarity）和普遍性（typicality）的比值来表示。相似性是指检材语音和样本语音在我们比较的语音维度上相似或不同的程度有多大。二者越相似，它们源自同一说话人的可能性就越大。同时，所选语音特征在所有人（或一定群体，如按方言或性别进行分类）的说话特点中可能比较常见，也可能比较罕见，这种普遍性越低，语音证据的证明力就越强。

LR 可以用公式（1）表示：

$$LR = \frac{p(\mathrm{E}|\mathrm{H_p})}{p(\mathrm{E}|\mathrm{H_d})} \quad (1)$$

该公式中，p 表示概率（probability），E 表示证据（evidence），H_p 表示起诉假设（prosecution hypothesis），H_d表示辩护假设（defense hypothesis），当起诉假设正确时，分子表示获取已知语音证据 E 的概率；当辩护假设正确时，分母表示获取相同证据的概率。如果 LR 大于 1，说明有相对更多的证据支持检材语音与样本语音是同一人所说，反之，如果 LR 小于 1，说明有相对更多的证据支持二者是不同人所说。分子代表比对的相似性层面，分母则代表普遍性层面，如果检材语音与样本语音的相似程度高，那表明二者是同一人所说的可能性也相对较大，反之亦然；当普遍性较高时，检材语音来自其他说话人的可能性也就会相对较大，反之，当普遍性较低时，除了样本语音的说话人以外，其他说话人涉案的可能性也相对较小。如果 LR 等于 1，说明检材语音与样本语音源自同一说话人和不同说话人的可能性是一致，此时的语音证据就起不到作用了。

LR 方法的一个优点是可以对特征的价值进行量化，在综合各个特征的 LR 数值的基础上，给出语音证据力度的量化值（各个 LR 数值相乘），用完全数字的形式表现出来（数字 LR 形式）。鉴于 LR 的数字形式可能不能被法庭所充分理解，有学者提出了 LR 相应的文字化分级形式（文字表述 LR 形式）[68]P240、[5]P61，或者是将 LR 数值取对数，换算成对应的较小数量级的数值，见表 1。至于在法庭上采用何种形式的 LR 体系进行表述，目前没有一致性的意见[18]P62。

表 1　LR 数值（及对数形式）相对应的文字表述形式的证据强度

LR	Log LR	文字表述形式
>10000	>4	非常强的（very strong）证据支持
1000-10000	3-4	强的（strong）证据支持
100-1000	2-3	适度强的（moderately strong）证据支持
10-100	1-2	适度的（moderate）证据支持
1-10	0-1	有限的（limited）证据支持
1-0.1	0 - −1	有限的（limited）证据反对
0.1-0.01	−1 - −2	适度的（moderate）证据反对
0.01-0.001	−2 - −3	适度强的（moderately strong）证据反对
0.001-0.0001	−3 - −4	强的（strong）证据反对
<0.0001	<−4	非常强的（very strong）证据反对

由于LR体系是基于贝叶斯理论(Bays’ Theorem)的方法，在此有必要对贝叶斯理论做简要介绍。贝叶斯理论可以用公式（2）进行表示：

$$\frac{p(\mathrm{H_p}|\mathrm{E})}{p(\mathrm{H_d}|\mathrm{E})} = \frac{p(\mathrm{E}|\mathrm{H_p})}{p(\mathrm{E}|\mathrm{H_d})} \times \frac{p(\mathrm{H_p})}{p(\mathrm{H_d})} \quad (2)$$

后验概率　似然比　先验概率

（posterior odds）（LR）（prior odds）

可以看出，后验概率是先验概率与 LR 的乘积。后验概率是整合案件中所有证据之后得到的概率（事实审判者最后的内心确信程度）。先验概率则表示事实审判者在鉴定专家展示以 LR 形式表现的语音证据（其他证据也一样）之前对上述两个竞争性假设之间的内心确信程度。要得到后验概率，必须知道先验概率，而先验概率的获取可能会受到案件中其他证据或案情的影响，如 DNA、指纹、证人证言等。显然，在贝叶斯理论框架下，声纹鉴定专家没有能力也不应该获取先验概率，继而也无法对后验概率做出评价。关于

LR 体系在语音证据运用上的详细介绍，可以参考[5, 8, 10, 20, 25]。

3.4 英国立场声明形式

英国立场声明形式是指由英国专家 Peter French 在 2005 年的 IAFPA 年会上提出，后经约克大学和剑桥大学的声纹鉴定专家讨论，最后于 2007 年形成的一份关于语音证据如何在英国法庭上表述的立场声明，除一人之外，英国所有的声纹鉴定专家都签名表示支持，该声明可以在期刊[11]和网络[69]上查询到。按照该形式，专家需要作出两方面的判断：一致性（consistency）和独特性（distinctiveness）。首先判断检材语音与样本语音的一致性，一致性是指检材语音和样本语音在来自同一说话人方面是否是一致的（consistent）或符合的（compatible），一致性判断有以下三种可能的形式：

一致 （consistent）

不一致（not consistent）

无结论（no decision）

1、如果在一致性方面得出的是“**一致**”的结果，专家还要进一步对语音特征的独特性进行整体评估，并将其分为以下 5 个等级：

格外特殊（exceptionally distinctive）——其他说话人同时享有这些特征组合的可能性是极其微小的。

高度特殊（highly distinctive）

特殊（distinctive）

适度特殊（moderately distinctive）

不特殊（not distinctive）

这其中有一个例外情况，即对少数有独立证据（比如视频监控）显示已知说话人出现并参与了谈话的闭集比对（closed set comparison）的案件，如果几个声音之间的差别足够明显，可以做出**绝对认定**的结论。

2 如果在一致性方面得出的是“**不一致**”的结果，专家既可以得出明确的排除结论也可以采用可能性等级的形式进行表述。

3.5 中国大陆现有的形式

目前中国大陆声纹鉴定意见为可能性等级表述形式，可能性等级为 5 级，该观点在众多文章[4, 54]和著作[55, 64-66]中都有所体现。近年来公布（推荐）实施的两个鉴定方法是关于此问题最为详细的介绍，一个是公安部物证鉴定中心制定的“语音同一认定方法（IFSC 11-01-01-2010）”（下称公安部方法）[70]，另一个是司法部司法鉴定管理局公布的《录音资料鉴定规范（SF/ZJD0301001-2010）》（第 3 部分《语音同一性鉴定规范》）（下称司法部规范）[71]。据笔者所知，我国大陆地区的公安、检察和安全机关下属的开展声纹鉴定业务的鉴定机构基本上都采用了公安部方法，而面向社会的开展声纹鉴定业务的鉴定机构大多采用了司法部规范。公安部方法（左）和司法部规范（右）均将声纹鉴定的意见划分成 5 个等级并给出了具体的划分标准：

（+2）	认定同一	肯定同一
（+1）	倾向认定同一	倾向同一
（0）	无结论	无法判断是否同一
（-1）	倾向否定同一	倾向不同一
（-2）	否定同一	否定同一

虽然具体叫法上与国外部分国家采用的可能性等级表述有所不同，其内涵是基本一致的[20]。值得注意的是，上述两种方法在具体等级的划分标准上有所不同，公安部方法对等级划分有明确的特征“数量”要求，如对“认定同一”的要求是：

“此结论要求检材、样本中可供比对的音节有 10 个以上，每个音节有 3 条以上有效共振峰；所有可供比对音节的特征符合率超过 90%。或者可供比对的音节有 6 个以上，每个音节有 4 条以上有效共振峰，特征符合率超过 95%。如果检材有严重伪装，则不能下此结论。”

相比之下，司法部规范给鉴定人员的自由度较大，没有硬性的“数量”要求，以“认定同一”为例：

检材语音与样本语音：存在**足够的**符合特征，且符合特征的价值充分反映了同一人的发音特点；没有本质的差异特征；同时差异或变化特征能得到合理的解释。

4. 对各种意见表述形式的评价

4.1 二元判决形式

理论上说，就说话人进行鉴定，其结论无外乎两种——认定或否定，但在实际办案中，鉴定人员会受到各种条件的局限，这种局限性一方面体现在说话人语音自身的变化性上，另一方面诸如噪音、传输信道（如电话）等外在因素也会引起语音质量的下降，有时还会使听觉和/或声

[20] 尽管两个方法中并没有将最高等级的意见表述为类似于“几乎肯定”或“接近确定”等叫法的形式，但其实质都表示是达到了鉴定人最高等级的内心确定，无论鉴定意见表述如何绝对，终究还是鉴定人自身的一种专业判断，与事实审判者的定案结论不同。

学分析受限或无法进行，同时，检材语音的时长如果很短的话，常常也不能充分体现说话人的语音特点。因此，基于上述众多局限性，针对所有情况只对语音做出绝对认定或否定的结论是不客观的，实践中鉴定人常常做出多个等级的可能性判断。Gold 与 French[58]在调查中提到巴西和中国的学者至少还有人在使用这种二元判决形式，但是这种形式并非中国大陆声纹鉴定从业人员的主流选择（详见脚注 17）。

4.2 经典的可能性等级表述形式

在排除了二元判决形式之后，我们似乎必然转向可能性等级表述形式，这种依据文字描述来区分各种可能性等级的形式，其优点很明显：无论是对专家而言，还是对法官、陪审团和普通民众来说都是很容易理解的，同时也被世界上大多数国家（无论何种司法体系）的政府机构和专家所广泛使用。采用这种意见表述形式的绝大多数专家都使用听觉-声学分析法进行鉴定，少数专家也有使用其他方法的，但都存在人工分析的成分。

尽管该形式被广泛使用，但并不代表它没有任何瑕疵，其不足之处突出表现在两个方面：主观性较强和逻辑上存在缺陷。

第一，该形式有较强的主观性。首先，在听觉分析中对各种语音特征的评价很多都是基于心理的主观感知，即使是在受过专业训练的语音专家之间也会存在一定的不确定性。对听觉特征之间的相似或特异程度做出数字形式的量化，继而做出统计分析，是一件非常困难的事情，或者说基本是不可能的，因为不同鉴定人（普通人也一样）对同一特征的感知未必完全相同[21]。其次，在对多个参数、多种方法的分析结果进行权重和综合分析时，主观性也是不可避免的。通过上文的分析，我们知道听觉分析和声学分析相辅相成，缺一不可，但是并不是说这两种分析方法一定会得出意见一致的结果，如果两者出现矛盾应该如何取舍？同样的困惑还可能会出现在解析分析与整体分析、人工分析与自动识别分析产生不一致结果的情况中。对此，我们常常依据现有的语音学/语言学理论作为评判标准，对特征的符合和差异程度做出是本质的/非本质的判断，其实这种是否“本质”的判断依然不是想象中的客观。

再次，可能性等级划分的主观性体现在三个方面：（一）等级划分的具体数量不统一，不同国家/专家之间使用的数量形式不尽相同，选择几级的可能性表述是由专家自己决定的[22]；（二）不同等级之间的划分没有具体的界限，很多情况下，鉴定意见的等级划分主要取决于专家对所选语音特征数量和质量的评估，由专家最后形成的不同的内心确认来表述意见的不同等级。尽管有的可能性程度划分有“客观”依据可循，如美国 IAI[41]和录音证据委员会[42]公布的表述形式中有细化的划分标准，但是未见有得出此标准的科学论证[23]。

意见主观常常成为 LR 体系拥护者批评经典可能性等级表述形式的主要论点之一，其实，我们有必要将鉴定意见表述形式的主观性和鉴定方法的主观性区分开来，上文提到的除了说话人自动识别之外的其他方法无一不是具有主观性的，正如 Broeders[16] P238 所言：“只要明确该方法是主观的，这种主观的判断就不应该受到责备，原因在于它就是主观的。关键的问题不是专家得出的结论是主观的还是客观的，而是该方法是否可信。”方法的主观性并非必然意味着分析是不准确的，对于实践中大多数案件而言，不同经验丰富的鉴定专家分析同一语料得到的鉴定意见常常是相同的。以 Eriksson 为例，他与其他专家进行合作分析的时候，就很少出现结论不一致的情况[18]。

第二、逻辑上存在缺陷是可能性等级表述形式备受批评的另外一个缺点。可能性等级表述形式的一个问题是让鉴定专家重点分析语音材料之间的相似性（和差异性），然而对特征之间的相似性在整个参考背景人口中的分布情况（即上文提到的普遍性）却关注太少[14] P388。而且在一些鉴定专家中普遍存在一种误解，即认为只要确定两种语料之间具有较强的相似性，就可以做出认定意见[25]P719，面对质疑一些鉴定人可能会说“我只是被要求对检材语音和样本语音进行比较，并没有被要求要把检材语音与所有人的声音

[21] 操作性较强的做法可以参照 Hollien 提出的“结构化方法”，从 0-10 级对语音特征之间的感知相似性进行判断，0 表示不相同，10 表示相同[36]P78-85。

[22] 很多国家对此并没有强制性规定。20 世纪 90 年代中期，Nolan 曾经主持 IAFP（IAFPA 的前身）的一个委员会研究探索设计一种对于协会成员来说意见一致的表述方式的可能性，但是最后他放弃了，因为 Nolan 认为“与意见一致的范围表述相比，这种占主导地位的真实的混乱状态能够更好地反映当前的技术水平。[13]”

[23] 如果能够证明该划分标准有充分的科学基础，无疑能够提高意见表述的客观性。

都进行比较，因此检材语音与整个人口中其他人的声音是否相似是无关的。”[13]实际上，如果只关注特征之间的符合程度（相似性）而忽略该特征符合在整个参考人口中的分布的话，我们便会不可避免地出现逻辑上的错误，即如果该符合特征在人口分布中非常罕见，由此得出认定意见的准确性会比较高，如果该符合特征在（除了被鉴定人以外的）其他人的声音中也非常普遍的话，再得出认定意见的出错率就会很高。正如 LR 形式反映的认定说话人的两个竞争性假设，忽略任何一个假设，都会在逻辑上出错。

4.2 似然比（LR）形式

相对于经典的可能性等级表述方法而言，LR 体系在逻辑上有没有缺陷，该方法要求在说话人鉴定中同时考虑两个竞争性假设的做法使其在逻辑上是正确的。该形式不对检材语音和样本语音是否是同一人所说做出绝对的后验性的判断，而只是对语音证据的力度进行描述，且使用量化的“客观”数字说话，既能有效降低鉴定的主观性，又能避免鉴定专家“代替”事实审判者做出最后的结论[24]。假如 LR 体系支持者们果真能够达到这些目标，那将是声纹鉴定领域的巨大成就，然而，该体系是否真如其支持者所言是最科学、最可靠的形式呢？事实可能并非如此，原因主要有以下三点：

第一、LR 体系支持者使用的鉴定方法具有主观性，并无创新。必须强调的是，任何意见表述形式都不等同于具体的鉴定方法，LR 形式也不例外，评价该形式是否准确可靠，要先从其所依据的方法入手来看。Morrison[8]P301-303 与 Morrison[10]曾在介绍各种鉴定方法的基础上，对其与 LR 体系的适用性进行过较详细的评论，作者认为声学分析方法和自动识别的方法可能适合 LR 体系，相反，由于听觉分析法（和听觉-声学分析法）与声谱比对分析法都是基于经验的主观性判断，从而认为这些方法与 LR 体系并不相容。LR 体系的支持者之所以青睐声学分析方法，原因在于声学分析能够提供“客观的”量化数据，是最容易进行统计计算的部分，然而声学分析法并非完全客观。如上文所述，在鉴定实践中声学分析与听觉分析是互补的、共生的（再强调一下，绝大多数专家都坚持应该将听觉分析与声学分析结合使用，而不是单独使用任何一种方法！）。如果通过听觉分析发现检材语音和样本语音中体现出根本性的方言差异，基本可以排除两者是由同一人所说，就没有进一步分析的必要了，如果检材和样本中说话人的情绪状态差异较大（如愤怒的大声说话与正常讲话），语音特征之间往往变得不具可比性，也就是说，声学分析最终选择的分析对象往往是在听觉分析中被判断为是相近或相同的特征，这种基于听觉分析结果的声学分析不会是完全客观的，原因很简单，因为被 LR 体系所排斥的听觉分析是主观的，很难想象 LR 支持者认识不到听觉分析与声学分析之间的密切关系[25]，之所以排斥听觉分析的原因多半是由于很多听觉特征是定性的，无法满足 LR 体系要求特征必须被量化的要求。同时，即使是声学分析中，鉴定人使用的软件不同，同种软件的算法、设置不同等，都会带来一定的误差或错误，换句话说，鉴定人员在声学测量时所做的类似选择都带有一定的主观性，继而也会影响到结果的“客观性”。在声纹鉴定领域，LR 形式最早同时也是主要应用在说话人自动识别上的，问题的关键是，尽管说话人识别技术取得了很大的进步，在少数国家（如西班牙和法国）也得到了认可，但是到目前为止，依旧无法广泛且准确地应用到鉴定实践中去，其识别结果还远达不到为法庭提供可靠证据的程度。由此可见，在实际鉴定中，LR 形式支持者使用的鉴定方法并无创新，与可能性等级表述形式的拥护者相比，前者对听觉分析的排斥，必然导致鉴定结果不准确的可能性增加。如前文所述，目前在众多鉴定方法中，听觉-声学分析方法仍然占据主导地位，那些认为构建可以被视为具有可行性的、严密的、排他的定量 LR 鉴定方法之前，仅仅是时间和研究问题的观点是不现实的，更不用说该方法的可靠性了[12]。

第二，众多特征，尤其是听觉特征很难被量化。稍微看一下 LR 支持者的研究成果就会发现，其研究主要集中在声学分析中的共振峰频率和轨迹等方面，对其他特征少有论及。本文在 2.1 部分比较详细地列举了听觉分析方法需要（或能够）分析的特征，French 等[12]P146-147 更是明确列

[24] 在英国立场声明中，众多签名人表示在原则上接受 LR 体系的现代思想，但是对量化的 LR 形式能否在现实中实现持否定态度，详见下文分析。

[25] Morrison 似乎也承认与完全自动的程序相比，声学分析中的人工指导会提高测量的准确性[10]。

举了声纹鉴定中常用的 11 种特征，很明显，这些特征的涵盖范围很广，有语音学的、语言学的也包括非语言学的特征，远远超过了仅对共振峰的分析结果，问题的关键是其中的很多特征是很难对其实施量化统计的。在实际鉴定中，鉴定专家必须要考虑到送检材料的各个方面，进行全面而系统的分析。尽管已经有研究发现部分元音的共振峰模式具备较强的话者区分能力，但这并不意味着鉴定人员就可以忽略其他方面的特征，因为基于其他很多特征的分析结果可能是决定性的，而且不一定与对共振峰的分析结果一致[26][12]。

第三、相关参考人口统计数据缺乏。创建或获准使用有关言语特性的人口统计数据，是对 LR 中普遍性进行量化的必要前提。然而，目前鲜有与语音有关的参考人口数据却是不争的事实，对此有几个方面的问题值得大家注意：首先，相关人口如何定义？LR 支持者承认这是一个紧迫的现实理论问题，认为相关的参考人口应该因个案而定，一般来讲，需要考虑说话人的性别、所说语言、方言等因素建立参考数据库，至于参考人口中取多少样本为宜的问题，答案是取决于要求的精度[7]。如此原则性的说法操作性很差，当然 Rose[73]推荐了一种收集参考数据的方法，即让鉴定专家们将他们的参考人口数据汇集到一起，从而满足每个特殊案件的需要，当然可能还需要收录其他说话人的语料。由于不同专家的选择标准和数据范围很难保持一致，因此这种方法看似容易，但其随意性很强，而且从可行性的角度讲，这种方法能够考虑到可以被量化的特征只有非常有限的一部分，忽视了在对话交流过程中其他丰富而复杂的信息，以及对其进行解析分析的可能性。而且 Rose 的例证只是对 30 个说话人的数据进行了分析，从说话人讲“yes”单词中提取元音共振峰轨迹的角度说明了收集数据的可能性，并未论及其他特征。其次，更为重要的是所建数据库中应该包括哪些特征、控制哪些因素等问题难以给出准确的具有可操作性的答案。如上文所述，鉴定中可能涉及到的特征涵盖范围很广，远远超出了共振峰特征的范畴，试图收集和分析足够多的能够包含全部（或绝大部分）特征的参考数据基本是不可能的。收集数据过程中同时必须考虑的一个问题是录音方式（或信道问题），实践中的录音方式非常多，有麦克风、录音笔、录音机等直接录音的，也有固定电话、手机（GSM 、CDMA、3G、小灵通）、IP 网络电话之间不同组合通话录音的（以后还会出现新的形式，如最近流行的微信留言形式），很难想象制作参考数据时能够包含上述全部录音方式。而且说话人的说话状态（大声说话、吸烟和醉酒等）和健康情况等因素都必须加以控制。再次，在像中国这样语言环境非常复杂，包含众多语言、方言和次方言的国家，要想建立全面的参考人口数据是非常困难的[27]，同时技术、资金都是非常大的限制因素。最后，还需要考虑的是参考数据的“保鲜”问题，不可否认所有的语言和方言都处在不断变化之中，因此任何参考数据就其适用性而言，其“保质期”都是有限的，如果将检材语音与过时的参考数据进行比较的话，无疑会带来很大的风险，对于检材语音中的某一关键特征来说，参考数据库中的数据可能不一定会有充分的代表性[12]。上述担心是必要的，因为众多因素都会对说话人的语言学/语音学特征模式产生显著影响，无法想象汇集或创建能够包含各种语音特征、考虑各种影响因素的合适的参考人口统计数据是可行的，即使是从长远的角度看。

综上可见，尽管LR体系在逻辑上没有错误，但是由于声纹鉴定专家一般无法量化大部分语音特征，使得 LR 的准确计算并不现实，而且“对于绝大多数语音特征而言，都缺乏相应的参考人口数据，那种用数字表示的似然比只会对分析提供虚假的量化表象，实际上分析还是严重依靠（人为）判断”[14]P388。尽管如此，我们也不能完全排斥 LR 体系（包括贝叶斯原理），其现代思想已经得到了广泛认可，并在 DNA 证据中得到了很好的应用，然而，将此模式强加给包括声纹鉴定在内的其他所有的法庭科学证据是否合适是我们值得考虑的问题，正如很多专家提到的：对

26 Nolan 曾经报道过对双胞胎进行区分的案例，由于双胞胎的声道结构非常相似，以至于二者的语音（听觉和声学）参数非常一致，如果只从声学上进行分析的话，很难进行区分，可以设想如果只分析共振峰参数的话很可能会得到数值很大的 LR，从而得出错误的鉴定意见（支持同源假设，即二者是同一人），最终区分还是主要依靠听觉分析完成的：二者一致性地在/r/音位上发音不一致，一个发成唇齿近音[ʋ]，另一个发成齿龈后近音[ɹ][72]。

27 这里讲的建立参考人口数据库的做法与从事普通方言学研究的专家所做的方言调查（或者建立方言数据库）是不同的，前者是以服务办案为目的，后者则是要研究某个方言的语音、音系、语法等完整的语言学特点，通常包含的发音人数也很少。

于声纹鉴定而言，贝叶斯原理的价值或许不在于直接应用，而在于给鉴定人员提供一个有用的概念性框架，使得专家和事实审判者之间责任区别开来[13, 16, 17, 25]。

对于声纹鉴定专家来说，尽管的确应该向使意见描述更严密的方向努力奋进，但是我们也必须要知道什么是现实的，什么是不现实的，这可能也是促使英国立场声明产生的动机之一。

4.4 英国立场声明形式

在英国立场声明形式产生之前，通常情况下，在英国的法律体系中，语音学家常用诸如经典的可能性等级表述方法来表达专家意见，之所以实现这种转变，主要是受贝叶斯理论（包含 LR 体系）的影响，英国的专家们不仅认识到经典的可能性等级表述形式中存在的逻辑缺陷——只考虑到两个竞争性假设中的一个假设，同时也意识到了 LR 方法中存在的诸多无法克服的现实问题（见 4.3 部分），继而在广泛讨论的基础上提出了英国立场声明形式。值得肯定的是，该形式吸收了贝叶斯理论的主要思想，要求鉴定人同时考虑认定说话人过程中的两个竞争性假设，然而正如 French 等在给 Rose 和 Morrison 的回应[12]P144 中表述的：“文件中提出的框架为的是提醒鉴定专家需要判断检材语音和样本语音中特征的独特性（相当于普遍性），这就意味着要与更多的人进行比较，虽然这种比较是非正式的，常通过分析人员的经验和一般的语言学知识进行的，而不是正式的、定量的。”不难看出，这是一种折中的做法，在既有鉴定方法（主要是听觉-声学分析法）的基础上融合了贝叶斯原理的基本思想来表述鉴定意见，该形式依旧不可避免的存在一定的主观性，不等同于定量化的 LR 体系。由此，也受到了 LR 体系支持者的批评。

第一，该形式中两个阶段——一致性和独特性——的划分相当但不等价于 LR 体系中的相似性和普遍性，两者不是并列平行的，而是连续有序的，独特性分析只是在判断为“一致”的基础上再进行，而且两者的划分等级标准也不相同（前者分为三级，后者则分为五级），两者也不直接相关。对此，Rose 和 Morrison[7]批评说该方法实际上无法测量出检材语音和样本语音来源于同一说话人还是不同说话人的可能程度是否相等。

第二，能否做出绝对的认定/否定意见？按照贝叶斯原理的规定，对于声纹鉴定专家来说，任何绝对肯定和否定都是基于后验概率的，在逻辑上都是错误的，因为先验概率并不可知。对此，LR 体系支持者和英国立场声明形式拥护者们互相指责对方都曾犯过这样的逻辑错误。首先英国立场声明形式在谈到“不一致”的结果时，认为“当语料之间不一致时，得出语料是不同人所说的判断在逻辑上是没有缺陷的。”[11]P141 对此，其拥护者后来承认这种论述是有问题的，但是强调这种做法在实际情况中是合理的，同时指出其批评者在论述中也存在同样的缺陷[12]P144，即“当然，可以想象，在特定情况下，语音比对也可以得出明确的排除结论，如一个幼小儿童的声道不可能产生典型成年男性较低的共振峰，但是这种情况下，两个语音听起来会明显不同，以至于向鉴定专家咨询这样的案件几乎是不可能的。”[7]P150 对此论述的反对意见是 LR 体系支持者自己也得出了一个绝对的（明确的）结论，而不是通过得出一个比率来支持说话人是一个成年男性而非一个儿童的假设[12]P145。英国立场声明形式中还规定了闭集鉴定的特殊例外情况，“在这种案件中，比对任务就变成了哪个人说了什么话。此时，如果几个声音之间的差别足够明显，我们认为可以做出绝对认定的判断。”[11]P142。Rose 和 Morrison[7]P151 批评该做法依然违背贝叶斯原理，并认为应该将此闭集情况同样作为开集来进行检验。其实，在实践办案中常常遇到推定某些语料是被鉴定人所说的情况，比如送检人常提出的鉴定要求是对一段语料中的男性/女性/被称为“XX”的/打电话的/报警的等类型的说话人的声音是不是张三所说，其中检材语音往往都是对话的形式，鉴定专家在选择分析对象（特定语音片段）的时候，不得不将无关的（对应上文）“女性/男性/被称为‘YY’的/接电话的/接警的”说话人的声音排除掉，这个过程实际上就是推定某些语音是一个人（与是哪个具体的个体无关）说的，其他语音不是这个人说的，这种推定的绝对认可与排除是必须的，否则鉴定专家将无法适从，到底应该对哪些语音进行分析呢[28]？总不能让事实审判者充当送检人的角色，来指定就某一句话进行鉴定吧？对于检材语音中包含多个对话人的情况（嫌疑人可能是一个或几个），更是需要

[28] 可以想象如果两个（或多个）对话人的声音非常接近，如双胞胎，这种情况下的推定将变得非常困难，幸运的是这种情况很少在实际案件中发生。

对其中的“不同”语音进行分类，然后再根据鉴定要求对其中的一个或多个分类后的语音（语音片段）进行进一步的分析，这其中都暗含了（闭集）绝对肯定和否定的思想。从这个角度看，英国立场声明形式中规定的闭集鉴定的特殊例外也不无道理。

第三，没有解决针对多个特征进行综合评价的问题。鉴定中需要使用多个特征是一个不争的事实，如果多个特征指向一致还好说，但是如果出现“九个特征上判定为一致，而一个特征上判定为不一致时，该如何处理？”[7]P151 对于这样的情况，英国立场声明形式和经典的可能性等级表述形式似乎都没有给出很好的说明，一般的做法是按照现有的语音学/语言学模型进行检验，看特征之间的相似或差异是否可以得到合理解释，然而这种解释无疑可能会在不同水平的鉴定人之间出现差异，这也是主观性的体现。

第四是没有明确给出不同等级间的划分依据。不难看出，英国立场声明形式没有给出区分不同等级的具体标准，语音特征组合之间的相似/差别多大才能判定为“一致/不一致”？特征组合需要多特殊才能达到“高度特殊”？不同鉴定人基于自己的经验及所掌握的语言学知识，对同一特征组合能否得出同样的“独特性等级”？英国立场声明形式的拥护者无法准确回答这些问题，当然他/她们并不否认这些问题的存在[12]P145，然而问题还是应该回归到其拥护者使用的鉴定方法上来，比如听觉-声学分析法会带有鉴定专家的主观判断，由此方法得出的英国立场声明形式（其他形式也一样）的鉴定意见也不可避免的带有主观成分。换句话说，就目前掌握的知识程度和技术方法来讲，实际上很难给出一个明确的划分依据。

4.5 选择鉴定意见表述形式中的价值问题

由于目前我们所掌握的技术水平和鉴定方法的限制，加之不同鉴定专家在鉴定中选择不同特点的鉴定方法（或组合），使得各种意见表述形式都存在着各自的优点和不足，迄今还没有任何一种表述形式是无争议的，那些被称为无论是“最科学的”还是“最可靠的”叫法都是自封的，违背了科学的精神。对于任何法庭科学来说，其鉴定意见不仅要体现在其“科学”的一面，还必须能够同时被“法庭”（或特定国家的“司法体系”）这一领域所利用，在“法庭”与“科学”二者结合的时候，必然不是 1+1=2 这样简单的数学运算，会涉及到很多社会学方面的价值判断。即使不同国家之间的技术水平相同，如果其他条件不同的话，做出的价值选择也可能不同。对于声纹鉴定（其他法庭科学也一样）而言，使用何种意见表述形式，不仅是一种科学技术问题，同时还是一种法律上的价值判断问题，在做价值选择的时候，需要考虑以下几点：表述的科学性、现实性（能否达到预期目标）、实用性、代价高低、事实审判者（法官或陪审团）的接受程度、与本土法律体系相融合的协调性等。合理的意见表述形式应该尽可能多地满足上述众多价值选项。

LR 形式在逻辑上没有错误，符合法庭科学的现代思想，然而由于目前其所用鉴定方法的限制，加之缺乏量化所必须的参考人口数据，使得在目前的条件下该方法的“现实性”大打折扣，LR 支持者反复强调其优势是用量化的数值只对证据的力度做出评估，然而，很难想象对于既非语音学家/信号处理专家，又非统计学家的事实审判者来说能够真正理解一个数字（如 LR=123）的内涵[29]，那些让统计学家对每个案件都出庭帮助解释 LR 的建议也不具可操作性。同时，这种只关注“完全量化”的数据，忽视法庭对其接受程度的做法，可能会陷入“物理嫉妒”或“科学至上”的怪圈[75]，现实中并不可取。不得不提的是，LR 形式的支持者常常以“Daubert”规则[30]作为必须采用 LR 体系的主要法律依据，殊不知“Daubert”规则的适用并非像 LR 体系支持者宣扬的那样死板，实际情况是“审判法官可以考虑 Daubert 案中提到的一个或多个更为具体的因

---

[29] Coulthard 也有同样的担心：即使能够计算出似然比，我们仍需要知道如何评估它的重要性。作者担心对于外行的陪审团来说，他/她们是否真能成功应付（cope with）似然比，或者甚至说[似然比]是否又简单地引入了更多的不确定性。迄今为止，没有研究表明对于个体陪审团成员来说，他/她们是否能够成功应对大范围的似然性数字（large likelihood or probability numbers）。尽管专家的报告中完全可以有准确的数字，但是陪审团并不看报告，而且大多数律师对这些似然性（probabilities）也并不满意，到头来，陪审团还是不得不得出一个语义编码的（semantically encoded）结论。[74]P483-484 与此同时，无法保证 LR 体系的不同鉴定人针对同一案件会得出完全一致的 LR 数值，因为鉴定人在选择特征、提取数据、选择参考人口数据库时不可避免地具有一定的主观性。

[30] “Daubert”案（Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993)）的法官认为专家证言的可采性取决于以下 4 方面的因素：1、形成专家证言所依靠的科学理论与科学方法是否建立在可检验的假设之上；2、形成专家证言所使用的科学理论与科学方法是否与现有的专业出版物当中记载的原理相同；3、有关理论的已知的或者潜在错误率以及该理论现存的研究标准；4、指导相关理论的方法论及研究方法为相关科学团体接受的程度[76]P32-33。

素，如果这么做将有助于确定证言可靠性的话。可靠性的标准是‘灵活的’，Daubert 案关于具体因素的清单，既非必要性也非排他性地适用于所有的专家或者每个案件。”[31][77]P222 最新修改的美国联邦证据规则第 702 条[32]也并未将“已知的或者潜在错误率”写进条文中去。其实，由于价值选择的不同，在美国只有部分州法院采用了该规则[33]（截止 2004 年，有 11 个州采用了“Daubert”规则[78]），其他国家也并无此同样要求的证据规则。尽管我们不应该排斥吸收其中的科学成分完善自己国家的证据采信规则，但是将其作为任何国家法庭科学从业者的行动纲领也不是应有的态度，更何况是片面而错误的理解。

经典的可能性等级表述形式易于理解，实用性很强，目前被众多国家/专家广泛采用也证明了这一点，然而其在逻辑上存在着缺陷，也受到不少学者的批评，但是这种批评并非是广泛的，至少目前是这样。尽管其表述形式是基于后验概率的，在 LR 体系支持者看来，鉴定专家侵犯了事实审判者的最终裁定权，但是实践中并非所有的事实审判者都反对专家的这种“侵权”行为，至少在中国大陆的情况是，一方面很多时候部分法官会期待这种“侵权”的发生，因为他们更喜欢“明确/绝对”的鉴定意见，另一方面，即使是鉴定专家给出了“明确/绝对”的“鉴定结论”，该“结论”实际上只是专家的判断意见，不一定就会成为法官最后定案的裁定结论，鉴定专家很可能需要就其鉴定意见进行出庭作证，接受当事人（鉴定结果对其不利的一方）、律师和法官多方面的质疑，专家意见是否得到采信的最终决定权还在法官手中。在采用可能性等级表述形式的国家（如德国[61]），经过长期的司法实践，法官常常习惯于这种后验的表述形式，并对此表示满意。如果事实审判者基于该形式的实用性保持这种满意态度，那么改革或摒弃这种逻辑上存在缺陷的形式将失去根本动力，“逻辑正确”的价值选择将让位于“实用性”的价值选择。

英国立场声明形式是一种“折中”的价值选择。之所以做出这种选择，是因为它必须纠正可能性等级表述形式中存在的逻辑缺陷，同时基于现有科学技术、鉴定方法在“科学性”上的局限性，认识到 LR 形式主张的“完全量化”的“客观”表述并非完全客观，且在实践中不具现实性，所以不得不仅在概念上引入了贝叶斯原理的现代思想，使得该形式在逻辑上变得正确。尽管该形式并未改变其表述存在主观性的本质，但是其现实性和实用性却得到了英国各界的广泛认可。

各国在基本国情、司法体系和技术水平等方面存在的差异，决定了在科学证据上采取的价值取向也不尽相同，无论何种形式的声纹鉴定意见归根结底还是要为本土法律服务，为现实服务，那种脱离实际仅停留在理想层面的鉴定意见不应该成为我们的价值选项。

4.6 我国现有形式是完美的吗？

公安部方法和司法部规范的颁布对于规范我国的声纹鉴定技术起到了的非常重要的作用，因为之前（2006 年左右）很少有对该问题进行过详细的文件化的描述。然而，正如上文所述，作为典型的可能性等级表述形式都存在逻辑上的缺陷和表述上的主观性，我国现行的两个方法中对鉴定意见的表述也都存在这样的问题，针对此问题，目前的情况是，除了个别 LR 体系支持者对其进行批评之外，大部分声纹鉴定从业人员和法官对现有模式表示满意，并未对其提出整改要求。当然，在我们认识到国内外现有鉴定意见表述形式的优缺点之后，还是需要再做出一次价值选择，是固守或调整逻辑上有缺陷但实用性很强的可能性等级表述形式，还是转向逻辑正确但难以真正实现的量化的 LR 形式？或许英国立场声明形式会给我们一些启示。

5. 未来方向

声纹鉴定仍然是一门比较新的技术，尽管业内很多专家和学者已经积累了丰富的经验，得出了许多研究成果，但是还有很多领域值得做进一步的研究。今后重点研究的任务是寻找那些个体稳定性更强、人际差异性更大的语音特征；在现有鉴定方法的基础上，整合各种方法对说话人鉴定的技术优势，重点开展一些合作性的基础研究；有可能的话，我们更期待创建新的更可靠的

---

[31] 该结论（“Daubert”规则的适用问题）是由美国联邦最高法院在判例“锦湖轮胎”案（Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999)）中得出的，参见[77] P222。

[32] 美国联邦证据规则第 702 条是对专家证人证言的规定：在下列情况下，因知识、技能、经验、训练或者教育而具备专家资格的证人，可以以意见或者其他的形式就此作证：(a) 专家的科学、技术或者其他专门知识将会帮助事实审判者理解证据或者确定争议事实；(b) 证言基于足够的事实或者数据；(c) 证言是可靠的原理和方法的产物；以及 (d) 专家将这些原理和方法可靠地适用于案件的事实。参见[77] P212。

[33] 由于“Daubert”规则是联邦普通法对联邦规则（联邦制定法）所做出的解释，因此该规则不具有普遍效力，是否适用该规则由各州法院自行决定[76] P33。

分析方法，尽管这是很难的事情。对于传统的听觉-声学分析法而言，由于目前很多研究依旧是基于较少发音人、孤立的、零散的、不合作的研究，对提高声纹鉴定的技术水平贡献有限。一个比较好的研究方式是首先尝试建立一些较大规模的数据库，限定发音人的方言、地域、年龄、性别等基本条件，人数在百人(或数百人)以上，使用多种信道进行录音，包含多种实验设计，设定数据共享机制，在此基础上展开实验室间合作性的基础研究。剑桥大学 Nolan 教授最近主持的一个研究项目“语音的动态性变化研究”[34]就是很好的范例。另外我们也期待说话人自动识别技术在不远的将来取得更大的发展，技术的进步可能会得益于语音分析专家与信号处理工程人员的深入结合，在解决文本内容、信道鲁棒性、检索速度等关键问题的基础上，建立大规模的声纹数据库。然而，在可预见的时间内，实践中说话人自动识别技术依旧要依靠鉴定专家的广泛参与。总之，只有从特征上、方法上创新，才能真正提高声纹鉴定的科学性和准确性，鉴定意见的表述则应该是下位的问题，因为只有从技术方法的基础研究中入手，才有可能从根本上解决鉴定中的诸多问题，如特征变化、特征价值、特征数量、特征量化、不同方法各自及整合后的准确率等等，只要这些问题解决了，关于鉴定意见如何表述才是更加科学的、可靠的、客观的、可行的问题就不难找到答案了；反之，在没有鉴定方法科学性的基础上，只谈鉴定意见表述的科学性注定难以达到科学性的目标。

6. 结论

本文针对目前讨论比较热烈的声纹鉴定意见表述问题进行了评述，首先比较全面地介绍了目前实践中正在使用的 5 种鉴定方法，指出了各种鉴定方法的优缺点，然后对现存的 4 种形式的鉴定意见表述形式进行了介绍和评析。本文无意在此将鉴定意见到底应该如何表述的问题彻底解决，只希望本文的论述能够引起读者的思考，相信关于该问题的争论还会继续。最后将几点重要结论总结如下：

第一、经典的可能性等级表述形式简单易懂，目前被大多数国家的声纹鉴定人员所采用。然而，使用该表述形式的鉴定人主要考虑了检材语音和样本语音特征之间的相似性，没有将特征的相似性放到整个参考人口中去考查其分布问题（特征的普遍性），由此得出的鉴定意见在逻辑上存在一定的缺陷，而且主观性较强，因此也受到很多专家的批评。

第二、似然比（LR）形式由于在鉴定中同时考虑了两个竞争性假设，即对特征的相似性和普遍性都进行了考查，而且该形式不对检材语音与样本语音是否是同一人所说做出绝对判断，只是使用量化的 LR 数值对语音证据的证明力度做出评价，因此在逻辑上是正确的。然而，由于缺乏相关的参考人口统计数据，使得完全量化的 LR 方法在鉴定实践中不具现实性；同时，在目前现有的鉴定方法中，听觉-声学分析方法仍然占据主导地位，大多数非声学语音特征无法得到准确的量化统计，目前基于某种特定分类的小样本的参考人口数据，仅仅使用声学分析对部分特征进行量化便得出结论的做法是不负责任的。进一步说，由于语音的特殊性，那些认为构建可以被视为具有可行性的、严密的、排他的定量 LR 方法之前，仅仅是时间和研究问题的观点是不现实的。然而，LR 体系毕竟是个新生事物，对此我们也要保持足够的开放心态，其适用性与可靠性最终需要实践来检验。

第三、随着时代的进步和技术的发展，我们似乎不得不在“简单易懂但逻辑上有瑕疵的”形式与“逻辑正确且须具有切实可行性的”表述方式之间做出价值选择，由于完全量化的 LR 形式在复杂多变的语音证据上难以真正实现，我们可能不得不另寻途径吸收和体现贝叶斯原理的现代思想，或许在不久的将来，我们也可能会转向使用 LR 方法，但不是完全数字的量化形式，可能只是主观的、定性的。此时，英国立场声明形式[11]或许能够给我们一些启示。然而，不同国家的司法体系和现实国情不同，鉴定意见的表述可以有本土化的体现，做出不同的调整。

第四、要提高鉴定意见表述的科学性和可靠性，我们不仅要吸收贝叶斯原理的现代思想，更应该从语音特征和鉴定方法的创新中入手，我们期待不同语音分析专家之间、语音分析专家与

[34] 项目英文名称是“Dynamic Variability in Speech (DyViS)”，网址：http://www.ling.cam.ac.uk/dyvis/。该项目建立了一个包括 100 位年龄在 18-25 岁之间的讲英国南部标准英语口音的男性说话人的数据库，该数据库中包含了多种对话情景和录音方式，同时该数据库对于非商业用途的研究者来讲是免费开放的。该项目的详细介绍，可参见[79]。欧洲及其他部分国家的（无论是支持还是反对 LR 体系的）声纹鉴定研究人员经常使用这个数据库进行相关实验研究，以 2013 年的 IAFPA 年会为例，有 20%强的文章中的语料都是取自 DyViS 数据库。

信号处理工程师之间开展合作性的基础研究，在不久的将来有成形的研究出现，或许到那个时候关于鉴定意见如何表述才是更加科学可靠、客观可行的问题才更容易回答。

致谢

感谢施少培（正）高级工程师、李英浩副教授、王虹副教授、许毅教授对本文初稿提出的很有价值的修改意见，感谢 Michael Jessen 博士与 Paul Foulkes 教授在 IAFPA 2013 年会期间就此问题与作者的有益讨论。文中错误之处由本文作者承担，与上述专家无关。

参考文献

[1] Saks, M.J. and Koehler, J.J.The coming paradigm shift in forensic identification science[J]. Science, 2005. **309**(5736): 892-895.

[2] National Reasearch Council. Strengthening forensic science in the United States: A path forward[M]. Washington, DC: The National Academies Press, 2009.

[3] Law Commission. Expert Evidence in Criminal Proceedings in England and Wales(Law Com. No. 325) [M]. London: The Stationery Office, 2011.

[4] 王英利，李敬阳，曹洪林. 声纹鉴定技术综述[J]. 警察技术，2012(04): 54-56.

[5] Rose, P., Forensic speaker identification [M]. London and New York: CRC Press, 2002.

[6] Rose, P., Technical forensic speaker recognition: Evaluation, types and testing of evidence [J]. Computer Speech & Language, 2006. **20**(2-3): 159-191.

[7] Rose, P. and Morrison, G.S. A response to the UK position statement on forensic speaker comparison[J]. International Journal of Speech, Language and the Law, 2009. **16**(1): 139-163.

[8] Morrison, G.S., Forensic voice comparison and the paradigm shift[J]. Science & Justice, 2009. **49**(4): 298-308.

[9] Morrison, G.S., Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework [J]. Australian Journal of Forensic Sciences, 2009. **41**(2): 155-161.

[10] Morrison, G.S., Forensic voice comparison[A], in Expert Evidence[M], Freckelton I. and Selby, H. Editors. 2010.

[11] French, P.and Harrison, P., Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases[J]. International Journal of Speech Language and the Law, 2007. **14**(1): 137-144.

[12] French, P., Nolan, F., Foulkes, P., et al., The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison[J]. International Journal of Speech Language and the Law, 2010. **17**(1): 143-152.

[13] Nolan, F., Speaker identification evidence: its forms, limitations, and roles[A]. in Proceedings of the conference 'Law and Language: Prospect and Retrospect'[C]. 2001. Levi (Finnish Lapland).

[14] Nolan, F., Voice[A], in Identification: Investigation, trial and scientific evidence[M], Bogan P.S., and Roberts, A. Editors. 2011. 381-390.

[15] Foulkes, P. and French, P. Forensic speaker comparison: a linguistic-acoustic perspective[A], in The Oxford Handbook of Language and Law [M], Tiersma, P. and Solan, L. Editors. 2012. 557-573.

[16] Broeders, A.P.A., Some observations on the use of probability scales in forensic identification[J]. Forensic Linguistics, 1999. **6**(2): 228-241.

[17] Broeders, A.P.A., Forensic speech and audio analysis,forensic linguistics. A review: 2001 to 2004[A]. in 14th INTERPOL Forensic Science Symposium[C]. 2004. Lyon, France.

[18] Eriksson, A., Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work[A], in Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism[M], Neustein A.and Patil H.A., Editors. 2011, Springer. 41-69.

[19] Eriksson, A., Presenting evidence in court - some fundamental problems to be considered[A]. in Proceeding of International Association for Forensic Phonetics and Acoustics Annual Conference [C]. 2011. Vienna, Austria.

[20] 张翠玲，Rose, P. 基于似然率方法的语音证据评价[J]. 证据科学，2008(03): 337-342.

[21] 张翠玲. 法庭语音技术的最新发展[A]. 第二届证据理论与科学国际研讨会[C]. 2009. 中国北京.

[22] 张翠玲. 法庭语音比较的科学性和可靠性[A]，证据理论与科学:第三届国际研讨会论文集[C]，常林，张中主编. 北京: 中国政法大学出版社，2012.

[23] 张翠玲. 法庭语音技术研究[M]. 北京: 中国社会出版社，2009.

[24] 季云起. 声纹鉴定结论的科学性表述模式[J]. 西南大学学报（社会科学版）, 2009(05): 122-125.

[25] Jessen, M., 曹洪林, 王英利（译）. 法庭语音学[J]. 证据科学, 2010(06): 712-738.

[26] Köster, O. and Köster, J., The auditory-perceptual evaluation of voice quality in forensic speaker recognition[J]. The Phonetician, 2004(89): 9-37.

[27] Köster, O., Jessen, M., Khairi, F., et al. Auditory-perceptual identification of voice quality by expert and non-expert listeners[A]. in Proceedings of the 16th international congress of phonetic sciences (ICPhS XVI) [C]. 2007. Saarbrücken.

[28] Laver, J., The phonetic description of voice quality[M]. Cambridge: Cambridge University Press, 1980.

[29] Nolan, F., Auditory and acoustic analysis in speaker recognition[A], in Language and the law[M], Gibbons, J. Editor. Longman: London/New York. 1994: 326-345.

[30] Nolan, F., Forensic speaker identification and the phonetic description of voice quality[A], in A Figure of Speech: a Festschrift for John Laver[M]. Hardcastle, W.J. and Beck, J.M. Editors. Lawrence Erlbaum Associates, Inc.: New Jersey and London. 2005: 385-411.

[31] Kersta, L.G., Voiceprint Identification[J]. Nature, 1962. **196**(4861): 1253-1257.

[32] Tosi, O., Oyer, H., Lashbrook, W., et al., Experiment on Voice Identification[J]. The Journal of the Acoustical Society of America, 1972. **51**(6B): 2030-2043.

[33] Tosi, O., Voice identification: theory and legal applications[M]. Baltimore: University Park Press, 1979.

[34] Nolan, F., The phonetic bases of speaker recognition[M]. Cambridge: Cambridge University Press, 1983.

[35] Hollien, H.F., The acoustics of crime: The new science of forensic phonetics[M]. New York: Plenum Press, 1990.

[36] Hollien, H.F., Forensic voice identification[M]. San Diego: Academic Press, 2002.

[37] International Association for Forensic Phonetics and Acoustics, IAFPA Resolution - Voiceprints (http://www.iafpa.net/voiceprintsres.htm) [EB/OL], 2007. [cited 2013 August 24]

[38] 美国国家研究理事会, 丁宁（译）. 嗓音鉴别的理论与实践[M]. 北京: 群众出版社, 1989.

[39] Morrison, G.S., Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison (in press) [J]. Science & Justice, 2013.

[40] Koenig, B.E., Spectrographic voice identification: A forensic survey[J]. The Journal of the Acoustical Society of America, 1986. 79: 2088-2091.

[41] Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identification. Voice Comparison Standards[J]. Journal of Forensic Identification, 1991. **41**(5): 373-396.

[42] Cain, S. American Board of Recorded Evidence-Voice Comparison Standards. (http://www.forensictapeanalysisinc.com/Articles/voice_comp.htm) [EB/OL], 1998. [cited 2013 August 24].

[43] Maher, R., Audio forensic examination: Authenticity, enhancement, and interpretation[J]. Signal Processing Magazine, IEEE, 2009. **26**(2): 84-94.

[44] Nakasone, H. and Beck. S.D. Forensic automatic speaker recognition[A]. in Proceeding of Speaker Odyssey Speaker Recognition Workshop[C] 2001.

[45] Coulthard, M. and Johnson, A. An introduction to forensic linguistics: language in evidence[M]. London and New York: Routledge, 2007.

[46] Archer, C. HSNW conversation with Hirotaka Nakasone of the FBI: Voice recognition capabilities at the FBI -from the 1960s to the present (http://www.homelandsecuritynewswire.com/bull20120711-voice-recognition-capabilities-at-the-fbi-from-the-1960s-to-the-present) [EB/OL], 2012. [cited 2013 August 24].

[47] Branca, A. Zimmerman Case: Dr. Hirotaka Nakasone, FBI, and the low-quality 3-second audio file (http://legalinsurrection.com/2013/06/zimmerman-case-dr-hirotaka-nakasone-fbi-and-the-low-quality-3-second-audio-file/)[EB/OL]. 2013. [cited 2013 August 24].

[48] Solan, L.M. and Tiersma, P.M. Hearing voices: Speaker identification in court[J]. Hastings LJ, 2002. **54**(2): 373-435.

[49] Solan, L.M. and Tiersma, P.M. Speaking of crime: The language of criminal justice[M]. Chicago and London: University of Chicago Press, 2005.

[50] Tiersma, P.M. and Solan, L. The linguist on the witness stand: forensic linguistics in American courts[J]. Language, 2002. **78**(2): 221-239.

[51] Schwartz, R. Voiceprints in the United States -Why they

won't go away[A]. in Proceeding of International Association for Forensic Phonetics and Acoustics Annual Conference [C]. 2006. Göteborg, Sweden.

[52] 王英利. 关于声纹鉴定技术的若干问题[A]. 第九届中国语音学学术会议论文集[C]. 2010. 中国天津.

[53] Gfroerer, S. Auditory-instrumental forensic speaker recognition[A]. in Proceedings of Eurospeech 2003[C]. 2003. Geneva, Switzerland.

[54] 李敬阳. 说话人鉴定概述[A], 第一届全国视听技术检验学术交流会论文选[C], 公安部物证鉴定中心, 北京: 中国人民公安大学出版社, 2007: 281-286.

[55] 李敬阳, 音像物证技术（二）：声音及其鉴定[A], 物证技术学[M], 李学军, 刘晓丹主编. 中国人民大学出版社: 北京. 2011.

[56] Nolan, F., Forensic Phonetics[J]. Journal of Linguistics, 1991. **27**(2): 483-493.

[57] 李敬阳, 胡国平, 王莉. 声纹自动识别技术与声纹库建设应用[J]. 警察技术, 2012(04): 66-69..

[58] Gold, E. and French, P., International Practices in Forensic Speaker Comparison[J]. International Journal of Speech Language and the Law, 2011. **18**(2): 293-307.

[59] 司法部司法鉴定科学技术研究所, 2010 司法鉴定能力验证鉴定文书评析[M]. 北京: 科学出版社, 2011.

[60] Utlåtandeskalan (http://www.skl.polisen.se/Global/www%20och%20Intrapolis/Informationsmaterial/SKL/ Utlatandeskalan.pdf) [EB/OL]. 2008. [cited 2013 August 24].

[61] Jessen, M. Conclusions on voice comparison evidence in Germany and a challenging case[A]. in Proceeding of International Association for Forensic Phonetics and Acoustics Annual Conference [C]. 2011. Vienna, Austria.

[62] Boë, L.-J., Forensic voice identification in France[J]. Speech Communication, 2000. **31**(2-3): 205-224.

[63] McDermott, M.C., Owen, T. and McDermott. F.M. VOICE IDENTIFICATION: The Aural/Spectrographic Method. (http://www.tapeexpert.com/pdf/voiceidauralspectro.pdf) [EB/OL]. 1996. [cited 2013 August 24].

[64] 王宁敏. 司法语音与声学检验[M]. 北京: 中国检察出版社, 2009.

[65] 王虹, 案件言语识别与鉴定技术规范[M]. 北京: 中国人民公安大学出版社, 2012.

[66] 曹巧玲, 视频中的声音运用[A], 视频检验技术规范[M]. 杨洪臣编著. 中国人民公安大学出版社: 北京. 2012.

[67] Cambier-Langeveld, T., Current methods in forensic speaker identification: Results of a collaborative exercise[J]. International Journal of Speech Language and the Law, 2007. **14**(2): 223-243.

[68] Champod, C. and Evett, I.W. Commentary on APA Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6 (2): 228-41[J]. Forensic Linguistics, 2000. **7**(2): 239-243.

[69] Forensic-Speech-Science.info, (http://www.forensic-speech-science.info/)[EB/OL]. 2007. [cited 2013 August 24]

[70] 公安部物证鉴定中心, 语音同一认定方法（IFSC 11-01-01-2010）[S], 2010.

[71] 中国人民共和国司法部司法鉴定管理局, 录音资料鉴定规范（SF/Z JD0301001-2010）[S], 2010.

[72] Nolan, F. and Oh, T. Identical twins, different voices[J]. Forensic Linguistics, 1996. 3: 39-49.

[73] Rose, P., Going and getting it - Forensic speaker recognition from the perspective of a traditional practitioner-researcher[A]. in the Australian Research Council Network in Human Communication Science Workshop: FSI not CSI - Perspectives in State-of-the-Art Forensic Speaker Recognition[C]. 2007. Sydney.

[74] Coulthard, M., Experts and opinions: In my opinion[A], in The Routledge Handbook of Forensic Linguistics[M], Coulthard, M. and Johnson, A. Editors. Routledge. 2010: 473-486.

[75] Lindh, J., Eriksson, A. and Nelhans. G. Methodological Issues in the Presentation and Evaluation of Speech Evidence in Sweden[A]. in Proceeding of International Association for Forensic Phonetics and Acoustics Annual Conference [C]. 2010. Trier, German.

[76] 徐继军. 专家证人研究[M]. 北京: 中国人民大学出版社, 2004.

[77] 王进喜. 美国《联邦证据规则》(2011年重塑版)条解[M]. 北京: 中国法制出版社, 2012.

[78] Admissibility of Scientific Evidence Under Daubert (http://www.drtomoconnor.com/3210/3210lect01a.htm) [EB/OL]. [cited 2013 August 24].

[79] Nolan, F., McDougall, K., De Jong, G., et al., The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research[J]. International Journal of Speech Language and the Law, 2009. **16**(1): 31-57.

桂东南博白客家话声调建模[①]

关英伟　姚云

摘要 文章以松旺镇客家话声调基频为基础，建立了松旺镇客家话声调模型；利用该模型合成出客家话调值样本，并进行了听辨实验。实验结果表明，声调模型基本合理科学。该模型不仅能够在计算机上动态地、交互式地再现博白松旺镇客家话声调音高、音长和声调的凹凸变化特征的真实形态，而且还能构建和预测声调的发展变化轨迹。在方言保护和研究、方言合成等方面具有方法上的意义和应用价值。

关键词 松旺镇客家话　声调　基频　五度值　建模

博白县，古称白州。位于广西壮族自治区东南部，属于玉林市。东接广东省湛江市，南依北海市，西邻钦州市，北靠玉林市。地处东经109°38′-110°17′北纬21°38′-22°28′。古代曾经是百越人杂居的地区。随着客家人的迁入，在博白形成两种方言：地佬话和新民话。地佬话也称土白话，属于粤方言桂南系次方言，新民话也称涯话，属客家方言。博白新民话与陆川、合浦、浦北和广东省廉江、化州等县的客家方言连成一体，是广西客家方言连片地区较大方言之一。“地佬”一词是当地新迁来的客家人对原住民的称呼，“地佬话”即为原住民讲的话；“新民”，即客家人，“新民话”即为客家人讲的话。

博白县是客家人聚居地，绝大部分居民都是历代从江西、福建、广东迁移过来的客家人，根据第五次人口普查数据，全县总人口1247518人。新民客家话是当地人的主要交际语言，主要在博白县东、中、南部的凤山、新田、宁潭、文地、三江、英桥、大垌、那卜、沙陂，合江、东平、沙河、菱角、松旺、双旺、龙潭、大坝以及西北部的黄凌、三育、江宁等20多个乡镇通行，使用人口约96万，占全县总人口的77%以上。

松旺镇是博白县第二大镇，地处北部湾畔，位于博白县西南部，南邻龙潭镇，东连那卜、大垌镇，西靠沙河、菱角镇，全镇总面积183.5平方公里，全镇辖11个村委会、196个村民小组，总人口5.3万人，全镇通行客家话，除了沙河镇靠近顿谷的山桥和靠近江宁的长远两村村民讲地老话外，绝大多数人以讲客家话为主。讲地老话的人也能听懂客家话。

本文选择博白县松旺镇的新民话作为调查点和声调建模，主要考虑到松旺镇的客家话具有一定的代表性，之前没有人做过音系调查和整理工作。本文的主要发音合作人陈有，男，1985年生，广西师范大学研究生，广西玉林博白县松旺镇人，其家人都讲客家话，都是地地道道的客家人，家庭成员之间交流都使用客家话。

一、博白松旺镇客家话声韵调

1.1 声母（17个，包括零声母）

表1.1　广西博白（松旺镇）客家话声母表

方式 / 部位	清不送气 爆发	清送气 爆发	清不送气 塞擦	清送气 塞擦	清 擦	浊 擦	浊 鼻	浊 近
唇	p	p^h			f	v	m	
齿/龈	t	t^h	ts	ts^h	s		n	l
软腭	k	k^h				x	ŋ	
零声母		∅						

声母说明：（1）/v /的摩擦较轻，声带振动较弱。从语音波形上看，振幅较弱，呈现长三角形状，从语图上看，没有明显的共振峰，浊音横杠比较明显；（2）/n/与细音相拼时，音值接近/ȵ/；（3）/ts，ts^h，s/ 与细音相拼时，音值接近/tɕ，$tɕ^h$，ɕ/。（4）/ŋ/在有些字中自成音节，例如“吴蜈鱼女”。

声母特点：（1）古全浊声母已经清化，无论平仄，大都读同部位的清音声母或送气清音声母。例如：被/ p^hi^{45}/，棋/ k^hi^{24} /，跪/ k^hui^{31}/，下/xa^{33}/，跌/ tet^3/，习/ sip^5/；（2）古非敷奉母字除了少数字仍保留重唇音读法，其它大都清化。例如：飞/fui^{45}/，浮/fou^{24} /，纺/$faŋ^{31}$/;（3）古见溪群母字多读/k，k^h/。例如： 歌/kou^{45}/，捆/ k^hun^{31}/，柜/ k^hu i^{33}/，记/ ki^{33}/；（4）古精庄知章组字多读/ts，ts^h/，例如：租/ tsu^{45}/，只/tsa^{31}/，朝/ts^hau^{24}/，查/ ts^ha^{24}/；（5）古疑母日母字多读/n，

[①]本文为国家社科基金资助项目“广西汉语方言的发声类型研究”（ No：10XYY003）和广西研究生教育创新计划资助项目（No：2010106020501M86）成果之一。

ŋ/，例如：鹅/ ŋo^{24}/，语/ni^{45}/，二/ni^{33}/，硬/ ŋaŋ^{33}/。

1.2 韵母（64 个，含自成音节的 m，ŋ ）

表 1.2 博白松旺镇客家话韵母表

	阴声韵（22）				阳声韵（23）				入声韵（19）			
	-Ø-	-i-	-u-	- y -	-Ø-	-i-	-u-	- y -	-Ø-	-i-	-u-	- y -
1.	ɿ				m ŋ							
	a	ai	ia		am	iam			ap	iap		
2.		au	iau		an	ian	uan		at			
					aŋ	iaŋ	uaŋ		ak			
3.	o	oi			on	ion			ot			
		ou			oŋ	ioŋ			ok			
	e	ei	ie	ui(uei)					ep			
4.		eu			en		un(uen)		et	iet		
					eŋ				ek			
					im				ip			
5.	i	io		uo	in				it	iok	uok	
					iŋ				ik			
		ua										
6.	u	uai	iu		uŋ				ut	iup		
									uk	iuk		
7.	y							yn yŋ				

韵母说明：（1）/oŋ，ioŋ/韵主要元音/o/开口度略大，实际音值为/ɔ/；（2）/eu/韵发主要元音/e/时唇形要更展些。

韵母特点：（1）鼻音韵尾/-m，-n，- ŋ /和塞音韵尾/-p，-t，-k/保留完整，例如：担/tam^{45}/，星/sen^{31}/，瓶/p^{h}iŋ ^{24}/，鸽/kap^{3}/，雪/set^{3}/，吃/sik^{5}/。但是中古阳声韵尾与入声韵尾的相匹配格局已经有所变动，因为在曾摄和梗摄有部分字韵尾为/-n/；（2）古咸摄、深摄平上去声绝大部分字今读/-m/韵尾，相应入声字读/-p/韵尾。例如：范/fam^{33}/，金/ kim^{45}/，踏/ t^{h}ap^{3}/，涩/ sep^{3}/；（3）没有撮口呼韵母。普通话读撮口呼的字博白松旺镇客家话多读齐齿呼。

1.3 声调（6 个，不包括轻声）

表 1.3 博白松旺镇客家话声调表

调类	阴平	阳平	上声	去声	阴入	阳入
调型	高升	中升	中降	中平	中	高
调值	45	24	31	33	3	5

声调特点：（1）共有六个声调，其中平声、入声分阴阳，上声和去声不分阴阳。古平声清声母字大部分今读阴平，浊声母字大部分今读阳平；（2）古上声字绝大部分今仍然读上声;（3）古去声字绝大部分今仍然读去声；（4）古入声清声母字今读阴入，次浊声母字有的读阴入，有的读阳入，全浊声母字今读阳入。

二、博白松旺镇客家话声调建模

本文拟构建的声调模型，是根据声调基频数据构建的数学公式，它反映了声调的调型、调值和时长之间的关系特征。该模型不仅能够在计算机上动态地、交互式地再现博白松旺镇客家话声调音高、音长和声调的凹凸变化特征的真实形态，而且还能构建和预测声调的发展变化轨迹，节约存储空间。在方言保护和方言研究、方言合成等方面具有方法上的意义和应用价值。

2.1 实验说明

2.1.1 语音材料

语音材料为松旺镇客家话，于 2011 年 6 月在陈有家中录音，录音软件 Praat。采样率 22kHz，从松旺镇客家话的阴平、阳平、上声、去声、阴入、阳入六个声调中随即选出 10 个字词，六个声调共得到 60 个实验用字，每个字读两遍，六个声调共得到 120 个样本（6×10×2）。

2.1.2 录音和分析软件

本次实验的录音软件为Praat ，采样率为22kHz，单通道，采样精度为 16 位。全部数据使用 Excel 电子表格进行统计和分析。

2.1.3 声学参数的提取和处理

首先在 Praat 软件上对录音样本进行声调段标注，在标注层上对声调段进行确定和标记。声调段的确定为韵母段，如果实验字（词）的声母是鼻音或边音时，声调段从鼻音、边音后面开始计算。在语图上从元音共振峰起点算起。

接着对基频进行归一化处理，用“音高提取程序”提取每个声调的时长和每个声调 10 个时刻点的基频数据，计算每个音节在 10 个采样点上的原始基频数据的平均值。 再将每个发音人基频数据的平均值转换成对数，最后采用 T 值公式进行五度值转换，得到相对化和归一化的数据。

利用 Matlab 中 *linspace* 和 *plot* 函数对归一后的基频数据进行多项式拟合，做出声调系统图，并建立基频曲线模型，得到松旺镇客家话声调的五度值数学模型，并进行初步的语音合成实验。

2.2 声调基频曲线和五度值转换

2.2.1 基频曲线提取方法和步骤：（1）用归一化方法提取松旺镇客家话 6 个声调的基频数据；（2）将 6 个声调的基频数据的平均值转化为五度值；（3）用多项式拟合的方法得到基频曲线的数学表达式，并

建立模型；（4）绘制声调模型图

2.2.2 松旺镇客家话声调基频曲线

表 2.1 是根据上述实验方法得到的广西博白松旺镇客家话声调基频平均值和平均时长数据，图 2.1 是根据表 2.1 做的松旺镇客家话声调基频曲线。

表 2.1 博白松旺镇客家话声调基频平均值数据表

例字	十个采样点的基频频率平均值（单位 Hz）										时长 单位：ms
阴平	142.1	142.7	144.1	145.2	147.1	148.2	149.9	151.9	154.0	157.9	398.0
阳平	112.0	111.7	111.7	113.2	115.1	117.9	121.1	126.2	131.7	138.2	396.9
上声	125.5	122.5	120.2	116.3	113.6	110.8	107.6	104.1	100.7	99.1	349.3
去声	141.7	140.6	139.3	137.7	137.7	137.1	136.0	135.2	135.1	135.0	416.1
阴入	131.3	129.7	128.1	126.3	124.6	122.8	121.1	118.6	117.5	117.8	193.8
阳入	156.2	156.1	156.1	157.0	158.1	158.9	159.9	160.9	161.7	162.9	195.4

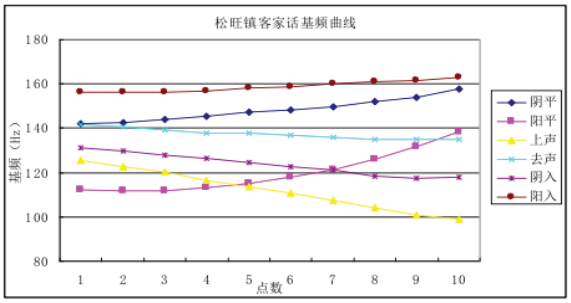


图 2.1 松旺镇客家话基频曲线

从图 2.1 我们可以看到，松旺镇六个声调的基频曲线的走向，阴平和阳平均为升调，升幅不大；阴平比阳平略高，阴平起点为 142.1 Hz，终点为 157.9 Hz，阳平起点为 112Hz，终点为 138.2Hz。上声为降调，起点为 125.5 Hz ，终点为 99.1Hz。去声、阴入和阳入均为平调，阳入分布最高，去声次之，阴入最低。从时长上看，松旺镇客家话入声调比起非入声调时长要短了一半，为 200ms 左右，其中阴入为 193.8 ms，阳入为 195.4ms；非入声调的时长都在 400 ms 左右，其中阴平最长，为 398ms，阳平次之，为 396ms，上声为 349ms。图 2.1 所展示的只是归一化的基频曲线，没有考虑时长的因素。从听感上，阴平、阳平、上声和去声较为舒缓，而阴入和阳入显得较为急促。

2.2.3 松旺镇客家话声调的五度值

从声调形成的生理特征看，声调的音高变化，与声带的松紧及单位时间内声带振动的频率有关，声带拉紧，振动快，频率高，声音就高，反之则低。声调的高低可以用基频来表示，基频是一个声学概念，在语音中基频是乐音周期变化的频率；在语言学上，声调是具有语言学意义即区别词汇意义的基频变化的模式。声调与基频有一定的关系，但不完全对应。用基频来表示声调并不完全符合人耳的听感，也不便于比较，所以，我们要把基频数据转换成符合听感的五度值。

五度标调法是赵元任先生创立的记录声调调值的方法，五度值所描写的调值是相对的，不管基频的绝对频率值是多少，也不管音域本身高低宽窄的变化有多大，一律都归并到相对的五度之中，这是符合人类对声调感知的客观实际的。转换公式见公式 1：

$$\frac{\log_{10}(X)-\log_{10}(Min)}{\log_{10}(Max)-\log_{10}(Min)}\times 4+1$$

（公式 1）

公式中 X 代表一个基频数据，Min 代表一组基频数据中的最小值，Max 代表一组基频数据中最大值。表 2.2 就是根据公式 1 得到的五度值数据。

表 2.2 博白松旺镇客家话声调五度值数据表

例字	五度标记法表示 （公式 $\frac{\log_{10}(X)-\log_{10}(Min)}{\log_{10}(Max)-\log_{10}(Min)}\times 4+1$，保留一位有效数字）										时长 单位：ms
阴平	3.9	3.9	4.0	4.1	4.2	4.2	4.3	4.4	4.6	4.8	398.0
阳平	2.0	2.0	2.0	2.1	2.2	2.4	2.6	2.9	3.3	3.7	396.9
上声	2.9	2.7	2.6	2.3	2.1	1.9	1.7	1.4	1.1	1.0	349.3
去声	3.9	3.8	3.7	3.7	3.6	3.6	3.6	3.5	3.5	3.5	416.1
阴入	3.3	3.2	3.1	3.0	2.8	2.7	2.6	2.5	2.4	2.4	193.8
阳入	4.7	4.7	4.7	4.7	4.8	4.8	4.9	4.9	4.9	5.0	195.4

根据表 2.2 数据，取横坐标为声调时长，单位 ms；纵坐标为五度值，利用 Matlab 中 *linspace* 和 *plot* 函数作出松旺镇客家话声调五度值示意图。见图 2.2：

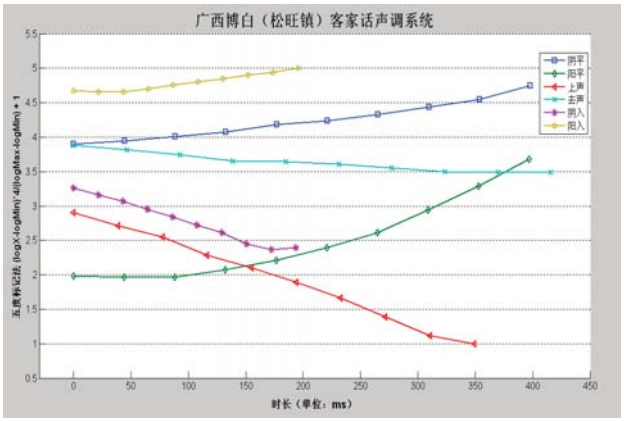


图 2.2 博白松旺镇客家话声调系统

图 2.2 中的 1-5 刻度分别表示 1-5 度，可以看出，阴平调和阳平调均为为升调，阴平调值曲线分布在 4 度-5 度之间；阳平调值曲线分布在 2 度-4 度之间；上声调为降调，调值曲线分布在 3 度-1 度之间；去

声为弱降，调值分布在 4 度-3.5 度之间，但从听感上并不明显，所以可以近似看做平调；阴入调为降调，调值曲线分布在 2.5 度-3.5 度之间；阳入调为升调，调值曲线分布在 4.5 度-5 度之间，阴入和阳入的声调时长均为 200ms。由于声调时长很短，我们在记调值时不考虑调型因素。由图 2.2 我们可以得到松旺镇客家话声调调值：阴平 45，阳平 24，上声 31，去声 33，阴入 3，阳入 5。

2.3 松旺镇客家话声调数学模型

博白松旺镇客家话声调都是升调、降调或者平调，没有曲拱特征，所以其声调曲线的解析函数可以近似为一次函数，用“T = A*X + B”公式来表示，其中 X 表示时长，T 表示声调五度值，A 和 B 为一次函数的系数。对应声调的形态来说，A 代表声调倾斜的情况，即斜率，B 代表声调初始值，A 如果为正值，声调就是一个升，为负值，声调为降。

利用 Matlab 中的数据拟合函数 *polyfit(x,y,n)*，令 X= linspace(0,时长,10)，单位为 ms，令 T=[声调的十个五度值]，n 为拟合多项式最高次次数，因为一次函数就可以满足要求，所以令 n=1，这样就可以得到一次函数“T = A*X + B”中 A、B 的值，从而得到松旺镇六个声调的五度值函数解析式：

阴平：$T_1 = 0.0020 * X_1 + 3.8344$，（ $0 < X_1 < M_1$； $273.8 < M_1 < 471.3$）；

阳平：$T_2 = 0.0043 * X_2 + 1.6551$，（ $0 < X_2 < M_2$； $299.7 < M_2 < 493.7$）；

上声：$T_3 = -0.0056 * X_3 + 2.9445$，（ $0 < X_3 < M_1$； $303.4 < M_3 < 390.5$）；

去声：$T_4 = -0.0010 * X_4 + 3.8375$，（ $0 < X_4 < M_1$； $316.2 < M_4 < 533.9$）；

阴入：$T_5 = -0.0050 * X_5 + 3.2620$，（ $0 < X_5 < M_1$； $157.4 < M_5 < 236.9$）；

阳入：$T_6 = 0.0018 * X_6 + 4.6135$，（ $0 < X_6 < M_1$； $72.7 < M_6 < 260.3$）；

其中参数 M_i 是各个声调时长的取值范围。图示如下：

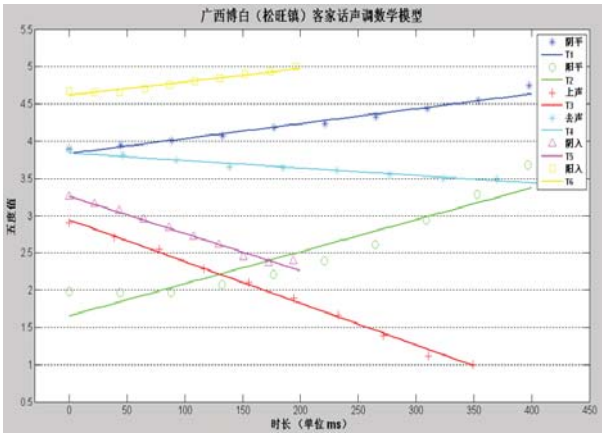


图 2.3 博白松旺镇客家话声调数学模型

图 2.3 模型（数学公式）中的曲线为函数解析式的图像，是对松旺镇客家话声调的模拟，不同形状的点代表声调归一的十个采样点的五度值。曲线所表现出的高低、曲折代表声调的调型，曲线的长短代表声调的时长。

2.4 松旺镇客家话声调空间

为了更形象地看到松旺镇客家话在声调系统中每个调类所占据的空间位置，我们以基频斜率为横坐标，基频平均值为纵坐标作出“基频斜率-基频均值”二维空间的松旺镇客家话声调散点图。

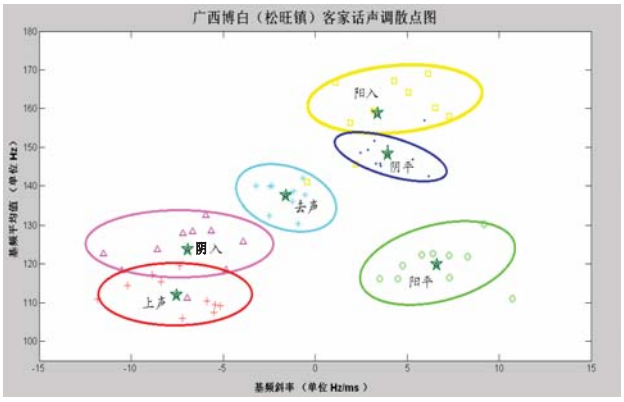


图 2.3 松旺镇客家话声调散点图

图 2.3 中共有 60 个点，分别代表 60 个声调例字。其中每个点的横坐标为该例字的基频斜率，纵坐标为该例字的基频平均值。基频斜率和基频平均值数据通过公式 2 和公式 3 得到。

$$\text{基频斜率} = \frac{\text{终点基频} - \text{起点基频}}{\text{时长}} \times 100 \quad \text{（公式 2）}$$

$$\text{基频平均值} = \frac{\sum \text{每个采样点基频}}{\text{采样点个数}} \quad \text{（公式 3）}$$

60 个声调例字的声调基频斜率和基频平均值数据见表 2.3

表 2.3 松旺镇客家话声调基频斜率和基频平均值数据

调类	声调例字	基频斜率	基频平均值	调类	声调例字	基频斜率	基频平均值	调类	声调例字	基频斜率	基频平均值
阴平	治	5.93	156.90	上声	哑	-7.23	105.88	阴入	鸭	-6.67	128.38
	方	3.94	150.27		府	-8.39	115.44		一	-5.66	128.52
	归	3.30	145.76		雨	-7.36	119.58		结	-7.20	127.96
	篾	3.22	151.70		古	-11.83	110.89		目	-5.93	132.66
	月	2.87	149.34		老	-5.89	110.39		北	-6.95	111.21
	胞	3.57	145.23		奶	-5.44	109.55		滴	-10.52	118.35
	该	6.18	142.58		宝	-10.19	114.48		节	-11.50	122.62
	绿	5.30	146.93		品	-5.52	107.53		黑	-3.92	125.82
	禁	2.47	148.58		土	-8.86	117.31		八	-4.85	118.45
	低	3.54	145.83		好	-5.17	109.27		出	-8.57	123.80
阳平	于	7.27	122.29	去声	岸	-0.54	137.72	阳入	服	166.78	1.09
	来	4.49	116.18		汇	-2.50	132.41		蜡	141.01	-0.44
	牛	3.52	116.14		雁	-1.25	136.08		立	145.57	2.20
	旋	5.81	122.34		劝	-3.21	140.06		吃	160.18	6.52
	苔	6.41	122.69		利	-0.66	142.04		食	159.44	3.18
	凿	9.19	130.26		貌	-1.32	138.04		石	167.17	4.28
	齐	10.73	111.05		内	-1.16	138.62		十	164.29	5.08
	长	7.31	116.54		外	-2.35	140.08		笛	158.07	7.27
	还	4.75	119.51		孝	-0.94	130.46		直	156.35	1.92
	霞	8.29	121.85		次	-2.46	139.87		物	169.00	6.15

图 2.3 中的六个圆圈分别代表博白松旺镇客家话的六个声调。声调散点图为我们展示了松旺镇客家话六个声调在“基频斜率”和“基频平均值”这一二维平面中的分布情况，由图可见，松旺镇客家话六个声调主要分布在二维空间的右下方的大部分空间中，而且每个声调相互独立。二维空间的左上方为空白状。这种分布表明：（1） 每个方言都有其独特的声调分布空间，一个语言（方言）声调系统的各个声调之间具有其独立的分布空间，不与其他声调混淆和交叉。方言声调的分布形状可以为方言的分区提供一定的理据。（2）二维平面中空白之处可以为方言声调的演变提供生成空间和可能，同时也可以为有更多声调的语言（方言）提供解释的理据。

2.5 声调变化轨迹的构建

语言是不断处在变化当中的，作为语言要素之一的语音也是不断处在变化之中的。声调模型的建立不仅仅只是满足语音合成的需要，还应该能够对语音的演变进行跟踪和构建，这样才能使人们对语音本质有更深入的了解和认识。我们对此进行了尝试。

2.5.1 参照值的确定

构建声调的变化，必须首先确定一个参照值。参照值就是用一些指标来量化声调并作为参照标准。描述声调的主要指标是声调的调型、调值和声调时长，参照值的确定就是在松旺镇客家话声调模型的基础上定义调型、调值和时长三个参数的值。首先是调型的确定。调型在模型中由基频斜率表示，由于松旺镇客家话的声调调型都是升调、降调和平调，调型比较单一，所以可以用一次函数的曲线斜率来代表调型，即用声调模型“T = A*X + B”中的 A 值来表示。其次是调值和时长的确定。调值是一个变化的量，它反映的是声调随时间的变化表现的高低升降的变化。可以用声调均值 C 来表示，参照值均值由公式（4）得到：

$$\text{调值：}\ C=\frac{T(0)+T(M)}{2} \qquad \text{公式（4）}$$

公式中(T0)代表声调起点的五度值，T(M)代表声调终点的五度值，M 代表时长。这样我们可以得到松旺镇客家话每个声调的参照值：A 斜率（调型）、C 调值、M 时长三个点的数组：见表 2.4。

表 2.4 松旺镇客家话声调参照数组数据

声调	参照值数组		
	调型 A（斜率）	调值 C	时长 M
阴平	0.0020	4.2324	398.0
阳平	0.0043	2.5084	396.9
上声	-0.0056	1.9637	349.3
去声	-0.0010	3.6294	416.1
阴入	-0.0050	2.7775	193.8
阳入	0.0018	4.7894	195.4

2.5.2 声调变化构建轨迹

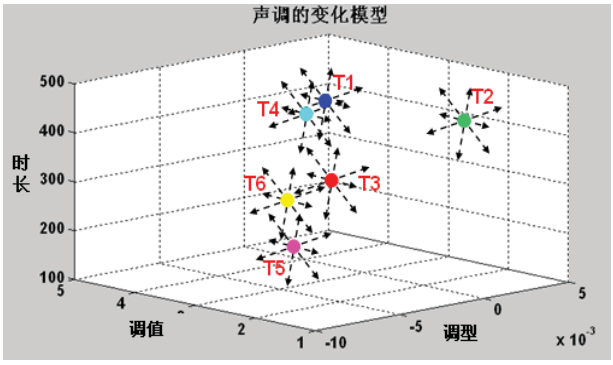


图 2.4 声调变化演示图

图 2.4 是松旺镇客家话上声变化构建的三维空间图例。图中六个点 T1—T6 是根据表 2.4 做出的松旺镇客家话的六个声调的参照点。我们可以看到，参照点周围的箭头表示声调可能变化的方向和轨迹。无论声调的调型、调值和时长发生什么变化，也无论其向哪个方向发展变化，其变化轨迹都可以反映到这一模型上，通过与参照数组的比较，我们就

能清晰地捕捉到其变化的轨迹。这将会有助于我们对声调本质更深刻的认识，也可以为声调演变提供解释的依据。

三、博白松旺镇客家话声调合成实验

语音合成，是将文字信息转化为可听的声音信息，相当于给机器装上了人工嘴巴。传统的田野调查，虽然可以录制方言的声音，记录方言的语音系统，但这种方式无论是在内容、存储、输出等方面都存在很大的限制，而且无法生成。通过计算机语音合成则可以在任何时候将方言文本转换成具有高自然度的语音，从而真正实现对方言的还原和保护。

为了检验松旺镇客家话声调模型的合理性和实用性，我们对松旺镇客家话声调模型进行合成实验测试。本文采用基频同步叠加的方法进行语音合成实验。

基频同步叠加技术(PSOLA)是合成效果较好的一种算法，其特点是能够在时域上调节语音波形的音高、音长和音强。其算法步骤主要分为三步：首先进行基频同步分析，将原始语音信号与一系列基频同步的窗函数相乘，得到有重叠的短时信号；然后对这些短时信号进行适当的时域变换，得到相应的与目标基频曲线同步的一系列合成短时信号；最后将合成的短时信号重叠相加得到合成的语音。

3.1 实验设计和语料

实验采用 Praat 软件进行语音合成，Praat 软件对基频处理的界面如下图：

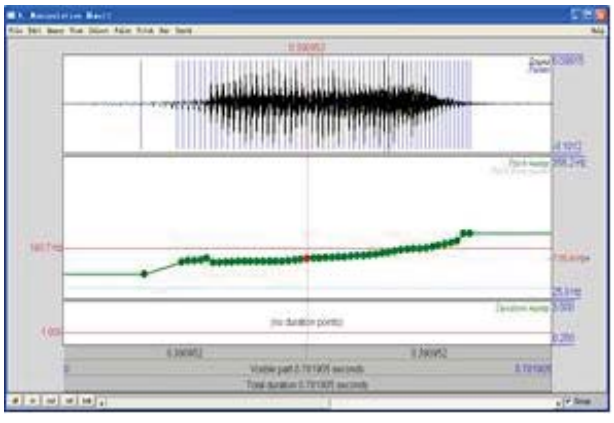


图 3.1　Praat 软件基频处理界面

图 3.1 是 Praat 软件处理基频的界面，上方是语音波形图，每条竖线是一个脉冲，两条脉冲间隔的时间表示声带一个开闭周期，这个周期的倒数就是基频值，对应下方图中的一个基频点。我们合成语音就是对下边的基频点的位置按照声调模型的要求重新布置。

我们首先从 6 个调类各选出 5 个语音样本，一共得到 30 个语音样本，见表 3.1。然后用得到的松旺镇客家话单音节声调模型的参数来改变这些语音样本的音高，之后用基音同步叠加合成（PSOLA）的方法合成出 30 个新的语音文件，最后进行人工听辨实验并和真实的语料进行对比。

表 3.1 广西博白（松旺镇）客家话单音节语音合成样本

调类	字	音节	调类	字	音节	调类	字	音节
阴平	治	/i^{45}/	上声	哑	/a^{31}/	阴入	鸭	/ap^{5}/
	方	/faŋ^{45}/		府	/fu^{31}/		一	/jit^{5}/
	月	/nie^{45}/		雨	/ji^{31}/		目	/muk^{5}/
	该	/koi^{45}/		宝	/bo^{31}/		黑	/xet ^{5}/
	低	/tei^{45}/		好	/xo^{31}/		八	/pat^{5}/
阳平	来	/loi ^{24}/	去声	岸	/am^{33}/	阳入	服	/fut^{5}/
	牛	/ŋeu^{24}/		雁	/jan^{33}/		蜡	/lap^{5}/
	旋	/sian^{24}/		利	/li^{33}/		食	/sit^{5}/
	苔	/t^{h}oi^{24}/		内	/nui^{33}/		笛	/t^{h}ek^{5}/
	还	/van^{24}/		次	/ts^{h}ɿ^{33}/		物	/vut^{5}/

3.2 实验结果

3.2.1 合成样本

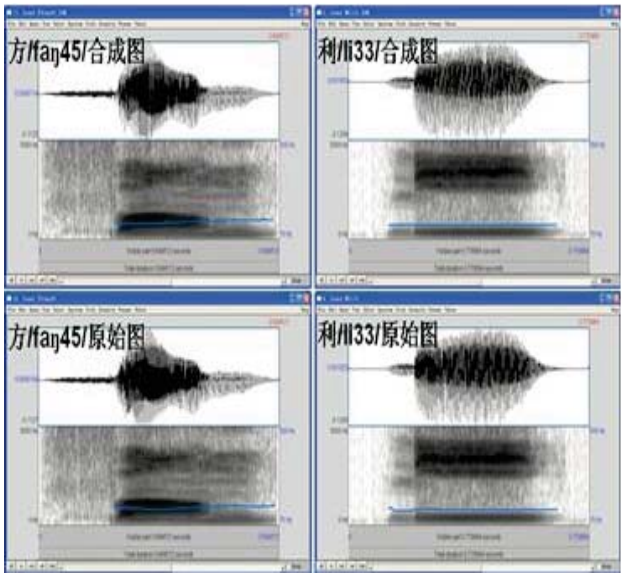


图 3.2 “方”、“利”合成语音与原始语音对比语图

图 3.2 是利用声调模型合成的松旺镇客家话“方”和“利”以及与其相应的原始语音语图。上方两幅子图分别是“方”和“利”合成语音的语图，下方两幅子图原始语音语图。从合成语图上可以看出，原始语音的基频曲线中间会有起伏波动，显得

更加自然，而合成的基频曲线比原始语图的基频曲线显得光滑平直，不够自然。为了检验其自然度和合理性，我们对合成结果进行了听辨实验。

3.2.2 测试结果分析

实验目的：对 30 个合成音节的调值的准确度和自然度进行判断。

被试：23 人，都会说客家话，听力正常。

方法：将 30 个合成音节与 30 个原始音节随即排序，每个音节之间有 2 秒的间隔，要求被试在测试

表上填写相应的结果，调值正确的在 1 处画圈，不正确的在 0 处画圈。自然度我们采用优、良、中、差四个等级，要求被试在相应的等级处画圈。然后汇总进行统计分析。

表 3.2 松旺镇客家话基频语音合成测试结果

调类	调型	调值	合成调值	合成声调自然度			
			准确率	优	良	中	差
阴平	高升	45	87%	19%	81%		
阳平	中升	24	89%	10%	90%		
上声	中降	31	87%	13%	87%		
去声	中平	33	92%	9%	91%		
阴入	中	3	82%	9%	82%	9%	
阳入	高	5	79%	7%	79%	14%	

从听辨结果看，平调（去声）的合成效果最好，准确率为 92%，自然度优秀和良好分别为 9%和 91%。其次为升调（阳平），准确率为 89%，自然度优秀和良好分别为 10%和 90%。再次为降调（上声）。短促调阴入和阳入的合成效果不如舒声调，同为升调的阴平，合成效果也稍逊色于阳平，准确率和自然度都不如阳平。

合成语音声调调值的测试结果总体表明，我们构建的声调模型是有效可行的，但由于本实验中采用 Praat 软件进行语音合成，对基频参数的控制主要采取手工调制的办法，参数控制的精确性都受到很大限制，合成出的音节自然度和基频参数控制的精度还有待提高。

参考文献：

[1] 邓玉荣. 广西贺县（莲塘）客家话音系 [J]. 方言，1994，(4).

[2] 孔江平.《论语言发声》[M]. 中央民族大学出版社，2001.

[3] 罗常培，王均.《普通语音学纲要》[M]. 商务印书馆，2004.

[4] 石锋. 论五度值记调法 [J]. 天津师大学报（社会科学版），1990.

[5] 王洪君.《汉语非线性音系学（增订版）》. 北京大学出版社，2008.

[6] 王士元，彭刚.《语言、语音与技术》[M]. 上海教育出版，2003.

[7] 吴宗济，林茂灿.《实验语音学概要》[M]. 高等教育出版社，1989.

[8] 杨锋. 标准壮语单音节、双音节基频曲线建模研究 [D]. 广西：广西大学，2008.

[9] Peter Ladefoged . Vowels and Consonants [M] . Blackwell Publishers. 2001.

[10] Peter Ladefoged. A COURSE IN PHONETICS [M]. Harcourt College Publishers. 2001.

新闻朗读的呼吸节奏与音高的关系初探

作者 张春连1，作者 孔江平2

（1. 北京大学 中国语言文学系 北京 100871；2. 北京大学 2，北京 100871）

文摘：本文在汉语音高相关研究的基础上，通过呼吸带传感器所记录的呼吸节奏变化，来分析呼吸节奏与音高（基频）之间的关系。本文按照大、中、小呼吸单元（一、二、三级呼吸）来对语料进行统计分析。一般情况下一个自然段为一个大呼吸单元，自然段里的复句对应中级呼吸单元，分句或句子成分对应小呼吸单元。我们依据呼吸重置研究了基频重置的情况，总体而言：1）呼吸重置和基频重置是规律对应的；2）在三个级别的呼吸节奏中，呼吸的重置（截距）相对比较稳定；3）一级呼吸重置中的基频重置并不大于二、三级呼吸重置的基频重置；4）基频的斜率和时长成正比；5）基频的斜率和截距成反比。

关键词：新闻朗读；呼吸节奏；基频；韵律

中图分类号：H017

1.引言

对于语调的定义大体有广义和狭义两种。广义的语调不仅涉及音高变化还包括其他超音段成分，比如音长、音强等。狭义的语调是指音高方面的变化。赵元任先生被认为对汉语语调研究做出了重要贡献。他在文章中曾指出：“英语和汉语两种语言中都有一项语言要素的分类，是嗓音基频音调的时间函数，统称为语调” [1] 并在《北平语调的研究》中提出了“代数和”主张[2]，其意思据林茂灿分析包括上升语调和下降语调作用于句末音节，上升语调使得句末音节的音阶抬高，下降语调使得句末音节的音阶降低，但调型不变。这是赵元任语调学说的核心内容之一。另外他在《国语语调》中提出了“橡皮带比喻”来说明音程（音域），“语调里最要紧的变化就是音程跟时间的放大和缩小。这种变化最好拿一个机械的比方来解释。”“高的更高，低的更低”，音域扩大，像拉橡皮带。[3] 林茂灿（2004）指出，汉语中以超音段上的音高F0（及其时长）来体现语调。汉语是声调语言，声调的声学特征也主要表现在音高上。[4] 目前学界相关研究有声调和语调的关系，汉语韵律边界的[1]声学表现，基频高低线的构建，韵律等级及边界处的声学线索，设计特定声调组合的实验室语句的方法研究汉语普通话语调降阶的规律，汉语自然对话音高等。胡明杨认为，汉语北京话的语调问题不是音高变化，即升降的问题，而是字调的起点高低问题，或者说调阈的高低问题，句末音高的高低是其音阶的抬高或降低，不是其音高变化问题。[5] 在劲松的试验中，他认为对汉语语调做贡献的是话语最后节奏单元重读音节的音高变化。

本文主要从语调的狭义定义（即音高方面的变化）出发来研究呼吸和语调的关系。由于音高是基频的表现形式，为了与已有的基频重置等术语相一致，下文主要使用基频这一术语。从谭晶晶等人的研究可以看到，呼吸和时长等超音段成分存在一定的相关性，所以，我们可以将时长等成分也考虑进来，将时长归入语调的内容，从不同方面来看呼吸和语调的关系。

根据谭晶晶等人[7－9]的研究，从语篇大尺度信息角度着眼，可以将语流分为大、中、小三级呼吸单元，分别对应于自然段、复句、分句或句子成分。研究还发现呼吸边界在音高和时长方面的一些声学表现。从音高方面说，不同呼吸边界后音节的高音点和低音点的差距都很显著，呼吸边界两端高音点和低音点的音高重置和呼吸重置幅度显著正相关，不同呼吸级别的音高重置程度有显著差异。张锦玉（2010）对语篇中的停延和呼吸特征做了研究，结果是呼吸参数在不同韵律等级下有显著差异，呼吸斜率和相应位置的停延时长也有较高的相关性。

本文希望通过对呼吸带传感器纪录的呼吸信号的节奏变化的分析，和语音基频模式的变化进行比较，看它们之间是否存在一定的相关性。

基金项目：国家自然科学基金资助项目(61073085)
作者简介：张春连（1989－），女（汉族），山东，硕士研究生。
通讯作者：孔江平，教授，E-mail：jpkong@pku.edu.cn。

2. 研究方法

本研究使用生理信号采集仪器对呼吸信号和语音信号进行采集，通过信号处理提取出语音的基频、呼吸重置的时长和截距，最终利用这些参数对汉语韵律和呼吸的关系做进一步的分析和研究。

2.1 实验语料

本研究采用的语料是从北京大学中文系语音乐律实验室自行设计录制的韵律语料库中的一部分。该语料库的语料包括传统文学体裁上的四种类型：诗歌，散文，小说，戏剧。本研究所选的新闻虽不在文学体裁范围内，但由于其写作模式较稳定，不带明显感情色彩，语音材料较易获得等特点，在韵律研究中采用较多。由于时间有限，本文所选用的新闻语料为语料库40篇新闻中的七篇，它们分别为新闻5，新闻7，新闻8，新闻12，新闻16，新闻18，新闻19。每篇约含250个汉字，一般为2－4个自然段。在句子命名时，我们按其呼吸信号来界定，第一个数字按其所在段落来定。例如，第三段的句子第一个命名数字为3，第一个命名数字后为其所在段落里句子的序号，例如第三段的第五个句子的整体编号为35，第三段的第11个句子的编号为311。由于每个新闻都不超过10段，所以此方法并不会造成混乱。此处所指的句子为“，”或“。”等类似的标点符号所隔开的句子成分。

本次研究所选语料的发音人是22岁吉林女性，录音时为北京大学电视台一名新闻播音员，受过较好的新闻播音训练，语音纯正。

2.2 实验设备

本次研究的录制在北京大学中文系隔音室进行。录音使用的主要设备是澳大利亚AD Instrument公司生产的肌电脑电仪及呼吸带传感器。录音和分析软件使用的主要是肌电脑电仪自带的Chart5。本次研究使用了3路信号进行信息采集：1通道为麦克风采集的语音信号，2通道为电子声门仪（EGG）采集的嗓音信号，3通道为MLT1132呼吸带传感器采集的呼吸信号。呼吸带传感器测量的是呼吸导致的胸腹腔的收缩和扩张变化。可以通过电压值的变化来反映呼吸气体体积的变化，进而获得呼吸节奏的变化。在二维图谱上，横轴是时间坐标，纵轴是振幅坐标。振幅的变化对应呼吸的变化，上升表示吸气，下降表示呼气。

呼吸信号处理使用的程序是北京大学中文系语音乐律平台的子程序，在Windows平台下用Matlab编程实现。基频提取方面本文使用的是由北京大学中文系语言学实验室编写的matlab程序。

2.3 实验方法

首先介绍一下呼吸信号二维图谱的参数定义。横轴是时间，纵轴是由呼吸导致的电压变化，我们可以称之为呼吸曲线。呼吸曲线的上升表示吸气，下降表示呼气。重置时长指的是从开始吸气到开始呼气之间经过的时间。重置幅度指的是一次吸气过程中呼吸信号数值的变化幅度。

处理呼吸信号时，首先对其进行归一化处理。然后对归一化处理后的呼吸信号进行平滑滤波。之后标记波峰和波谷，保存参数，对要提取呼吸参数的wave文件进行批处理，直接把文件夹下面的wave文件的标记数据读取到excel表格中。

基频提取的具体的方法是，先将 wave 文件的采样频率降至 11025Hz，导入 matlab 程序，然后标记声调段，通过自相关方法提取声调周期标记，对语音样本进行平均。这里涉及到样本的时长归一化处理，时长归一化后一个音节共取 20 个点，最后得到声调基频的平均数值。

3. 研究结果

通过信号处理得到了一个型的数据库，根据得到的数据，本文是从：1）基本统计数据；2）数据趋势；3）相关分析三个方面来讨论本研究的初步结果。

3.1 基本统计数据

我们对三个呼吸级别的时长、基频的斜率和截距的最小值、最大值、均值和标准差分别进行了统计和研究。其中，时长表示某个呼吸级别的时间长度。基频斜率表示某呼吸级别的基频倾斜度，截距表示某呼吸级别里基频的浮动范围。

表一是一级呼吸数据统计表。从表中可以看出，截距的最小值为195.89，最大值为223.32，平均值为212.32，标准差为8.84。时长的最小值为3.54，最大值为15.24，平均值为6.67，标准差为3.11。斜率的最小值为-0.06，最大值为0，平均值为-0.03，标准差为0.02。从这些数据看，截距的数据比较稳定，差别不大，而时长和斜率的数据有较大的变化。

表1　一级呼吸数据统计表

	N	Minimum	Maximum	Mean	Std. Deviation
截距	18	195.89	223.36	212.32	8.84
时长	18	3.54	15.24	6.67	3.11
斜率	18	-.06	.00	-.03	.02
Valid N (listwise)	18				

表二是二级呼吸数据统计表，从表中可以看出，截距的最小值为189.78，最大值为257.05，平均值为223.11，标准差为18.79。时长的最小值为1.27，最大值为6.35，平均值为2.57，标准差为

1.29。斜率的最小值为-0.38，最大值为-0.01，平均值为-0.14，标准差为0.10。从这些数据看，截距的数据和一级呼吸差不多，但标准差要大一些。同时时长变小和斜率变大。

表2　二级呼吸数据统计表

	N	Minimum	Maximum	Mean	Std. Deviation
截距	28	189.78	257.05	223.11	18.79
时长	28	1.27	6.35	2.57	1.29
斜率	28	-.38	-.01	-.14	.10
Valid N (listwise)	28				

表三是三级呼吸数据统计表，从表中可以看出，截距的最小值为176.19，最大值为268.97，平均值为227.82，标准差为23.50。时长的最小值为0.36，最大值为2.54，平均值为1.26，标准差为0.47。斜率的最小值为-0.72，最大值为-0.02，平均值为-0.29，标准差为0.17。从这些数据看，截距的数据和一级呼吸差不多，但标准差变得更大。同时时长变得更小、斜率变得更大。

表3　三级呼吸数据统计表

	N	Minimum	Maximum	Mean	Std. Deviation
截距	57	176.19	268.97	227.82	23.50
时长	57	.36	2.54	1.26	.47
斜率	57	-.72	-.02	-.29	.17
Valid N (listwise)	57				

我们知道，一级呼吸包含二级呼吸和三级呼吸，二级呼吸包含三级呼吸，只是斜率的计算方法不同。从结果可看出，截距相对稳定，和呼吸级别没有太大的关系，但呼吸级别越高时长越短，但斜率越大。

3.2 数据趋势

本节将不同呼吸级别的斜率、时长和截距按斜率排序，然后画成线条图，从这些图可以看出每个呼吸级别数据之间的基本关系。

从图一可看出，一级呼吸的截距随斜率的升高而下降，时长随斜率的升高而升高。斜率和截距成反比，而斜率和时长成正比。

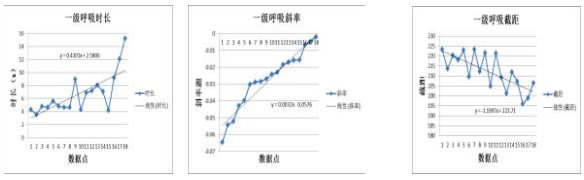


图1　一级呼吸数据图

从图二可看出，二级呼吸的截距随斜率的升高而下降，时长随斜率的升高而升高，斜率和截距成反比，而斜率和时长成正比。

图2　二级呼吸数据图

从图三可以看出，三级呼吸的截距随着斜率的升高是下降的，同时，时长随着斜率的升高是升高的，很显然，斜率和截距成反比，而斜率和时长成正比。

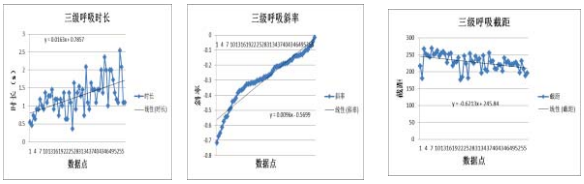


图3　三级呼吸数据图

从以上的分析可以看出，三个级别的呼吸，具有共同的性质，主要体现为：1）截距的数据相对稳定；2）斜率和时长成正比；3）斜率和截距成反比。

3.3 相关分析

从上一节可看到时长、斜率、截距在三个呼吸级别上具有一定的线性关系，但没有看到有复杂的关系，因此我们在这一节对这些数据进行定量的相关性分析。

从表四可以看出，一级呼吸的斜率和截距在0.01 级别上为负相关，数据为-0.704；斜率和时长在 0.01 级别上也为正相关，为 0.707。

表 4　一级呼吸数据相关分析表

		截距	时长	斜率
截距	Pearson Correlation	1	-.453	-.704**
	Sig. (2-tailed)		.059	.001
	N	18	18	18
时长	Pearson Correlation	-.453	1	.707**
	Sig. (2-tailed)	.059		.001
	N	18	18	18
斜率	Pearson Correlation	-.704**	.707**	1
	Sig. (2-tailed)	.001	.001	
	N	18	18	18

**. Correlation is significant at the 0.01 level (2-tailed).

从表五可以看出，二级呼吸的斜率和截距在0.01 级别上为负相关，数据为-0.752；斜率和时长在 0.01 级别上也为正相关，为 0.632。

表 5　二级呼吸数据相关分析表

		截距	时长	斜率
截距	Pearson Correlation	1	-.178	-.752**
	Sig. (2-tailed)		.364	.000
	N	28	28	28
时长	Pearson Correlation	-.178	1	.632**
	Sig. (2-tailed)	.364		.000
	N	28	28	28
斜率	Pearson Correlation	-.752**	.632**	1
	Sig. (2-tailed)	.000	.000	
	N	28	28	28

**. Correlation is significant at the 0.01 level (2-tailed).

从表六可以看出，三级呼吸的斜率和截距在0.01 级别上为负相关，数据为-0.395；斜率和时长在 0.01 级别上也为正相关，为 0.572。在 0.05 级别上，时长和截距为正相关，为 0.263。

表 6　三级呼吸数据相关分析表

		截距	时长	斜率
截距	Pearson Correlation	1	.263*	-.395**
	Sig. (2-tailed)		.048	.002
	N	57	57	57
时长	Pearson Correlation	.263*	1	.572**
	Sig. (2-tailed)	.048		.000
	N	57	57	57
斜率	Pearson Correlation	-.395**	.572**	1
	Sig. (2-tailed)	.002	.000	
	N	57	57	57

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

有一点需要说明，斜率的数值越小，基频下倾的越陡。同时肺部气流流出的也越快。

4 结论

研究结果表明，1）在三个级别的呼吸节奏中，呼吸的重置（截距）相对比较稳定，这和肺的生理机制有关；2）一级呼吸重置中的基频重置并不大于二、三级呼吸重置的基频重置；3）基频的斜率和时长成正比；4）基频的斜率和基频的截距成反比；5）分析发现少量呼吸重置和基频重置不完全对应的现象，即基频发生重置时呼吸并不一定会出现重置。

参考文献

[1]　赵元任. 英语语调（附美语变体）与汉语对英语调初探［A］. 赵元任语言学论文集［C］. 商务印书馆, 2001.

[2]　赵元任.　北平语调的研究[A]. 赵元任语言学论文集[C]. 商务印书馆, 2001.

[3]　赵元任. 国语语调[A]. 赵元任语言学论文集[C]. 商务印书馆, 2001.

[4]　林茂灿. 汉语语调与声调［J］. 语言文字应用, 2004,（3）.

[5]　胡明杨. 关于北京话的语调问题［A］. 北京话初探［M］. 北京: 商务印书馆, 1987.

[6]　王茂林，林茂灿，李爱军. 汉语自然对话音高研究［J］. 声学学报, 2008,（2）.

[7]　谭晶晶，孔江平. 新闻朗读的呼吸节奏初探. [Z]. 第七届中国语音学学术会议暨语音学前言问题国际论坛, 2006.

[8]　谭晶晶，李永宏. 汉语普通话不同文体朗读时的呼吸重置研究[J]. 第九届全国人机语音通讯学术会议论文集[C]. 2007年10月.

[9]　谭晶晶，李永宏，孔江平. 汉语普通话朗读时的呼吸节奏研究[J]. 清华大学学报(自然科学版)，2008年第S1期.

[10]　张锦玉. 普通话语篇停延率与呼吸特征初探[J].　第九届中国语音学学术会议论文集[C], 2010.

[11]　石锋. 实验音系学探索［C］. 北京大学出版社, 2009年6月.

[12]　林茂灿. 汉语语调实验研究［C］. 中国社会科学出版社, 2012年7月.

[13]　曹剑芬. 汉语声调与语调的关系［J］. 中国语文, 2002,（3）.

Preliminary Study on the Relationship between Respiratory Rhythm and Fundamental Frequency of News Reading

Author Zhang Chunlian1, Author Kong Jiangping2

1. Department of Chinese Language and Literature, Peking University, Beijing 100871, China

2. Department of Chinese Language and Literature, Peking University, Beijing 100871, China

Abstract: On the previous study of Chinese intonation, we now analyzed the relationship between respiratory rhythm and fundamental frequency by the record of respiratory sensor. The corpus is classified into three kinds of units, they are the big units, the middle units and the small units（or first respiratory, second respiratory, third respiratory）. Normally one paragraph is one piece of big unit, and the complex sentences in the paragraph are middle units, and clauses or sentence elements are small units. After comparison of the respiratory resetting and fundamental frequency resetting, we found that they are regularly corresponding to each other. On the whole: 1) the respiratory resetting and the fundamental frequency resetting is regularly corresponded; 2) in the three kinds of respiratory rhythm styles, the resetting period (intercept) is relatively stable; 3) the fundamental frequency of the resetting of the first respiratory is not as large as the second and third one’s; 4) the slope and length of the fundamental frequency is proportional to each other; 5) the slope and intercept of the fundamental frequency is inversely proportional to each other.

Key words: news reading; respiratory rhythm; fundamental frequency; cadence

Relationship between Fundamental Frequency and Speakers Physiological Parameters in Chinese-speaking young adults

Cao Honglin[1, 2], Kong Jiangping[2], and Wang Yingli[3]
[1]Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, Beijing, China
[2]Department of Chinese Language and Literature, Peking University, Beijing, China
{caohonglin|jpkong}@pku.edu.cn
[3]Center of Criminal Technology, Public Security Bureau of Guangdong Province, Guangzhou, China
wangyingli776@sina.com

Whether fundamental frequency (F0) of speech can show a reliable (negative) correlation with human speakers' body size (and shape) remains controversial, especially when the age and gender variables are controlled. No significant correlations between them were found in most previous studies. And many studies only focused on body size and the average F0 of vowels (and/or passages). However, a few studies, such as (Evans *et al*. 2006), found weight was significantly negatively correlated with mean F0 (r= -0.34, p= 0.02). The authors also found a significant negative relationship between mean F0 and certain body shape measures. Additionally, (Graddol *et al*. 1983) found that not mean F0 but median F0 and height had a significant negative correlation, but only among male speakers. Nearly all of the studies are about non-tone languages, tone language, like Chinese are investigated rarely. The purpose of this study is to examine the relationship between the mean, median, mode, standard deviation (SD) of F0 and 7 physiological parameters (body size and shape) of Chinese-speaking young male and female speakers.

70 male speakers (aged 19-38) and 74 female speakers (aged 20-34) were recruited for this study. They were students (accounting for a large proportion), teachers, physicians and public servants. All speakers were able to speak Mandarin Chinese well. The speech material was a short passage entitled "North Wind and the Sun". The speakers were required to be familiar with the material first and then read the text at a normal speed, in a comfortable way. The same equipment was used to record the material in sound-attenuated rooms at Peking University and the Second Hospital of Dalian Medical University. All recordings were made at a sampling rate of 22 kHz and 16 bit depth. The average duration of materials was 36.0s for male speakers and 34.8s for female speakers, respectively. WaveSurfer (Sjölander *et al*. 2000) was used to extract F0 values of the text using settings as follows: ESPS method, max pitch value 400Hz (for males)/500Hz (for females), min pitch value 60 Hz, frame interval 0.01s. Any occurring F0-tracking errors were corrected manually. The mean, median, mode and SD of each speaker's long-term F0 were calculated. 7 physiological parameters: height, weight, body mass index (BMI), anterior neck length, shoulder breadth, neck circumference, chest circumference were measured according to the general anthropometric methods (Xi *et al*. 2010). The range of height was 155.0-197.6cm (mean 176.3, SD 8.9) for male speakers and 144.0-180.0cm (mean 161.1, SD 8.6) for female speakers, respectively. The distribution of heights of both genders tended to be normal. Pearson correlations were calculated to estimate the relationship between all of the four types of F0 and speakers' 7 kinds of physiological variables.

The results showed that, for male speakers, no significant correlations were found between the mean, median, mode F0 and physiological parameters. However, significant negative relationships were found between SD F0 and measures of speakers' height, weight, shoulder breadth, neck circumference and chest circumference. By contrast, the findings for female speakers were more complex. Mean F0 was significantly negatively correlated with height, but not for the other 6 physiological parameters. Median F0 was found to be only significantly negatively correlated with height and shoulder breadth. No significant correlations between mode F0 and physiological parameters were found. Significant negative relationships were also found between SD F0 and height, weight and shoulder breadth. Not surprisingly, F0 showed more reliable negative correlations with physiological parameters when the data were pooled across gender classes than when analyses were performed separately for each gender.
For example, Figure 1 shows the relationship between SD F0 and speakers' height in the form of a scatter diagram. Linear regression analysis was made for both genders. It can be seen that SD F0 are negatively

correlated with speakers' height not only when both male and female speakers were involved, but also when gender was controlled. Similar results will be shown in the present study. Some possible interpretations and implications for speaker profiling of the findings will be discussed.

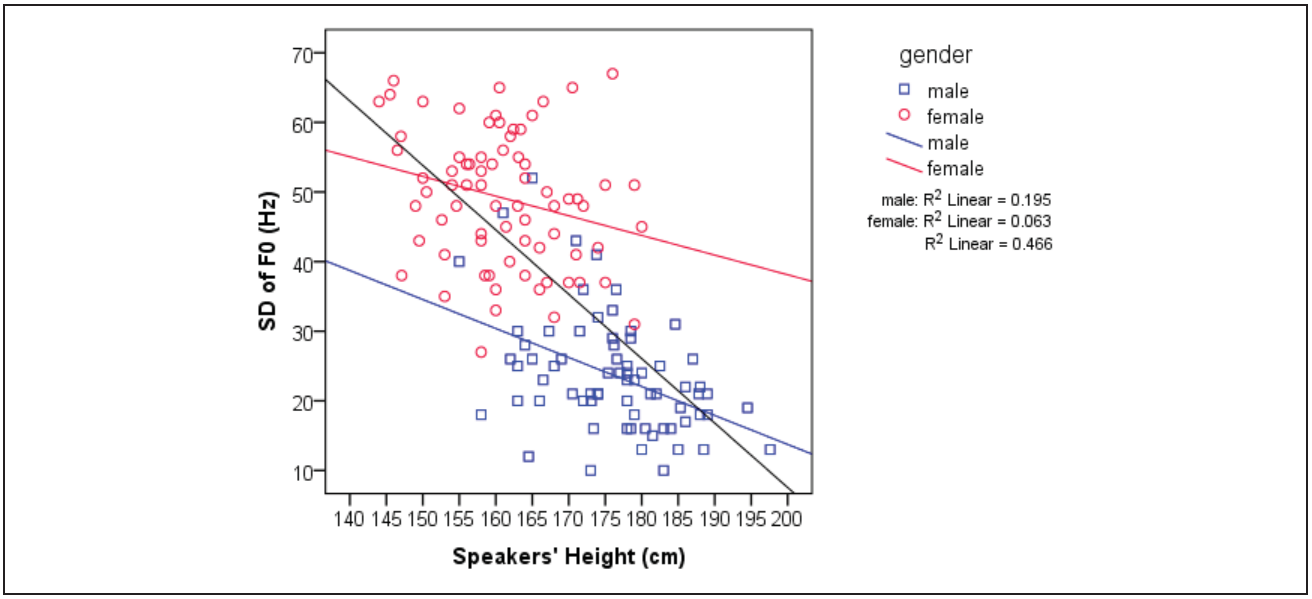


Figure 1 Scatter diagram of speakers' height and SD F0 for both genders

References

Evans, S., N. Neave and D. Wakelin (2006).

Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology* **72(2)**: 160-163.

Graddol, D. and J. Swann (1983).

Speaking fundamental frequency: Some physical and social correlates. *Language and Speech* **26(4)**: 351.

Sjölander, K. and J. Beskow (2000).

Wavesurfer - an open source speech tool. *Proceedings of ICSLP '* 2000. Beijing: 464–467.

Xi, H-J. and Chen, Z. (2010).

Anthropometric Methods (2nd edition). Beijing: Science Press. (in Chinese)

Speech Length Threshold in Forensic Speaker Comparison by Using Long-Term Cumulative Formant (LTCF) Analysis

CAO Honglin[1, 2], KONG Jiangping[2]

[1] Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China;
[2] Dept. of Chinese Language and Literature, Peking University, Beijing, China;
caohonglin@pku.edu.cn; jpkong@pku.edu.cn

ABSTRACT: Long-Term Formant distribution (LTF) is a relatively new method in forensic speaker comparison, by which the results have been proved to contain important speaker-specific information. However, few studies have been carried out for the fundamental issue that how long the speech sample should be collected. The current paper investigated the speech length threshold (SLT) by using Long-Term Cumulative Formants (LTCF) analysis, which was one of the LTF methods. The speech sample for each speaker was segmented into one-second length subsamples. Pearson's correlation coefficients were calculated for LTCF values of the whole speech sample and new set of speech samples that were formed by adding the immediately following subsample onto the speech sample before it with a start from the first subsample. The results show that SLT can be placed at about 70 seconds natural speech recordings (approximate 20 seconds only vocalic samples in duration), which are adequate to represent the whole vocal tract resonance characteristics.

Keywords: long-term formant; speech length; correlation coefficients; forensic speaker comparison

1. INTRODUCTION

Formant features are of great importance in forensic speaker identification, as they contain lots of vital speaker-specific information. Recently, Nolan and Grigoras [1], in a case study, proposed a new method of formant analysis, which was called the Long-Term Formant distribution (LTF). Instead of selecting specific vowel targets, this method captures the information from all vocalic portions, from which formant structures are visible and reliable, leading to a long-term distribution for each formant. It can summarize the whole resonance characteristics of individual vocal tract and reflect an individual's anatomy and articulatory habits [2]. Many advantages of the LTF method are found, including relatively time-efficient application, high inter-expert reliability, anatomical motivation (LTF2 and LTF3 are negatively correlated with speaker height) and language independence [3, 4]. Meanwhile, the LTF values (and the bandwidths) of F1 to F3 can also be applied to automatic speaker recognition effectively [5].

As for any long-term feature (e.g. Long-Term Average Spectrum and Long Term F0 distribution) in forensic phonetics, the speech length threshold (SLT) is an important influencing factor. In previous studies, speeches in varied lengths were used. For instance, in [6, 7], both spontaneous and reading materials were adopted. The durations of the two style recordings were, on average, 178s (range from 79s to 313s) and 39s (range from 31s to 54s) respectively. After editing for LTF analysis (i.e. only vocalic portions with clear formant structure remained), the durations decreased correspondingly, and the values were, on average, 40s (range from 12s to 83s) and 12s (range from 8 to 16s) for the two style recordings respectively. In [5], 22s and 11s length of samples (after editing) were investigated for training and test set respectively. In [2], 30s of vowel/approximant materials were selected for LTF analysis. Are these durations above long enough for LTF analysis? According to [8], it is recommended that, using the Catalina (Version 3.0h) software, speech (raw materials without editing) lengths should be longer than 10 seconds. Additionally, Moos [6, 9] compared the standard deviation (SD) of LTFs (F1 to F3) of different packages (durations were cumulative from 1s to 10s with an interval of 0.5s). The author found that there was a net duration threshold value of available speech material beyond which LTF's were saturated, which could be placed at around 5s to 8s of pure vocalic stream, depending on the formant and the speaking condition. About 6 seconds of pure vocalic stream (equivalent to 27 seconds of dialogue or 19 seconds of read speech) were, on average, enough to produce reliable LTF values [7].

Most studies about LTF analysis were based on English and German speech and focused on Long Term Average Formants (LTAF). Actually, there are two types of LTF analysis including LTAF and Long Term Cumulative Formants (LTCF) [8]. LTAF shows the distribution of each formant individually, while LTCF represents the

vertical addition of all LTAFs (cf. Fig. 1). The present study, based on Chinese reading materials, examines the correlations of LTCFs of speech with a variety of durations , in order to test the discrimination ability and SLT by using LTCF analysis.

2. METHOD

2.1. Material

Six short passages were selected in this study. They are short introductions of six famous cities in China: Beijing, Shanghai, Shenzhen, Hong Kong, Xi'an and Guangzhou (chosen from [10]). All of texts contain 2,919 syllables (In Standard Chinese one character stands for one syllable [11]).The length of the speech materials is about 10 to 20 times the length of general studies' materials, which can be shown from Table I (cf. the duration information of the references described above). The reason for this is to guarantee the hypothesis that this speech length is adequate to represent the speaker's LTCF distribution characteristics is reasonable.

2.2. Speakers and Recording

The subjects included three male native speakers of Standard Chinese, aged 30 to 34, none of whom had any noticeable voice and speech disorders (owing to the data processing process was very time-consuming, only three speakers were investigated eventually). They were required to be familiar with the materials first and then read the text at a normal speed, in a comfortable way. The SONY ECM-44B microphone was used to record the materials in a sound attenuated room at Peking University. All recordings were made at a sampling rate of 22 kHz and 16 bit depth.

2.3. Data Processing

First of all, a short C program using dynamic programming algorithm was carried out on eliminating all voiceless information. Then Wavesurfer [12] was chosen to edit and analyze the materials. The recordings were edited by hand to eliminate all nasal consonants and other vocalic portions where the formant structures were unclear. The durations of the original recordings and resulting samples of the three speakers were shown in Table I. Subsequently, formant tracking was applied to all resulting samples by using the LPC-based algorithm. The first four formant values were obtained automatically, checked and, if necessary, corrected manually. These procedures were all applied with Wavesurfer, with the settings as follows : LPC order : 12, number of formants : 4, analysis window length (hamming): 0.049s, pre-emphasis factor: 0.7, frame interval: 0.01s, down-sampling frequency: 10 kHz.

Table 1. The durations of the original recordings and resulting samples of the three speakers. “m” and “s” stand for minute and second respectively.

Speaker	CH	YQ	YF
Original duration	11m30s	12m57s	17m26s
Resulting duration	3m21s (201s)	3m42s (222s)	4m52s (292s)
Percentage	29%	29%	28%

A MatLab program was used to extract formant values, generate histogram, display the distribution of values and calculate the correlation coefficients. As discussed above, many studies focused on the mean or mode values of LTAF (often F1 and F3). However, the shape, like kurtosis and skewness, of the distribution of each formant is also very useful, because the LTAF distributions are often non-symmetrical. In the present study, we use the LTCF values for calculating, other than the mean (or mode) value of each LTAF. The differences of LTCF and LTAF are shown in Fig. 1. The LTCF represents the vertical addition of all LTAFs, i.e. the LTCF shows the cumulative frequency of occurrence of all values of F1 to F4. For one LTCF, it can be represented by an array, which is composed of the number of the frequency of occurrence (Y-axis), with a fixed X-axis range. The array of one LTCF is easily-acquired and conveniently-calculated. Meanwhile, it includes the shape information, even though it doesn’t differentiate the detailed information of each LTAF distribution.

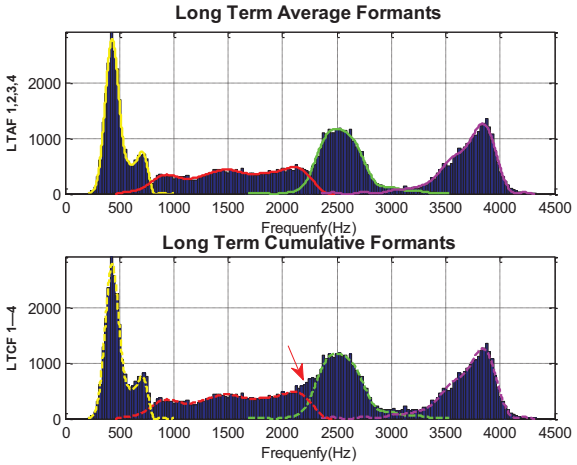


Figure 1. An illustration of LTAF and LTCF of Speaker YQ. All histograms are displayed in 25 Hz bins.

It is hypothesized that the whole resulting speech samples, whose lengths are shown in Table I, are adequate enough

to summarize the resonance characteristics of the three speakers' vocal tracts. The whole speech sample for each speaker is segmented into one-second length subsample. New set of speech samples (called cumulative samples) that are formed by adding the immediately following subsample onto the speech sample before it with a start from the first subsample are studied. For each speaker, the LTCF values of cumulative samples are calculated. For example, the length of 1s, 2s…200s and 201s cumulative samples of speaker CH are analyzed. And then, Pearson's correlation coefficients between the LTCF values of the whole speech sample and the LTCF values of the cumulative samples are calculated. To ensure the LTCF values of different samples can be calculated, the range of the whole formants frequencies is normalized to 200-4500Hz, which can cover the extreme of the F1-F4 values of the three speakers by the square.

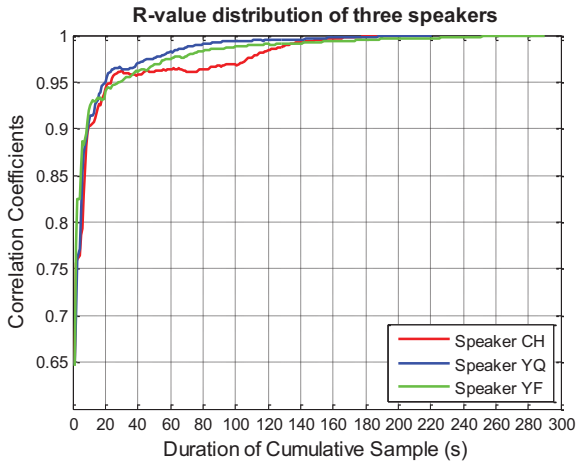


Figure 2. An illustration of Pearson's correlations (r-values, p>0.000) between the LTCF values of every cumulative sample and the LTCF values of the whole speech samples of three speakers.

3. RESULTS

Pearson's correlations (r-values) between the LTCF values of every cumulative sample and the LTCF values of the whole speech samples of the three speakers are shown in Fig. 2 in the form of line chart. The red, blue and green curves represent speaker CH, YQ and YF respectively. It shows , for example, that the r-value (for all r-values, p>0.000) of speaker CH increases from 0.65 to 0.90 rapidly, when the duration is between 1s to 9s. From 10s to about 28s of the cumulative duration, the r-value increases from 0.90 to 0.96 on medium speed. In the range of 29 to 103s, the change of r-value is relatively stable (the r-values range from 0.96 to 0.97). From 104s to the end of the material, another increase appears. To get a better view of the duration differences between the three speakers in terms of certain r-values, some special duration information is also displayed in Table II, when the r-values reach 0.80, 0.85, 0.90 and 0.95. There are at least two findings that can be derived from Fig. 2 and Table II. The first important one is that all three speakers' r-values move up at a decreasingly rapid rate, as the duration increases. The second finding concerns the individual differences of the three speakers: the distributions of the three curves are not exactly the same. For instance, when the duration is 5s, the r-value of speaker YQ is 0.80, but the value that speaker YF can reach is 0.85 instead; when the r-value reaches 0.95, no less than 18s and 24s long vocalic samples are needed for speaker YQ and YF respectively; it seems that less vocalic samples are needed for speaker YQ than speaker CH to reach a certain r-value.

Table 2. The duration information (in second) of three speakers, when r-values reach 0.80, 0.85, 0.90 and 0.95.

Speaker	r = 0.80	r = 0.85	r = 0.90	r = 0.95
CH	7	8	9	21
YQ	5	6	9	18
YF	3	5	9	24
mean	**5**	**7**	**9**	**21**

Fig. 3 shows the mean r-values distribution of three speakers. Specifically, for each speaker, r-values between the LTCF values of every cumulative sample and the LTCF values of the whole speech samples of the same speaker (intra-speaker correlations) are calculated. The red curve represents the mean r-values of the three curves which are shown in Fig. 2. Additionally, for each speaker, by contrast, the same LTCF values of cumulative samples are also compared to the other two speakers' whole speech samples' LTCFvalues (inter-speaker correlations), and then 2*3=6 groups of r-values are gotten. The blue curve represents the mean values of the 6 groups. The magenta and green curves represent the distributions of the mean values plus and minus the SD of the 6 groups respectively. Markedly, the red curve is much "higher" than the other three curves, which means that it is easy to discriminate the three different speakers using the correlation coefficients calculated by LTCF method. Meanwhile, the discrimination ability of correlation coefficients of LTCFs can be significantly improved, as the cumulative durations increase. It can be seen that, approximately, 20 seconds in duration is the boundary between the stable and non-stable change of the dis crimination ability.

To get a better view of the change rate of r-values as the duration increases, Fig. 4 is generated. In Fig. 4, the red

and blue curves represent the SD values of every three consecutive points (along the X-axis) of the red and blue lines in Fig. 3respectively. The range of the X-axis is limited to 1s to 100s so that the two curves can be more clearly illustrated. As can be observed from Fig. 3-4, the r-values are volatile when the length of the vocalic sample is les s than about 20s . The SDs of the average r-values of both intra- and inter-speaker correlations are under 0.0025 constantly when the duration is longer than about 20s.

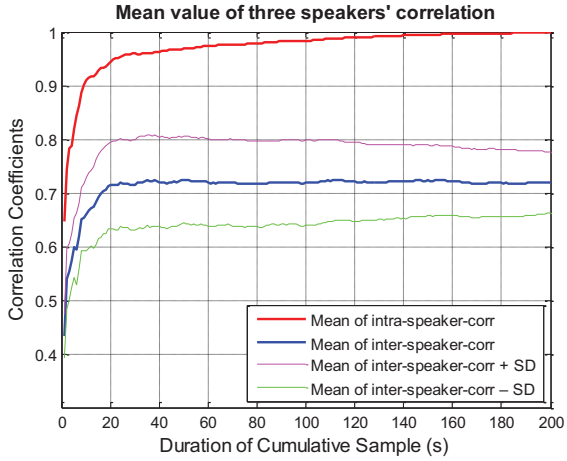


Figure 3. The mean r-values distribution of three speakers. The range of the X-axis is limited to 1to 200 seconds.

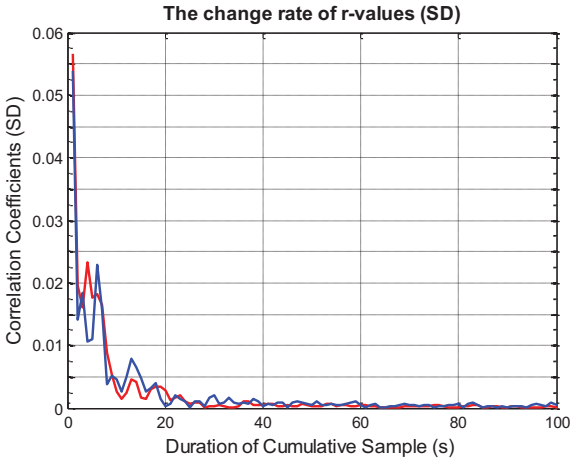


Figure 4. The change rate of r-values. The red and blue curves represent the SD of the three consecutive points of the red and blue lines in figure 3, respectively.

4. DISCUSSION AND CONCLUSION

In this paper, about 20s vocalic samples in duration are found to be able to have a stable discrimination ability of the three speakers by using LTCF analysis, i.e. the LTCF values of 20s vocalic samples seem to be able to show a good representation of the whole resonance characteristics of individual vocal tract. Table I shows the proportional relationship between the duration of resulting vocalic samples and original recordings, which is 29% on average (cf. [7]).Based on this proportion, about 70s original speech (speaking at natural speed) in length is needed, which is equivalent to 20s vocalic samples. From a forensic point of view, 70s speech in length of one speaker is often available in civil cases and also in a few criminal cases. It doesn't mean that, however, the speech length less than 70s (or 20s vocalic samples) cannot be used. The results presented in this paper show that the r-values of all three speakers can reach at 0.90 when 9s vocalic samples (about 31s natural speech) are available. In a s mall sample size (number of speakers), 9s will be adequate to discriminate speakers. But as the sample size increases, more speech materials will be needed, since the possibility that the r-value of two speakers' LTCFs is very high (e.g. exceed 0.90) cannot be ruled out. Meanwhile, the relation of r-value and duration varies between the three speakers, which supports Moos' findings [7, 9].

Many previous studies on LTF analysis focused on the mean or mode value of every LTAF (often F1 and F3) based on English or German speech materials. The approach proposed in this paper uses LTCF values based on Chinese reading materials instead. So the results of these studies are hard to be compared directly. To some extent, for different languages, it can be speculated that 20s vocalic speech should be adequate for LTF analysis (both LTAF and LTCF). Anyhow, the longer the better.

It is worth mentioning that another direction for future research will be calculating the relation of each single LTAF other than the whole LTCF, because it was found that the dis crimination abilities of LTAFs of different formants were not the same (e.g. Moos [6, 7, 9] found that LTF value of F3 was most valuable). It will be useful not only for the quantitative analysis on LTAF distribution of every formant, but also for the new understanding of SLT by using LTF analysis.

In conclusion, the present paper proposed a new method to discriminate speakers using LTCF analysis based on three male speakers' long Chinese reading materials. For each speaker, Pearson's correlation coefficients between the LTCF values of cumulative samples and the LTCF values of the whole speech sample are calculated. The results show that this method can distinguish the three speakers effectively and the more speech materials are available , the steadier the dis crimination ability is . Markedly, on average, the speech length threshold can be placed at about 70 seconds natural speech recordings (approximate 20 seconds only vocalic samples in

duration), which are adequate to represent the whole resonance characteristics of individual vocal tract and are enough to produce reliable LTCF values. The findings reported here also have significant theoretical value to forensic casework. However, owing to only three subjects and reading other than spontaneous speech were studied, this research has some limitations. In the future studies, the effects of sample size (number of speakers), different speaking styles, languages and contents of the speech materials still have to be further investigated, which are very important factors in forensic applications.

5. ACKNOWLEDGMENTS

This research is funded by the Ministry of Sciences and Technology of China. Grant No: 2009BAH41B00. Many thanks to Dr. Michael Jessen for introducing the LTF method used in BKA. Thanks to Dr. Zhu Lei for writing the C program to edit the original materials. We also thank Dr. Lee Yinghao and Zhang Wei for their comments for the early draft of the article.

6. REFERENCES

[1] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," International Journal of Speech Language and the Law, vol. 12, pp. 143-173, 2005.

[2] K. McDougall, F. Nolan, P. Harrison and C. Kirchhübel, "Characterising speakers using formant frequency information: a comparison of vowel formant measurements and Long-Term Formant analysis," in 21th Annual Conference of IAFPA, Santander, 2012.

[3] M. Jessen and T. Becker, "Long-term formant distribution as a forensic phonetics feature," Journal of the Acoustical Society of America, vol. 128, p. 2378, 2010.

[4] M. Jessen, "Forensic phonetics," Language and Linguistics Compass, vol. 2, pp. 671-711, 2008.

[5] T. Becker, M. Jessen and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models," in Interspeech,2008.

[6] A. Moos, "Forensische Sprechererkennung mit der Messmethode LTF (long term formant distribution)," MA, Universitt des Saarlandes, 2008. unpublished.

[7] A. Moos, "Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech," The Phonetician , vol. 101, pp. 7-24, 2010.

[8] C. Grigoras, Catalina Forensic Audio Toolbox (Version 3.0h) User' s Manual, 2007. unpublished.

[9] A. Moos, "Long-Term Formant Distribution (LTF) Based on German Spontaneous and Read Speech," in 17th Annual Conference of IAFPA, Lausanne, 2008, pp. 5-6.

[10] Overseas Chinese Affairs Office of the State Council, Common knowledge about Chinese geography (English edition). Beijing: Higher Education Press, 2007.

[11] H. Wang, Chinese non-linear phonology (in Chinese). Beijing: Peking University Press, 2008.

[12] http://www.speech.kth.se/wavesurfer/

Loudness and Pitch of Kunqu Opera[1]

Li Dong, Johan Sundberg and Jiangping Kong

Abstract: Equivalent sound level (Leq), sound pressure level (SPL) and fundamental frequency (F0) is analyzed in each of five Kunqu Opera roles, Young girl and Young woman, Young man, Old man and Colorful face. Their pitch ranges are similar to those of some western opera singers (alto, alto, tenor, baritone and baritone, respectively). Differences among tasks, conditions (stage speech, singing and reading lyrics), singers and roles are examined. For all singers, Leq of stage speech and singing were considerably higher than that of conversational speech. Inter-role differences of Leq among tasks and singers were larger than intra-role differences. For most roles time domain variation of SPL differed between roles both in singing and stage speech. In singing as compared to stage speech SPL distribution was more concentrated and variation of SPL with time was smaller. With regard to gender and age, male roles had higher mean Leq and lower MF0 as compared with female roles. Female singers showed a wider F0 distribution for singing than for stage speech while the opposite was true for male singers. Leq of stage speech was higher than in singing for young personages. Younger female personages showed higher Leq while older male personages had higher Leq. The roles performed with higher Leq tended to be sung at a lower MF0.

Key words: Equivalent sound level; Sound pressure level; Fundamental frequency; Kunqu Opera; Task; Condition; Singer; Role.

INTRODUCTION

Kunqu Opera is a traditional performing art in China. It has been handed down orally since the middle of the sixteenth century and is revered as the ancestor of all Chinese Operas. It is commonly praised for its elegant phrases, wonderful stories and beautiful melodies and is performed by at least ten artists, Jing, Guansheng, Jinsheng, Laosheng, Fumo, Zhengdan, Guimendan, Liudan, Fuchou, and Xiaochou, each with a special voice timbre (Wu, 2002). The roles can be divided into five groups:

1, Sheng (Young man roles) recites and sings in both modal and falsetto register. Both Guansheng, who wears an officer's hat, and Jinsheng, who wears a headband change their voice quality according to the age and identity of the personages. A Guansheng performer acts as a young king or a gifted scholar, and his voice quality has been described as "broad and bright" having "a heavy oral resonance". Jinsheng performers often act in love stories, and sing with a brighter, lyrical voice.

2, Dan (Female roles) includes Laodan (Old woman role), Zhengdan (Middle-aged woman role), Guimendan (Young woman role) and Liudan (Young girl role). To portray their different ages and identities, D panerformers sing with different voice qualities; in general, the older the personage, the greater the proportion of modal voice. Thus, Laodan performers recite and sing with loud modal voice, Liudan performers with falsetto voices, while Zhengdan and Guimendan use both these registers.

3, Jing (Colorful face roles) performers sing with their faces painted in different colors depending on the identity of the personage. The voice quality has been described as "resonant and vigorous". Often they use a series of special effects to display different characters, such as voice bursts and "intense resonance".

4, Mo (Old male roles), including Laosheng (Old man role) and Fumo (Second Old man role), recite and sing in modal register. Laosheng performers play the roles of middle-aged or elderly gentlemen. The Fumo performer introduces the story at the beginning of the performance.

5, Chou (Buffoon roles), including Xiaochou (Clown role) and Fuchou (Second clown role), recite and sing with register shifts between falsetto and modal. Fuchou pays more attention to expression than to voice. Xiaochou is a comical role performed with a loud and clear voice.

Summarizing, the voice timbres mirror the ages, characters and identities of the various personages. The voice qualities deviate dramatically from both

[1] 本文已发表于 Journal of Voice 第一卷第一期 14-19 页。

conversational speech and Western operatic tradition, which have been well described in previous research (see e.g. Fant, 2004; Kong, 2001; Sundberg, 1987). By contrast, few attempts have been made to describe in scientific terms the acoustic characteristics of Kunqu Opera roles, although these characteristics possess a general relevance from the point of view of voice science, illustrating the flexibility of the human voice and exemplifying how the voice can be used in artistic, musical and dramatic contexts. The present study investigates 1) differences among roles; 2) differences between singing, stage speech and reading lyrics; 3) intra-role differences between songs; and 4) differences between singers of the same role. The investigation focusses on two primary acoustic properties of the voice, loudness and fundamental frequency (F0), in five Kunqu Opera roles, two female (Young girl and Young woman), and three male (Colorful face, Old man and Young man).

METHOD

Four female and six male professional performers of Kunqu Opera, age range 25 to 47, volunteered as subjects, two performers in each of five roles, see Table 1. Their professional experiences varied between 7 and 27 years. The singers were told to sing just as on stage. As there are no songs that are common to all these roles, the singers were asked to perform three or four songs of their own choice that belonged to their repertoire at the time of the recording. The songs had duration of between 2 and 3 minutes and differed in emotional color. The two Young girl singers sang only three songs, because one of the songs was very long. The singers also recited a section of stage speech. In addition, all singers read, in modal voice, the lyrics of the songs chosen, duration between 2 and 3.5 minutes. The language differed from Mandarin Chinese but was identical with what they used in their roles on stage, which actually corresponds to ancient Chinese.

TABLE 1. Ages, in years, of the ten performers

Roles	*Young girl*	*Young woman*	*Colorful face*	*Old man*	*Young man*
Singer 1	45	47	27	46	45
Singer 2	41	27	25	44	27

Young girl singer 2 and Old man singer 2, who both are performers of the Northern Kunqu Opera Theater, could be recorded in an anechoic room, about 3.6x2.6x2.2 m, as they lived in Beijing, the city where the research was carried out. The other singers, who were performers at the Kunqu Opera Theater of the Jiangsu Province, had to be recorded in an ordinary room, about 4x5x3 m. Audio was picked up by a Sony Electret Condenser Microphone placed off axis at a measured distance that varied between 15 and 21 cm for the different singers. All sound level data were normalized to 30 cm. The signals were digitized on 16 bits at a sampling frequency of 20 kHz and recorded on single channel wav files into ML880 PowerLab system. Sound pressure level (SPL) calibration was carried out by recording a 1000Hz tone, the SPL of which was measured at the recording microphone by means of a TES-52 Sound Level Meter (TES Electrical Electronic Corp., Taiwan, ROC). This SPL value was announced in the recording file together with respective microphone distance.

Two programs were used for analyzing the recordings. WaveSurfer-1.8.8p3 was used to measure the fundamental frequency (F0). After converting the files into the smp format and eliminating pauses longer than 10 ms from the recordings, the Soundswell Core Signal Workstation 4.0 was used to analyze the equivalent sound level (Leq). The distribution of SPL values was determined by means of the Soundswell Histogram module. Statistic analyses were completed using SPSS 18. Given the small sample Leq and mean F0 ($N \leqslant 8$), the mean values were compared by T-test. For the larger sample of time variation of SPL ($N>360$), a Mann- Whitney U test was employed.

RESULTS

Figure 1 shows the Leq for the different singers and tasks. The within subject averages across read texts, and songs are listed in Table 2 together with the values pertaining to stage speech. With regard to the reading of the lyrics the intra-subject variation was rather small, while the variation between the different songs was larger, means 2.5 dB and 4 dB, respectively. There were clear Leq differences between the songs sung by the same singer, which does not seem surprising, since

the Leq of singing would depend on the character of the song.

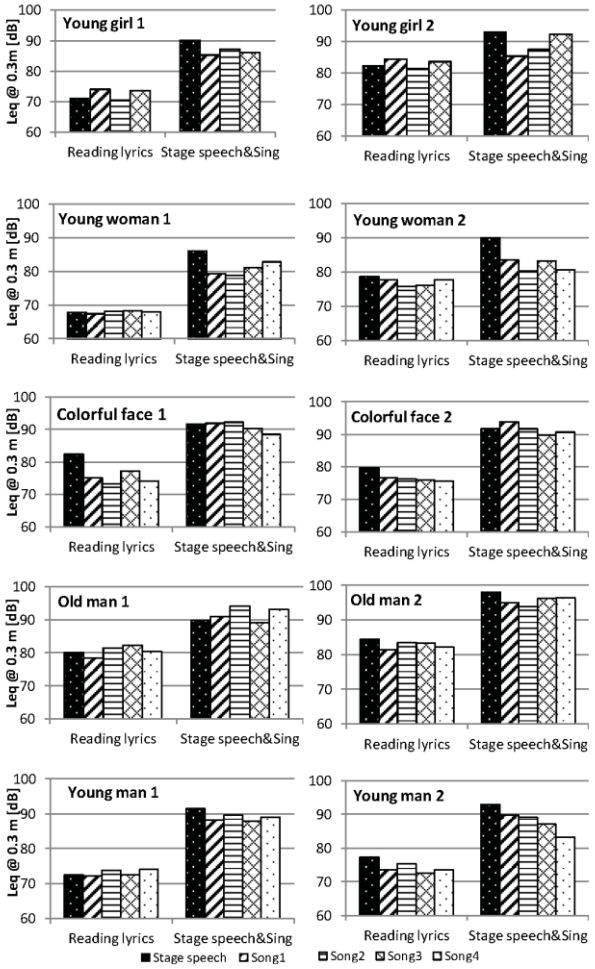


FIGURE 1. Leq @ 0.3m for reading lyrics, stage speech and singing. In each panel, the left group of columns shows the Leq values of the reading of the different lyrics, and the right group the Leq values of stage speech and three or four songs.

As can be seen in Table 2 the Leq of singing was, on average across subjects, 12.3 dB (SD 3.6 dB) higher than that of reading lyrics. Stage speech showed even higher Leq values, average 15.1dB (SD 3.6 dB). For all roles, the Leq differences between reading lyrics and singing were significant, and also between reading lyrics and stage speech ($p<0.05$). The Leq of stage speech was higher than that of songs for all singers except Colorful face 2 and Old man 1. However, only Young woman role and Young man role showed significant differences between singing and stage speech ($p<0.05$). Thus, the Leq values of singing and stage speech were similar, but both were significantly higher than that of the reading of the lyrics. Also, the variation among singers was greater in singing than in stage speech.

TABLE 2. Leq and SPL @ 0.3m. averaged across the 3 or 4 songs sung and spoken by the ten Kunqu Opera singers. The columns marked Reading lyrics refer to the singers' reading of the song lyrics.

Subject	Reading lyrics		Singing		Stage speech	
	$Mean_{Leq}$	$Mean_{SPL}$	$Mean_{Leq}$	$Mean_{SPL}$	Values	$Mean_{SPL}$
Young girl1	73.4	65.52	86.3	77.48	90.2	78.38
Young girl 2	83	73.84	89.3	79.24	93	80.47
Young woman 1	68	60.91	80.7	74.04	86.1	78.39
Young woman 2	77.1	67.8	82.7	74.74	90	81.1
Colorful face 1	75.4	66.42	90.5	83.66	91.7	80.14
Colorful face 2	76.2	68.02	92.1	86.34	91.7	83.61
Old man 1	81.2	71.08	92.6	83.8	89.9	77.52
Old man 2	82.7	72.75	95.4	87.9	98	87.27
Young man 1	73.2	64.06	88.8	81.71	91.5	81.15
Young man 2	73.5	64.85	87.9	79.53	92.9	80.36

The Leq values for the two performers of the same role varied in many cases. With regard to reading lyrics the Leq values were significantly different between the two singers of the Young girl role and of the Young woman role, and with respect to singing the two Old man role singers showed significant differences ($p<0.05$). The female roles who spoke louder in reading the lyrics also recited the stage speech and sang louder. For male roles, by contrast, the Leq of stage speech and singing had little to do with the Leq of reading lyrics.

Comparing the five roles, there were some Leq differences, see Figure 2. As a whole, the Leq of male roles were higher than those of the female roles. In both singing and stage speech Colorful face and Old man showed the highest values and Young woman the lowest, and the differences between roles were much larger in singing. The Leq of most roles differed significantly for singing ($p<0.05$). Only Young girl role and Young man role did not show significant difference in singing. With regard to the age of the characters, the younger females and older males had higher Leq.

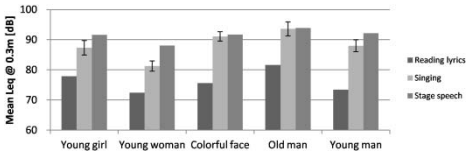


FIGURE 2. Leq @ 0.3m, averaged across subjects for the indicated conditions. The bars represent ± one standard deviation.

The mean values of SPL were about 10 dB lower than the Leq values, see Table 2. This is not surprising, given the influence of soft phonation and pauses on the SPL. The mean SPL will drop considerably if the recorded signal contains long soft or silent sections, while, under the same conditions the Leq will remain similar. The reason is that, unlike the SPL average, the Leq is calculated on the basis of linear sound pressure.

Therefore, the narrow distribution of SPL values will decrease the difference between the SPL average and the Leq. The SPL of singing had a more concentrated distribution compared with that of stage speech, see Figure 3. With respect to the roles, singers of the same role had similar distributions of SPL while singers of different roles showed differing distribution in singing but not in stage speech.

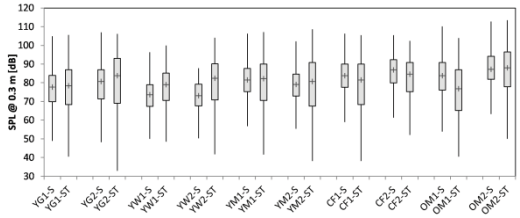


FIGURE 3. Box plot of SPL @ 0.3 m. Crosses inside boxes represent the medians. The box represents the value between the first and third quartile locations. The whiskers represent adjacent values. The horizontal axis labels are acronyms of the singers, Y young, G girl, W woman, M man, CF colorful face , O old, –S representing singing and –ST stage speech.

The SPL differences between adjacent voiced segments reflect the time domain variation of loudness. As can be seen in Figure 4, this difference was significantly larger for stage speech than for singing ($p<0.05$), indicating that the variation was greater in stage speech. For most roles, the variation of SPL with time differed significantly between roles in both singing and stage speech ($p<0.05$), see Table 3.

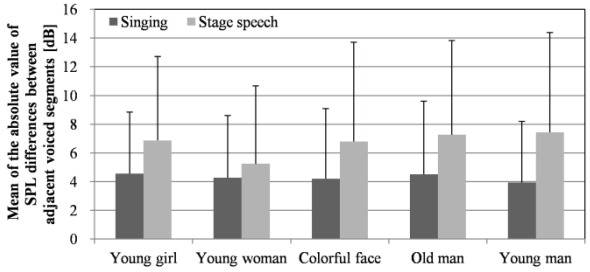


FIGURE 4. Means of the absolute values of SPL differences between adjacent voiced segments. The bars represent + one standard deviation.

TABLE 3. p-values according to a Mann-Whitney U test (significance level 0.05) of the difference in time variation of SPL between the roles in singing and in stage speech.

Role	Singing				Stage speech			
	Young woman	Colorful face	Old man	Young man	Young woman	Colorful face	Old man	Young man
Young girl	0.000	0.000	0.000	0.000	0.000	0.014	0.904^{ns}	0.927^{ns}
Young woman		0.000	0.160^{ns}	0.001		0.002	0.000	0.000
Colorful face			0.010	0.105^{ns}			0.025	0.020
Old man				0.184^{ns}				0.831^{ns}

The average fundamental frequencies MF0 of the different roles are shown in Figure 5. As expected, female roles showed higher means than male roles. For all roles, MF0 was lowest in reading and highest in stage speech. The differences were significant ($p<0.05$), except for singing and stage speech of the Old man role singers. MF0 for the different roles showed significant differences for singing ($p<0.05$) although MF0 for Young girl role and Young woman role were similar. For the male characters, the younger roles used higher MF0. The MF0 differences between singing and stage speech were much larger in the female than in the male roles.

The means and SD of F0 for each singer are listed in Figure 6. Female singers showed a wider F0 distribution for singing than for stage speech while the opposite was true for male singers. Between singers the SDF0 showed great variation for stage speech but small variation for singing, the latter reflecting mainly compositional characteristics. The singers of the same role showed similar SDF0 for the same task except for the Old man role. The female singers showed smaller SDF0 than male singers when reciting stage speech.

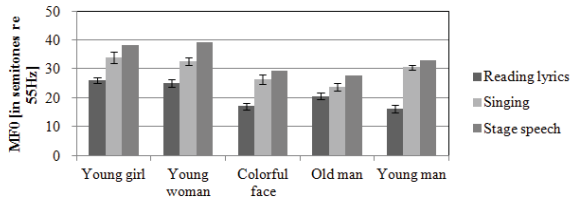


FIGURE 5. MF0 [in semitones re 55Hz], averaged across subjects for the indicated conditions. The bars represent ± one standard deviation.

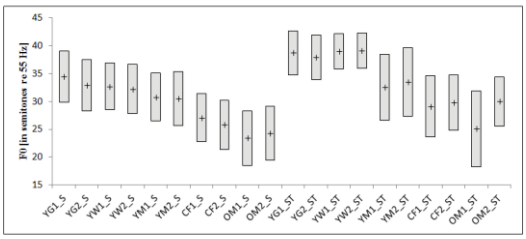


FIGURE 6. Mean and standard deviation of the distribution of F0 (in semitones re 55Hz) observed in the ten Kunqu Opera singers' singing of 3 or 4 songs and their stage speech. Crosses inside boxes represent the means. The box represents the positive and negative standard deviation. Y young, G girl, W woman, M man, CF colorful face, O old, –S representing singing and –ST stage speech.

Comparing the data shown in Figure 2 and 5, interesting relationships between mean Leq and MF0 can be observed for the different roles. Within roles there was a positive correlation, implying that the MF0 was high when the singers produced a high Leq. By contrast the roles performed with higher Leq tended to be sung at a lower MF0. It seems likely that these relationships between Leq and MF0 belong to the characteristics of the different roles.

DISCUSSION

Our analyses comprised no more than two singers for each of the five roles. On the other hand, all singers were professional and earned their livelihood from

singing, suggesting that they had well-established singing skills and well-controlled voices. Secondly, four songs with different emotions were enough for reflecting the variations of songs in the same role. The songs of Kunqu Opera could be divided into two groups, the south song and the north song. Typically, south songs are smooth while north songs are more excited. All roles include both south songs and north songs except Colorful face; at present most songs of that role are of the north song type. Thus, including several song samples for each role should have enhanced the credibility of the results. However, there was only one sample of stage speech for each singer, which might have limited the representatively of the findings. The variance of stage speech should be considered in the future.

As was shown in Figure 1, some singers' Leq was higher when they were reading the text of stage speech than when reading the lyrics of the songs, possibly because they were influenced by the speaking style of stage speech. In fact, they read the texts of stage speech more emotionally than the lyrics. When reading the lyrics of the songs, they perhaps adopted their voice habits of conversational speech.

Considering the age, dialect and the recording place, some point should be mentioned. In all roles singer number 1 was older than singer number 2, particularly for Young woman and Young man. The younger singers showed higher Leq than the older singers, especially in stage speech. Another factor is the dialect. Young girl 2 and Old man 2 both came from North China and their dialect was northern mandarin. The other singers came from South China, and their dialect was the Wu dialect, which sounds gentler than Mandarin. The style of north Kunqu Opera is bold while the style of south Kunqu Opera was gentle. Although the singers performed similar plays and used the same language when they were acting, they were probably influenced by their cultures and dialects. Also, it cannot be excluded that the different recording conditions between the north and south groups had an effect. Young girl 2 and Old man 2 were recorded in a sound treated booth with an abnormally low reverberation level, which may have caused them to increase vocal loudness. On the other hand, the Leq difference was small between the lyrics reading of Old man 2 and Old man 1, who were recorded in different rooms.

Previous research has found as strong positive correlation between Leq and MF0 in speech produced at different loudnesses (Gramming et al., 1988). The correlations differed between roles in singing. Leq and MF0 were only significantly correlated for the Old man role and Young man role ($p<0.05$, $R2>0.9$). The songs sung by each of the singers differed in emotional color, and this is likely to weaken the correlation. In speech, none of the correlations were significant (significance level is 0.05). However, the ranges were narrow both in Leq and in MF0.

In the tradition of Kunqu Opera, the Young girl and the Young woman roles are performed in falsetto register, while the Colorful face and the Old man roles use modal register. Mean Leq and MF0 were intermediate for Young man role, and this role uses modal voice in the lower pitch range and falsetto in the higher. The relationships between vocal register and Leq and MF0 in Kunqu Opera would be worthwhile to study more in detail in the future.

CONCLUSION

This study explored the differences in Leq, SPL and F0 between tasks, singers, conditions and roles. The inter-role difference was larger than intra-role difference. The singers of the same role showed a similar F0 concentration, not only for singing but also for stage speech. The variation of SPL with time differed between most roles in both singing and stage speech.

On average Leq of stage speech and singing were 15 dB and 12 dB higher than conversational speech as documented in the singers' reading of lyrics. The Leq of stage speech were higher than singing for all the singers of Young girl, Young woman and Young man roles. The between-role Leq differences were smaller in stage speech than in singing. In singing as compared to stage speech the SPL distribution was more concentrated and the time domain variation of SPL was smaller.

Mean Leq and MF0 varied systematically with the sex

and age of the singer. Male roles had higher mean Leq and lower MF0 than female roles. The F0 distribution of singing, expressed in semitones, was wider than that of stage speech for female singers and narrower for male singers. There was not much difference in F0 concentration between singers while singing. The female singers showed smaller SDF0 than male singers in stage speech. With regard to the ages of the characters, younger female personages showed higher Leq while older male personages had higher Leq. The roles performed with higher Leq tended to be sung at a lower MF0.

ACKNOWLEDGMENTS

The authors would like to thank the voice experts for their gentle participation in this investigation. This research was funded by the National Social Sciences Foundation of China and China Scholarship Council, grant numbers were 10&ZD125 and 201206010134 respectively. It was carried out during the first author's stay at the Department for Speech, Music and Hearing at KTH Stockholm, Sweden.

REFERENCES

[1] Sundberg J., Gu L., Huang Q., & Huang P. Acoustical Study of Classical Peking Opera Singing. J Voice 2012; 26(2):137-143.

[2] Wu X., Dictionary of Chinese Kunqu Opera. Nanjing: Nanjing University Press, 2002.

[3] Fant G., Speech acoustics and phonetics. Norwell, MA: Kluwer Academic Publishers, 2004.

[4] Kong J. On Language Phonation. Beijing: Central Nationalities University Press, 2001.

[5] Sundberg J., The science of the singing voice. DeKalb, Illinois: Northern Illinois University Press, 1987.

[6] Gramming P., Sundberg J., Ternström S., Leanderson R., & Perkins W. H. Relationship between changes in voice pitch and loudness. J Voice 1988; 2(2): 118-126.

Long-Term-Average Spectrum Characteristics of Kunqu Opera Singers' Speaking, Singing and Stage Speech[1]

Li Dong, Jiangping Kong, Johan Sundberg

Abstract: Long-term-average spectra (LTAS) characteristics were analyzed for ten Kunqu Opera singers, two in each of five roles. Each singer performed singing, stage speech and conversational speech. Differences between the roles and between their performances of these three conditions are examined. After compensating for Leq difference LTAS characteristics still differ between the roles but are similar for the three conditions, especially for Colorful Face (CF) and Old Man roles and especially between speaking and singing. The curves show no evidence of a singer's formant cluster peak, but the CF role demonstrates a speaker's formant peak near 3 kHz. The LTAS characteristics deviate markedly from non-singers' standard conversational speech as well as from those of western opera singing.

Key words: LTAS, Kunqu Opera, condition, role, speaker's formant, singer's formant cluster

Introduction

The voice timbres of Kunqu Opera singers are supposed to mirror the ages, characters and identities of the respective roles, which have been described elsewhere [1]. In our previous investigations, Kunqu Opera singers' stage speech, singing and conversational speech were found to differ with regard to equivalent sound level (Leq) and fundamental frequency (F0) [1]. These parameters were somewhat higher for stage speech than for singing, and both were significantly higher than for conversational speech. They also differed between roles. However, Leq and F0 differences would not be enough for describing all relevant acoustic characteristics of the specific voices of the different Kunqu Opera roles. Also spectrum differences would be important. Already Leq differences are typically accompanied by frequency dependent effects on the voice source spectrum [2-5]. Furthermore, at high F0 singers may vary the formant frequencies and the distances between them [6-8]. This affects the levels of formant peaks in the spectrum and hence also the voice timbre. Therefore, an exhaustive description of the vocal style of Kunqu Opera singing needs to analyze also spectrum characteristics.

Long-term-average spectrum (LTAS) is an effective tool for voice analysis. It represents the overall spectral characteristics of a voice and typically stabilizes after 30-40 seconds of running speech [9-14] and singing [15-18]. The LTAS contour reflects both the voice source and the vocal tract resonance characteristics. In singing as well as in speech an LTAS typically shows a peak near 0.5 kHz. The reason is that F1 is frequently located in this range. Classically trained western singers, such as bass, baritone, and tenor singers, typically display another pronounced peak in the high frequency part of an LTAS, between about 2.5 and 3.3 kHz [8, 15]. This peak has been referred to as the singer's formant cluster and has been explained as the result of clustering formants 3, 4 and 5 [15]. For professional voice users, such as actors and country singers, a prominent peak often occurs at a slightly higher frequency, near 3.5 kHz. It has been called the speaker's formant [12-14, 19]. It has been explained as the result of the closeness of F3 and F4 [14].

Both the singer's formant and the speaker's formant have been explained as the consequences of a reduction of the frequency distance between higher formants. Acoustic theory of voice production [7] predicts that the levels of two formants generally

[1] 本文已发表于 Logopedics Phoniatrics Vocology.

increase by 6 dB each if the distance between them is halved. Likewise, vowels with a high first formant, such as /a/, or a high second formant, such as /i/, have strong singer's formants, and vice versa. Formant frequencies are determined by vocal tract shape. For example, the singer's formant is highly dependent on the physiological configuration of the vocal tract, particularly the shape of the larynx tube and the area ratio between the larynx tube opening and the pharyngeal tube at the level of this opening [15].

The amplitudes and frequencies of the LTAS peaks just mentioned are influenced also by voice source. The amplitudes of the voice source partials depend mainly on the maximum flow declination rate which occurs during the closing of the vocal folds [7]. If the rate is slow, the amplitude of the partials in high frequency will be low, and vice versa. The type of closure also influences the amplitude of the partials. For example, in "breathy" phonation, in which the vocal folds fail to close the glottis completely, the amplitudes of the upper partials are decreased, which reduces the prominence of the singer's formant.

In this investigation, voice characteristics of Kunqu Opera performers of five traditional roles Young girl (YG), Young woman (YW), Young man (YM), Colorful face (CF), and Old man (OM) are analyzed in terms of LTAS. The aim was to investigate 1) whether the LTAS of Kunqu Opera singers are similar in speaking, singing and stage speech; 2) whether the Kunqu Opera singers demonstrate a singer's formant or speaker's formant LTAS peaks. Comparisons of LTAS of classically trained western singers and normal speakers and those of Kunqu Opera singers are made to illustrate the differences.

Method

Four female and six male professional performers of Kunqu Opera used in our previous study [1] were subjects also for the experiment, see Table I. The singers were told to sing 3 to 4 songs just as on stage. The total duration of the songs, which differed in emotional color, was 6-18 minutes. The singers also recited a section of stage speech, which lasted for 1-3 minutes. In addition, all singers read, in modal voice, the lyrics of the recorded songs. This reading took between 2 and 3.5 minutes. The language differed from Mandarin Chinese but was identical with what they used in their roles on stage, which actually corresponds to ancient Chinese in Ming Dynasty.

Table I. Information of ten performers

Roles	Singer	Age	Professional experiences	Gender
Young girl	1	45	25	Female
	2	41	21	Female
Young woman	1	47	27	Female
	2	27	8	Female
Colorful face	1	27	9	Male
	2	25	7	Male
Old man	1	46	27	Male
	2	44	25	Male
Young man	1	45	25	Male
	2	27	8	Male

YG singer 2 and OM singer 2, who work at the Northern Kunqu Opera Theater, were recorded in an anechoic room, about 3.6x2.6x2.2 m. The other singers, who are performers of the Kunqu Opera Theater of Jiangsu Province, had to be recorded in a quiet living room, about 3.5x5x3 m, the background noise is 35 dB(A) and the reverberation time is about 0.3 s. Although the room acoustic was quite different from a

typical Kunqu Opera stage, none of these highly experienced singers complained about difficulties to control their voices. A Sony Electret Condenser Microphone, placed off axis at a measured distance that varied between 15 and 21 cm for the different singers, was used to record the audio signals (critical distance of the room was about 75 cm). The signals were digitized on 16 bits at a sampling frequency of 20 kHz and recorded on single channel wav files into ML880 PowerLab system. Sound pressure level (SPL) calibration was carried out by recording a 1 kHz tone, the SPL of which was measured at the recording microphone by means of a TES-52 Sound Level Meter (TES Electrical Electronic Corp., Taiwan, ROC) and then announced in the recording file together with respective microphone distance. All sound level data were normalized to 30 cm.

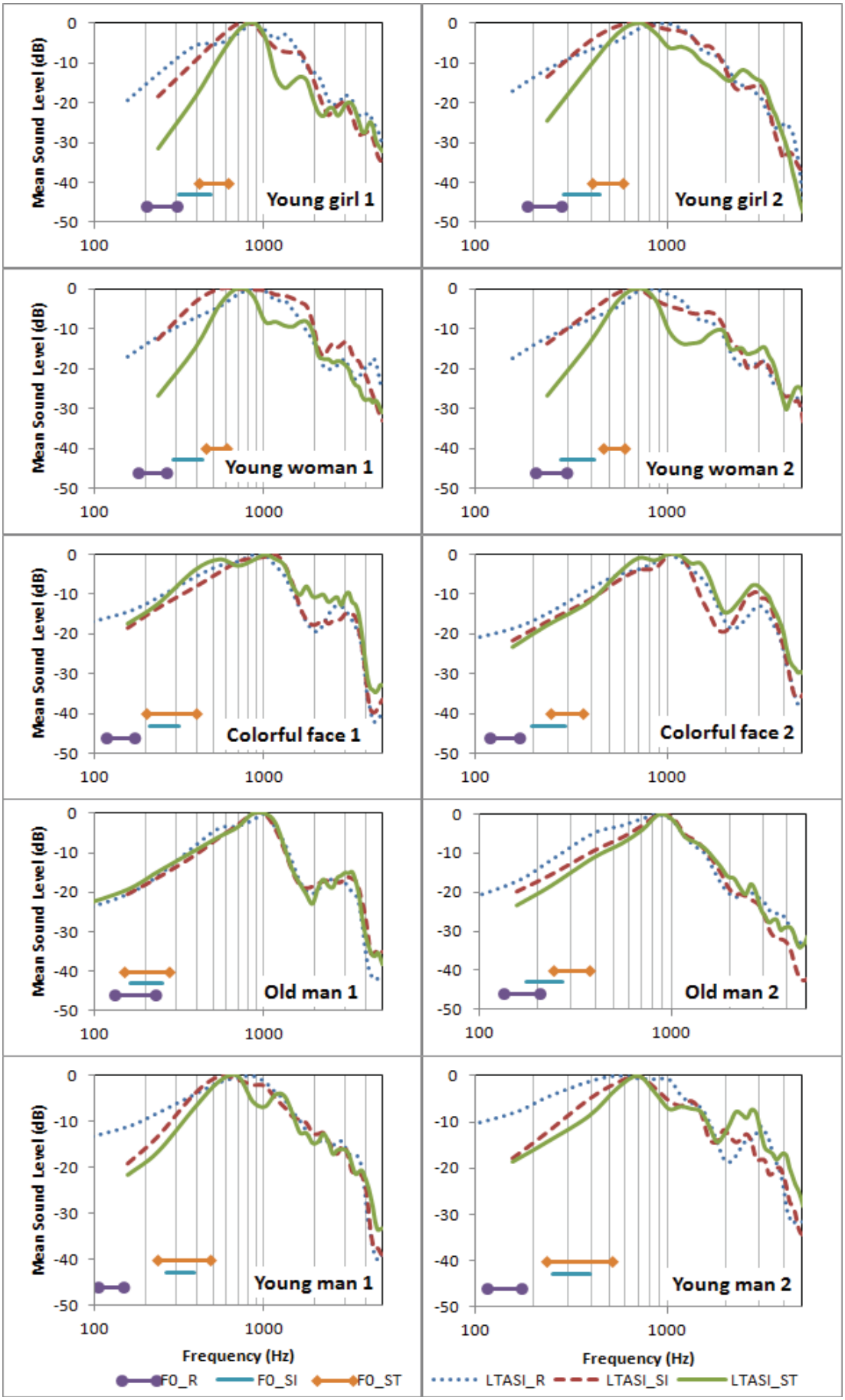
The LTAS analysis of the wav files was accomplished using the WaveSurfer software (1.8.8p3). The FFT window length was set to 128-point, the bandwidths of the analysis filters to 303 Hz and the frequency range to 0-10 kHz. After eliminating pauses longer than 10 ms from the recordings, LTAS were computed for each singer's entire recording in each condition. The recordings of singing were long and those of reading lyrics and of stage speech was rather short (1-3 minutes). Therefore, for each singer, LTAS was computed for each 40s long section of the recordings of singing so as to allow analysis of variation. Since the main sound energy appeared in the frequency range 0-5 kHz, the analysis was limited to this range. The curves for speaking and stage speech were adjusted so as compensate for Leq differences. This compensation was realized by multiplying the level values by the LTAS mean gain factors reported in previous research for different frequency bands [1, 5]. The gain factor increases with frequency in the low frequency range, keeps stable in the middle range (from 1.3 to 3 kHz) at 1.4 for male singers and at 1.6 for female singers, and decreases in the high frequency range. For instance, to compensate an Leq difference of 10 dB between two voice samples of a male singer, the LTAS level of the voice with lower Leq is increased by 10*1.0 = 10 dB in the 500 Hz frequency band, while the LTAS level in the 3000 Hz frequency band is increased by 10*1.4 = 14 dB. To obtain a quantitative measure of LTAS similarity, correlations (linear regression) were calculated between pairs of LTAS curves, using SPSS 18.

F0 was extracted using the WaveSurfer software. The extraction method was ESPS (Entropic Speech Processing System), using the algorithm of ACF (Auto Correlate Function); F0 was limited from 60 to 900 Hz; the analysis window length was 0.0075 s; and the frame interval was 0.01 s. The description statistics were accomplished using SPSS.

Results

After the LTAS had been compensated for Leq differences [1, 5], the differences between them for the three conditions were substantially diminished, especially for the CF and OM roles, see Figure 1. The LTAS curves for the three different conditions differ in a similar way for the two singers of the same role. For the female singers stage speech showed considerably less energy in the low frequency range, up to about 0.6 kHz. This would depend on their elevated F0 range. On the other hand, for the CF and OM roles, the LTAS curves of all three conditions are quite alike. For YG, YW and YM roles, the maximum peak in stage speech is located near or somewhat higher than in singing, and the peak is also narrower. The stage speech curve exhibits several peaks. Their center frequencies are close to harmonic. For example, for YG1, the center frequencies of the second, third, fourth and fifth peaks of the stage speech appear at 1.8, 2.4, 3.1 and 4.2 kHz, i.e., close to 3, 4, 5 and 7 times 600 Hz.

Figure 1. LTAS curves for lyrics reading, singing and stage speech (R, SI and ST, respectively). The horizontal lines correspond to the separation of the first and third quartiles of the F0 distribution


Mean Sound Level (dB)
0
-10
-20
-30
-40
-50
100
1000
Young girl 1
Young girl 2
Young woman 1
Young woman 2
Colorful face 1
Colorful face 2
Old man 1
Old man 2
Young man 1
Young man 2
Frequency (Hz)
F0_R
F0_SI
F0_ST
LTASI_R
LTASI_SI
LTASI_ST

Pairwise LTAS comparisons of conditions are listed in Table II by means of the linear regression. After the compensation for Leq differences, the data show higher correlations than the original data, especially between reading and singing and between reading and stage speech. This suggests that Leq variation was an important reason for the differences between three conditions. With regard to the correlations between the compensated data, all of them were significant, and for most singers, Reading lyrics and Stage speech showed the lowest similarity; the spectrum level of Reading lyrics and Singing were highly correlated (R2 > 0.9 in 8 of the 10 singers). Thus the LTAS curves of Kunqu Opera singers' singing and conversational speech show high similarity.

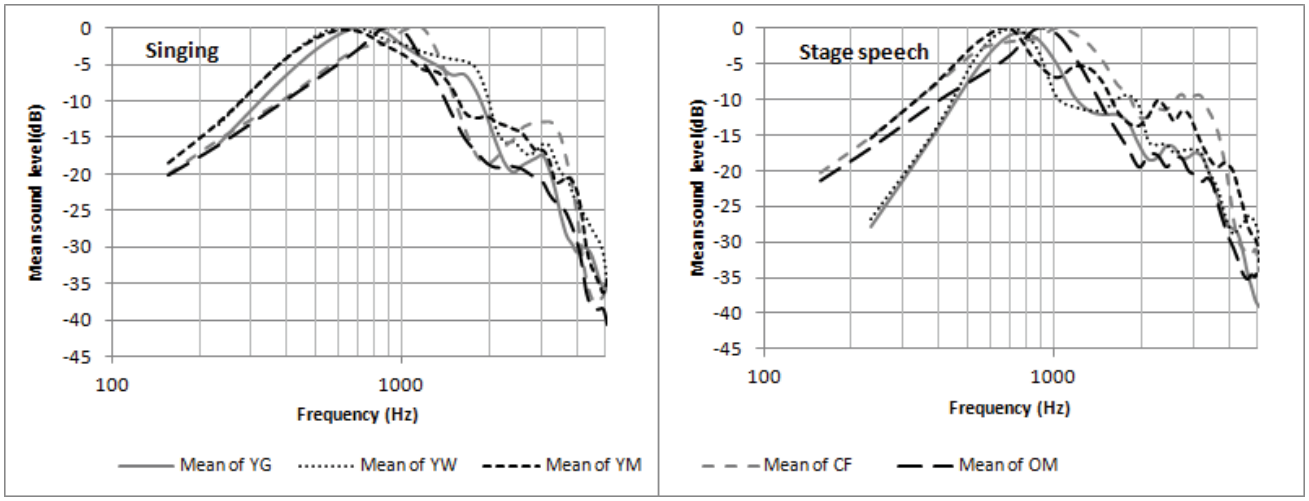
Table II. Correlations between three conditions for ten singers before and after compensation of the Leq differences (Original and Compensated, respectively) [5]. All regressions are significant

	Reading & Singing				Reading & Stage speech				Singing & Stage speech			
	Original		Compensated		Original		Compensated		Original		Compensated	
Singers	R^2	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2	Slope
Young girl 1	0.78	1.00	0.97	1.16	0.40	0.61	0.76	0.81	0.78	0.76	0.83	0.72
Young girl 2	0.91	1.05	0.93	1.04	0.73	1.01	0.86	1.03	0.84	0.99	0.88	0.96
Young woman 1	0.49	0.84	0.76	1.15	0.32	0.63	0.74	1.01	0.84	0.85	0.87	0.83
Young woman 2	0.92	0.99	0.96	1.10	0.47	0.64	0.73	0.70	0.67	0.74	0.77	0.70
Colorful face 1	0.86	1.03	0.97	0.92	0.77	0.90	0.95	0.82	0.96	0.90	0.96	0.89
Colorful face 2	0.74	1.05	0.93	0.92	0.67	0.90	0.91	0.83	0.94	0.87	0.94	0.88
Old man 1	0.93	0.87	0.97	0.77	0.96	0.92	0.97	0.83	0.97	1.02	0.98	1.07
Old man 2	0.79	1.24	0.96	1.24	0.71	0.96	0.92	0.96	0.95	0.80	0.96	0.78
Young man 1	0.81	1.03	0.98	0.94	0.71	0.81	0.93	0.72	0.96	0.82	0.96	0.77
Young man 2	0.64	0.84	0.87	0.84	0.37	0.54	0.76	0.58	0.86	0.79	0.89	0.70

The voice timbres differ between roles [1] and LTAS curves can reflect the voice timbre. Thus, it also seems relevant to examine how the LTAS differ between the roles. Although in the present study no more than two representatives of each role were analyzed, the average LTAS for a role seems worthwhile to study. It should be born in mind that our subjects were professional representatives of the respective roles and hence their voice must contain typical characteristics of that role. Furthermore, such an average LTAS will reduce the salience of individual characteristics. For example, of the two OM singers, one showed a marked peak near 3000Hz, while the other did not, so this peak is rather weak in the average LTAS. On the other hand a marked peak appeared in this frequency range in both CF singers' LTAS, so it became prominent in the average LTAS, thus suggesting that this may be a typical property of this role.

The left and right panels of Figure 2 show the average LTAS for each of the roles for singing and stage speech. All roles display a main peak between 0.7 and 1.1 kHz; for the CF and OM roles it appears at somewhat higher frequencies than for the other roles, for both singing and stage speech. The curves differ in steepness in the octave above the main peak. In singing it is more than 16 dB/oct for the CF and OM roles and much less for the three young roles, no more than 4 dB/oct for the YW role. In stage speech the spectrum slope in this octave is 8 dB/oct for the YM role, 12 dB/oct for the YG, YW and CF roles, and 17 dB/oct for OM roles. A second peak can be observed at 3 kHz. It is particularly marked for the CF role and the stage speech of YM role.

Figure 2. Mean value of the LTAS data of the two singers in the same role



To see the LTAS characteristics of Kunqu Opera singers' singing and stage speech, it is relevant to compare their LTAS with that of standard conversational speech, which has been reported in a previous study [5]. Figure 3 shows how the Kunqu Opera singers' LTAS curves deviate from this reference. For both singing and stage speech, the LTAS level around 1 kHz is higher than the reference. This applies to all roles. In the female roles' singing, the LTAS level between 1 and 2 kHz is much stronger than the reference. A marked valley occurred in the vicinity of 2 kHz for OM and CF roles. Between 1.5 and 4.5 kHz, there are between one and three peaks for most singers. The CF shows a positive deviation from the reference between 2.5 and 4.5 kHz and for the YM role a peak, particularly marked for stage speech, can be seen around 4 kHz. Less clear peaks can be observed near 3 kHz for YG role, YW role and YM role.

Figure 3. Differences between the LTAS of singing, stage speech of Kunqu Opera singers and standard conversational speech [5]. The LTAS of standard conversational speech was compensated for the Leq difference for the different singers

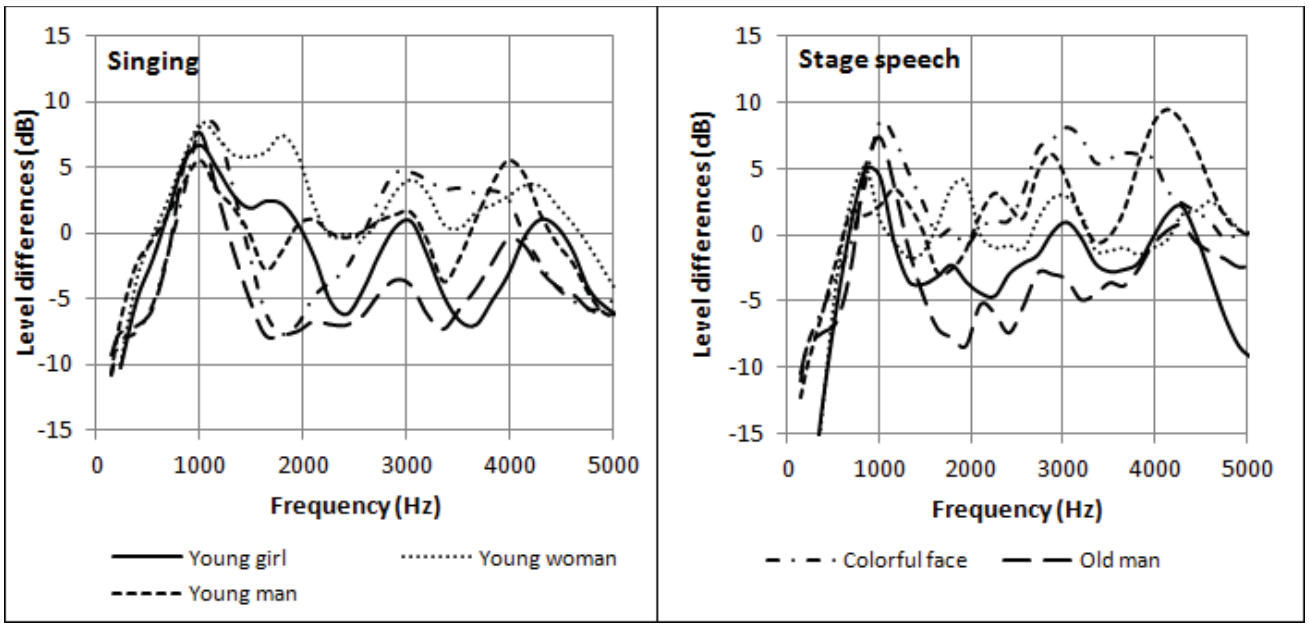
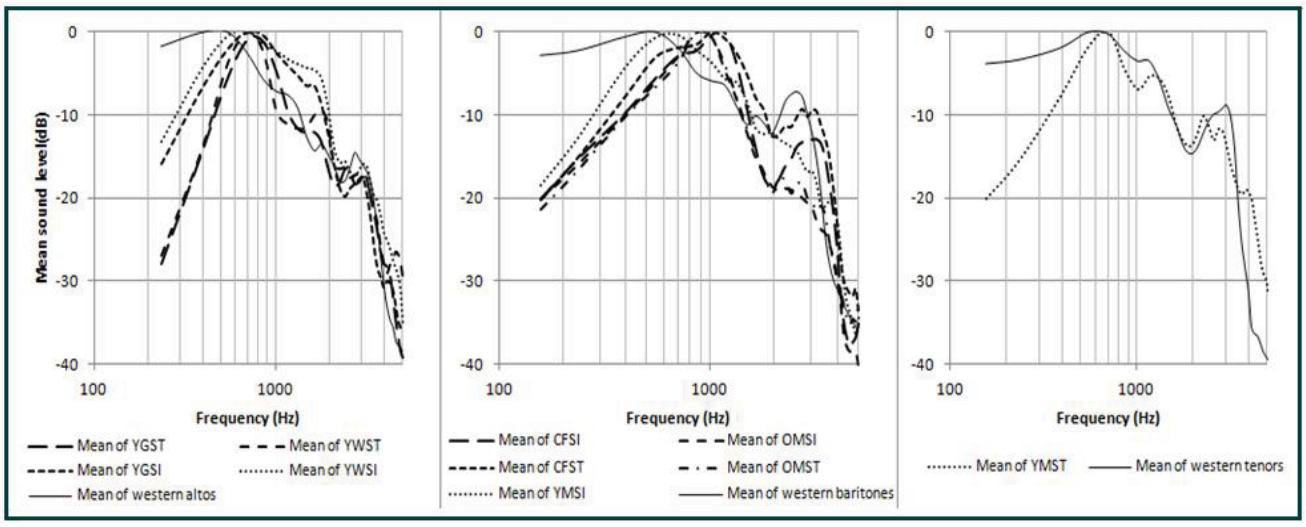


Figure 4 compares the LTAS curves of different Kunqu Opera singers with those of comparable western opera

singers [8], using similarity in pitch range as criterion: alto for YGSI, YGST, YWSI and YWST, baritone for CFSI, CFST, OMSI, OMST and YMSI, tenor for YMST. In both singing and stage speech, the main peak of Kunqu Opera singers' LTAS curves appears at higher frequency than for the western opera singers and the LTAS level below the main peak frequency is clearly lower. However, this may be because the LTAS curves of the western opera singers were derived from commercial recordings which were accompanied by an orchestra. In the female roles' singing, the LTAS level between 1 and 2 kHz is much stronger than in the case of western altos. The female Kunqu Opera singers and the western altos both display an LTAS peak near 3 kHz, which is somewhat higher in frequency and less marked in the Kunqu Opera singer voices. The LTAS curves of the CF role show a peak similar to that of western baritone singer's formant cluster, even though its center frequency is higher. Its level is comparable for stage speech but clearly weaker in singing. The LTAS curves of OM role's singing and stage speech and YM role's singing show no obvious peak in this frequency range. In YM role' s stage speech, two small peaks present between 2 and 3 kHz, while western tenor singer's formant cluster appears at higher frequency and is more marked.

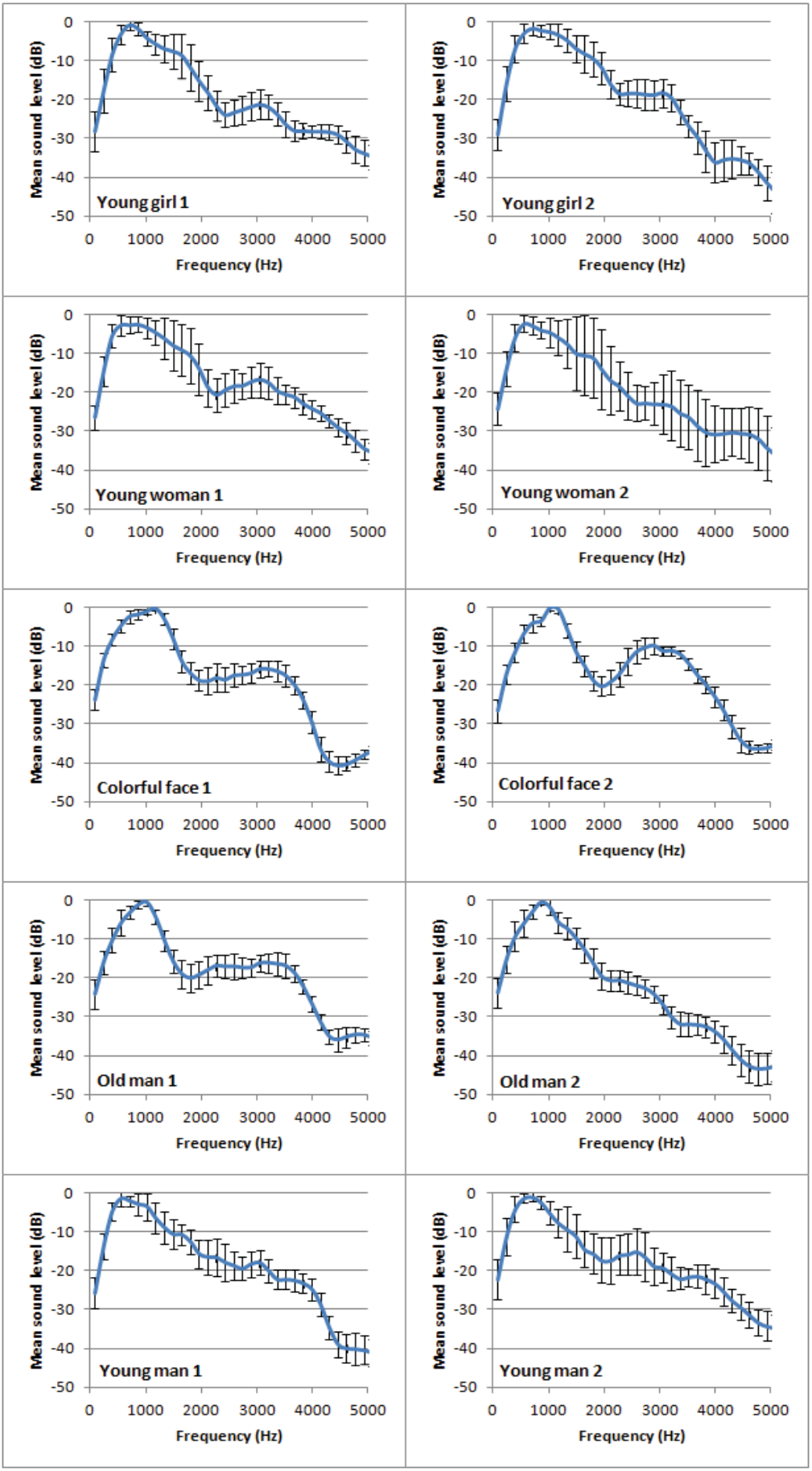
Figure 4. LTAS of singing, stage speech of YG, YW, CF, OM and YM roles and singing of western opera singers. SI: Singing, ST: stage speech



The standard deviations associated with the LTAS curves for the ten subjects' singing are shown in Figure 5. This standard deviation, henceforth SDLTAS, varies considerably between roles and singers. It is particularly wide for YW2 and particularly narrow for the OM and CF roles. For the female roles, the SDLTAS between 1 and 2.5 kHz is similar to the difference between their LTAS for singing and the LTAS for standard conversational speech, see Figure 3. This indicates that for these voices the LTAS curves vary considerably depending on what segment is chosen for analysis.

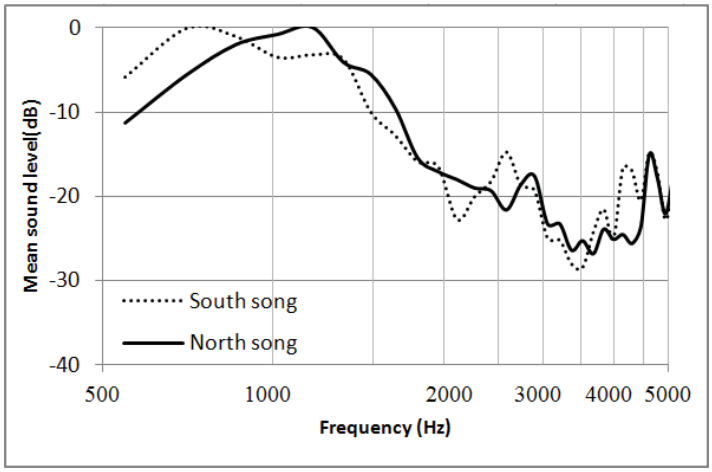
The SDLTAS in the frequency range of the singer's formant cluster is relevant for determining whether or not a voice possesses a singer's formant cluster; a low SDLTAS would imply that the spectrum level in the corresponding frequency range shows a small variation. In the case of the CF role, particularly in the case of singer 2, the SDLTAS is quite narrow in the frequency range of the singer's formant cluster. This means that these singers tended to mostly produce strong partials in this frequency region. The three young roles, especially YW 2 and YM 2, show large values of SDLTAS near 3 kHz.

Figure 5. LTAS curves and standard deviations of the different singers' singing


Young girl 1
Young girl 2
Young woman 1
Young woman 2
Colorful face 1
Colorful face 2
Old man 1
Old man 2
Young man 1
Young man 2
Mean sound level (dB)
Frequency (Hz)
0
-10
-20
-30
-40
-50
0
1000
2000
3000
4000
5000

Traditional Kunqu opera singing is performed without sound amplification and typically accompanied by a solo Kun bamboo flute. The singer's formant cluster in western operatic singing seems to have been developed in response to the sound quality of western orchestra, enhancing partials in a frequency range where the competition with the accompaniment is moderate. It is then relevant to ask if a similar relationship exists between the timbral quality in Kunqu opera singing and the Kun bamboo flute. Figure 6 shows LTAS curves, measured over several minutes of playing of the Kun bamboo flute for two types of music, "south song" and "north song". Both demonstrate three peaks below 5 kHz. The main peak appears in the low frequency range, near 700 and 1200 Hz. Both show secondary peaks between 2 and 3 kHz and between 4 and 5 kHz.

Figure 6. LTAS of Kun bamboo flute



Discussion

LTAS curves of most Kunqu Opera singers show one or more peaks in the high frequency range. Clear peaks in an LTAS curve may reflect either of three conditions or combinations of them: (i) stable formants frequencies; (ii) narrow formant bandwidths; (iii) partials in the corresponding frequency region. Since the frequencies of the higher formants are rather constant, the first condition is mostly met. Regarding the second condition, a long closed phase will make the bandwidths narrow, and with respect to the third, a high F0 implies wide separation of spectrum partials, so that the peaks at high frequencies may reflect both harmonic partials and formants. Conversely, an LTAS peak will be a sign of a stable formant when the F0 average is low or when the variation of F0 is great. Compared with CF role, YG, YW and YM roles, who all sing in a high F0 range, showed lower spectrum level at high frequencies. This may be a combined effect of formants and partials.

A tendency to cluster two formants will result in a peak at the center frequency of the cluster surrounded by valleys. The singer's formant is produced by clustering F3, F4 and F5 and the centre frequency of the peak appears between 2.5 and 3.3 kHz, depending on the voice classification. According to Bele [14], the speaker's formant is produced by lowering of F4 such that it approaches F3, and the center frequency is between 3.15 and 3.7 kHz. Both CF singers and one of the OM role singers show peaks near 3 kHz surrounded by valleys, while the other singers do not. The peak has wider bandwidth and lower level than the singer's formant in western baritones' LTAS. Thus, it is not comparable to the singer's formant cluster but similar to the speaker's formant.

Formant frequencies affect the shape of the LTAS curve, as mentioned. For Kunqu Opera singers, F2 in low vowels, e.g., /a/ and /ɑ/, produce a strong spectrum peak which tends to extend the main peak up to 2 kHz.

By contrast, the center frequency of the main LTAS peak in previously published studies of conversational speech and of western opera singing is typically located in a lower frequency range, about 500 Hz. In Kunqu Opera singer's front vowels F2 in singing is up to 2.5 kHz and close to F3. This will raise the level of the second marked LTAS peak and form a valley between the main peak and the second peak, as in the case of the CF singers, see Figure 1.

There may be several reasons for the absence of the singer's formant cluster in Kunqu Opera. (i) The presence of a singer's formant cluster reduces the differences between vowels, and text intelligibility may be particularly important in Kunqu Opera. (ii) The singer's formant cluster boosts the sound of the singer's voice so it can be heard over an accompanying orchestra. However, Kun bamboo flute, which is the most common accompaniment for Kunqu Opera, shows a peak in the frequency range of the singer's formant cluster, see Figure 6. Thus, it has an LTAS curve totally different from that of a western opera orchestra, which shows a rather low level around 3 kHz. Hence, a speaker's formant would be more effective than the singer's formant cluster to boost the singer's voice. For female roles, which show lower Leq than the male roles, the LTAS levels between 1.5 and 2 kHz are higher than that of the bamboo flute. This may help the female roles to cut through the sound of bamboo flute.

Conclusion

LTAS characteristics of Kunqu opera performers of the roles YG, YW, YM, CF, and OM were found to differ between the roles. CF role demonstrated a speaker's formant peak in their LTAS curves. In singing the LTAS curves for the performers of the three young roles showed a great variability near 3 kHz between consecutive parts of the song, as reflected in terms of large values of SDLTAS. This implies a great variation of voice timbre and/or vocal loudness. None of the roles showed a singer's formant cluster. For all roles the main LTAS peak showed wider bandwidth and appeared at a higher frequency in singing and stage speech than in non-singers' standard conversational speech. The singers' conversational speech differed considerably from their singing and stage speech, but the substantially lower Leq seemed to be an important reason for this difference. Thus, after compensating the LTAS curves for this difference, the characteristics of conversational speech, singing and stage speech became strikingly similar, particularly for the CF and OM roles. For all roles the similarity was particularly high between conversational speech and singing.

Acknowledgments

The authors would like to thank the voice experts for their gentle participation. Special thanks should go to Oh Hanna for her help with recording. This article was carried out during the first author's stay at the Department for Speech, Music and Hearing at KTH.

Declaration of Interest section

This research was funded by the National Social Sciences Foundation of China and China Scholarship Council, grant numbers were 10&ZD125 and 201206010134 respectively.

REFERENCES

[1] Dong L, Sundberg J, Kong J. Loudness and Pitch of Kunqu Opera. In press.

[2] Cleveland T, Sundberg J. Acoustic analysis of three male voices of different quality. In: Askenfelt A, Felicetti S, Jansson E & Sundberg J (eds). Proceedings of the Stockholm Music Acoustics Conference (SMAC 83):I, Stockholm: Royal Swedish Academy of Music, Publ Nr 1985. 46:1, 143-156.

[3] Bloothooft G, Plomp R. The sound level of the singer's formant in professional singing, J Acoust Soc Amer 1986; 79: 2028-2033.

[4] Hollien H. The Puzzle of the singer's formant, in Vocal fold Physiology. Contemporary Research and Clinical Issues, ed. D. M Bless & J. H Abbs, San Diego: College-Hill, 1983. 368-78.

[5] Nordenberg M, Sundberg J Effect on LTAS of vocal loudness variation TMH-QPSR, KTH, 2003; Vol. 45

[6] Sundberg J. Formant technique in a professional female singer, Acustica 1975; 32: 89-96.

[7] Fant G. Acoustic Theory of Speech Production, Haag: Mouton,1960.

[8] Sundberg J. Level and center frequency of the singer´s formant. J Voice 2001; 15: 176-186.

[9] Kitzing P. LTAS criteria pertinent to the measurement of voice quality. J Phonetics. 1986; 14: 477-482.

[10]. Löfqvist A, Mandersson B. Long-time average spectrum of speech and voice analysis. Folia Phoniatr (Basel). 1987; 39: 221-229.
[11]. Novak A, Dlouha O, Capkkova B, Vohradnik M. Voice fatigue after theater performance in actors. Folia Phoniatr (Basel) 1991; 43: 74-78.
[12] Leino T. Long-term average spectrum study on speaking voice quality in male actors. SMAC 93.1993; 206-210.
[13] Nawka T, Anders LC, Cebulla M, Zurakowski D. The speaker's formant in male voices, J Voice. 1997; 11: 422-428.
[14] Bele I. The Speaker's Formant, J Voice 2006; 20: 555-578
[15] Sundberg J. Articulatory interpretation of the "singing formant." J Acoust Soc Am. 1974; 55: 838-844.
[16]. Cleveland T. Acoustic properties of voice timbre types and their influence on voice classification. J Acoust Soc Am. 1977; 61: 1622-1629.
[17]. Dmitriev L, Kiselev A. Relationship between the formant structure of different types of singing voices and the dimension of supraglottal cavities. Folia Phoniatr (Basel). 1979; 31: 238-241.
[18] Sundberg J, Gu L, Huang Q, Huang P. Acoustical Study of Classical Peking Opera Singing. J Voice 2012; 26(2):137-143.
[19] Cleveland T, Sundberg J, and Stone R. E. Long-Term-Average Spectrum Characteristics of Country Singers During Speaking and Singing J Voice 2001; 15: 54–60

VAT Measurement of Amdo Tibetan Plosives

Sangta , Phonetics Lab of Peking University, Bejing, China, 100871

Abstract: By measuring the VAT of Amdo Tibetan plosives, there are “hard”, “comfortable”, and “soft” onset, corresponding to voiced, unaspirated voiceless and aspirated voiceless with voiced plosives respectively. “Hard” voice has negative value with a large number. “Comfortable” voice has VAT value distributed around zero; “Soft” voice has positive VAT value. The voiced plosives are the ‘hardest’ since there is a stronger airflow need to be built up before producing a voiced plosive. Parts of the VAT of voiced plosives are “soft”, of which the VAT is positive. The initiation of SP (sound pressure) and EGG (electroglottograph) of the unaspirated voiceless plosives are almost simultaneous(VAT is around zero both positive and negative), while the SP precedes EGG when producing aspirated voiceless since the stronger airflow of aspiration oscillated the vocal fords before the adduction of the vocal fords. There is no correlation with VAT to the place of articulation of the plosives.

Keywords: Voice Attack Time (VAT), Tibetan Amdo plosives, phonation

Introductions

Vocal Attack Time (VAT) is the time lag between the growth of the sound pressure signal and the development of physical contact of vocal folds at vocal initiation (Baken et al, 1998). Amdo Tibetan plosives contrast in voicing. Voiceless plosives contrast in aspiration. The voiced plosives could be pre-nasalized. The purpose of this measurement is to find out how those different articulation correlates to the VAT values. If it’s positive, the SP (sound pressure) precedes physical contact of vocal folds, measured by EGG (electroglottograph) signal, and it’s ‘soft’ onset; if it’s negative, physical contact of vocal folds is made before SP signal, it’s then “hard” onset. Simultaneously, the relationship between the place of articulation and the VAT is observed as well.

Methods

1. Stimuli

Five groups of stimulus syllables were constructed for the plosive initials: simple voiced plosives, pre-nasalized plosives, simple aspirated plosives, unaspirated voiceless plosives and complex voiceless plosives. Each group contains at least 40 stimulus syllables. Within each group, only the rhymes of the syllables are different. Detailed is shown in the table 1.

phonation types		place of articulation(IPA)			samples of the stimulus syllables
		bilabial	alveolar	velar	
voiceless	unaspirated (simplex)	/p/	/t/	/k/	པད། དུ། ཀ
	unaspirated (complex)	/^{h}p/	/^{h}t/	/^{h}k/	ཕོར། ཐུག ཁད།
	aspirated	/p^{h}/	/t^{h}/	/k^{h}/	ཁ། ཚོས། ཁོང་།
voiced	simplex	/b/	/d/	/g/	ཐོན་དགའ
	nasalized	/^{n}b/	/^{n}d/	/^{n}g/	འབའ་མདའ མགོ

Table 1: The categorization of the stimuli

2.Subjects

The data is obtained from a 29 years old healthy Amdo Tibetan Speaker who hadn’t had any history of previous or current voice, speech, language, or hearing problems.

3.Instrumentation

The SP was recorded using SONY ECM-44B microphone at a sampling rate of 44.1kHz and a resolution of 16 bits. The EGG data was collected by using Real-Time EGG Analysis (Model 5138) produced by Kay PENTAX.

4.Procedure

Both SP and EGG data were collected in the recording room of the linguistic lab at Peking University. VAT value was computed from the time lag of the cross-correlation function using a fully automated process accompanied by operator validation (Robert F. Orlikoff, et al, 2007). Observing the VAT validation by looking at the FOM(Figure of Merit) and disregarding the VAT value when the corresponding FOM is under 0.75. Observing the VAT value from the perspective of the linguist distinction thus categorizing the plosives based on the VAT.

Result and Discussions

The summary VAT date, FOM and F0 are shown in the table 2.

		Voiceless			voiced	
		aspirated	Unaspirated (simplex)	Unaspirated (Complex)	simplex	nasalized
Number of tokens		46	63	83	63	59
VAT	Mean(SD)	7.92(6.52)	-0.75(2.11)	0.5(2.2)	-82.05(79.55)	-66.11(78.52)
	interval	-1.68 to 68.7 30.11	-6.49 to 2.54	-3.83 to 5.28	-174.79 to 17.53	-179.96 to 12.36
	Median	5.4	-0.7	0.62	-127.075	3.65
FOM	Mean(SD)	0.99(0.02)	0.99(0.01)	0.99(0.02)	0.83(0.09)	0.84(0.08)
	median	0.99	0.97	0.98	0.84	0.83
F0	Means(SD)	165(8.04)	170(16)	175(14)	112(13)	122(15)
	Median	167	172	178	107.81	117.99

Table 2: VAT values (mean VAT with its standard deviation, intervals and the median), FOM(mean FOM with its standard deviation and the median) , and F0(mean F0 and the median) of the plosives

1. VAT of unaspirated plosives.

The FOM of all of the 146 tokens is above 0.96 except one is 0.791, whose corresponding VAT value is disregarded.

Unaspirated plosives initials can be divided into simplex and complex as mentioned above. From table 2, we can see that the VAT intervals of simplex and complex are overlapped; the difference of the mean and median of VAT is slight, so there is no significant difference between them. This is clarified from figure 1 that you can't separate the tokens by VAT value.

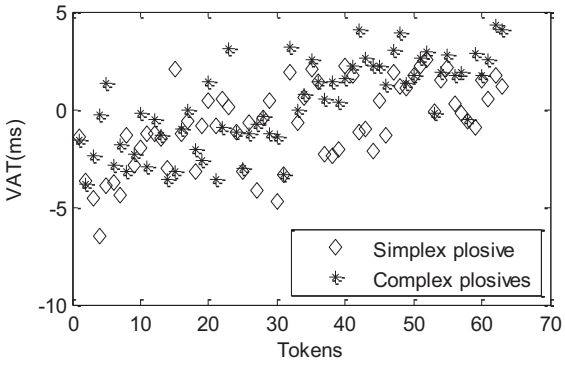


Figure 1: VAT of simple plosives and prefixed plosives

As a result, we observe the VAT of both groups together. Thus we can say that the VAT of unaspirated voiceless plosives is ranging from -6.49 milliseconds (ms) to 5.28ms, and its median is -0.08ms. This

means that the SP and EGG almost started at the same time for unaspirated plosives. We refer this type voice onset 'comfortable' voice.

2. VAT of aspirated plosives.

The VAT value of the aspirated plosives ranges from -1.7ms to 68.7ms (The confident interval is from -1.68ms to 30.11ms), among which 87% is distributed under 10ms. This is mostly overlapped with 'comfortable' voice (proposed by R.J.Baken et al,2007), which is ranging from-1.4 to 9.6ms. However, 24% of the tokens are overlapped with of 'breathy' voice (proposed by R.J.Baken et al,2007), which is ranging from 7.6 to 38.0ms. Anyway, we refer this type of onset as 'soft' voice. When comparing with the unaspirated plosives, the VAT of the aspirated plosives is much longer as figure 2 shows.

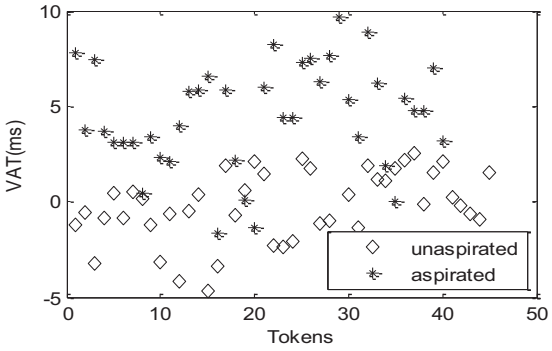


Figure 2: VAT of unaspirated and aspirated plosives

So the aspirated plosives have a more "soft" onset than the unaspirated plosives. The initiation of SP and EGG of the unaspirated voiceless plosives are almost simultaneous, while the SP precedes EGG when producing aspirated voiceless since the stronger airflow of aspiration oscillated the vocal fords before the full adduction of the vocal fords.

3. VAT of the Voiced plosives.

As mentioned above the voiced plosives can be categorized into simplex and nasalized. As you can see from table 1, there is not significance difference on the VAT values whether if it's nasalized.

The VAT of the voiced plosive is very complicated. The FOM is the poorest when comparing with it of the voiceless. There are 122 tokens totally. 20% of the tokens are under the 0.75 in FOM, so only those above 0.75 of FOM are used. The median of FOM is smaller than those of voiceless. The standard derivation is higher than those of voiceless as well. All those numbers indicate that the VAT value is not as validate as those of voiceless.

Regard the VAT value, it range from -179.96 to 17.53ms. The standard derivation is even close to the mean of the VAT, which means that discrete degree is very high as shown in the figure 3.

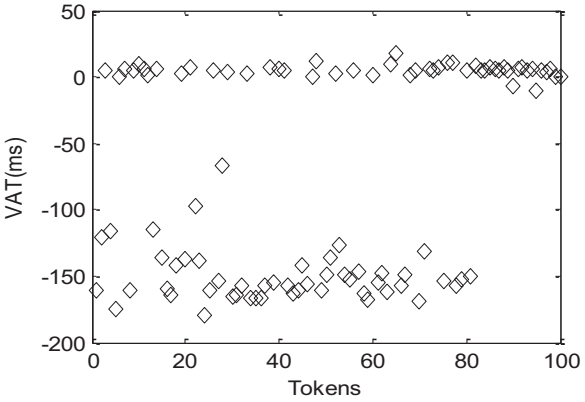


Figure 3: VAT of voice plosives

However, you can see from Figure 3 that the VAT of voiced plosives is distributed in two areas: One is around -150ms and the other is around zero. The latter happen to be overlapped with VAT of aspirated voiceless plosives as figure 4 shows. The reason could be that the voiced is starting to become devoice but it's not for sure at this stage.

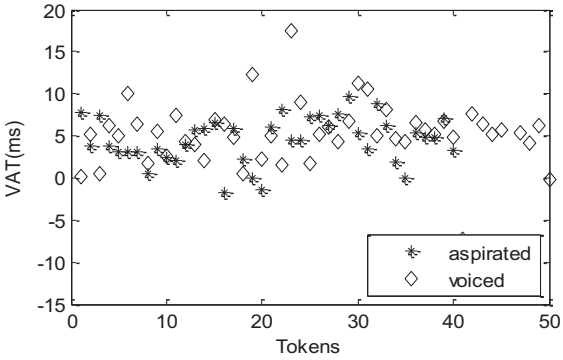


Figure 4: VAT of aspirated and Voiced plosives

So the onset of voiced plosives is separated in two extremes in the system: extremely "hard", meaning EGG Signal precedes SP Signal, and extremely "soft", meaning SP signal precede EGG signal. The exact reason of this separation is unknown, but we can sure there is complication of the phonation type when producing voiced plosives.

Regarding the negative value of VAT, the VAT value is very large, which means that EGG precedes SP, and the adduction of EGG last for a longer time. A stronger airflow is needed before producing a voiced plosive. Ron Baken pointed out that the negative VAT can all be canceled to zero, which is just qualitative. However, we can still separate those negative VATs based on the different types of plosives. Though unaspirated voiceless plosives also have negative VAT which are distributed near zero, it's still separated from the negative VAT of voiced distributed around -150. It is more than enough for linguistic purpose. Their actual value is not crucial here as we can separate them.

Conclusion

Based on the previous discussion, we can conclude that Amdo Tibetan plosives have different onset types, namely soft, comfortable and hard. Soft voice corresponds to aspirated and voiced plosives. Comfortable voice corresponds to unaspirated plosives while hard voice corresponds to voiced plosives. Soft voice has positive VAT. Comfortable voice has the VAT distributed around 0(both negative and positive). Hard voice has negative values around -150ms. The VAT pattern of Amdo Tibetan plosives is shown in figure 5.

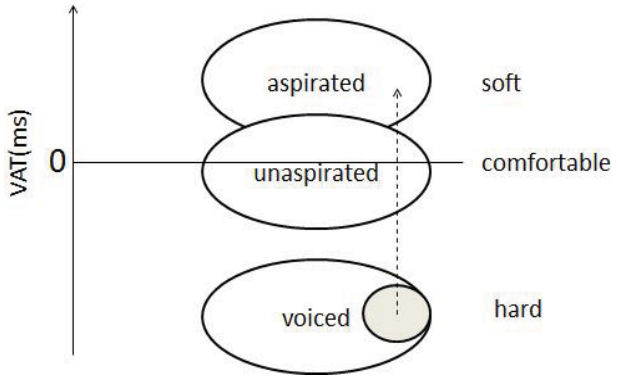


Figure 5: VAT pattern of Amdo Tibetan Plosives

The prefixal consonants of both the voiced plosives and voiceless plosives don't have any influence on the VAT value and there is no correlation between the VAT value and the places of articulation either. This experiment has only looked into the measurement of male speaker. The VAT value should be different for females, but we are looking at it from linguistic perspectives, so the pattern should be the same for females as well.

Reference

1. Roark RM, Watson BC, Baken RJ,Brown DJ, Thomas JM. Measures of Vocal Attack time for healthy young adult. J Voce. Doi:10.1016/j.jvoice.2010.09.009
2. Roark RM, Watson BC, Baken RJ. A figure of merit for vocal attack time measurement. .Doi:10.1016/j.jvoice.2010.10.012
3. Orlikoff RF, Deliyski DD, Baken RJ, Watson BC. Validation of a glottographic measure of vocal attack. J Voice. 2009;23:164-168
4. Estella P. ma, R.J. Baken, Rick M. Roark. And * p.-M. LI. Effect of Tones on Vocal Attack Times in Cantonese speakers. J. Voice. Doi: 10.1016/j.jvoic.2011.07.015
5. Hua, Kan (華侃). 2002. Zangyu Anduo Fangyan Cihui [A Vocabulary of AmdoTibetan]. Lanzhou: Gansu Minzu Press.
6. 孔江平，《论语言发声》，北京：中央民族大学出版社，2001.11

音位負擔量計量研究——以藏緬語為例

孔江平

北京大學中文系，中國語言學研究中心，

1. 引言

隨着人們利用基因研究人類演化的進展，已有比較充分的證據説明人類起源於非洲。如果這個結論正確，目前世界上説不同語言的民族就具有同一個祖先。從這個角度看，人類語言的演化經歷了相同的時間和同樣的演化路線，因此可以提出一個問題，語言的本質和演化的本質是甚麼？

從古人類學的研究成果看，相對於人類生理的進化，語言形成和演化的時間要短得多。雖然從目前古人類學的研究成果看還很難確定語言產生的時間，但現代言語科學技術的進步，利用古人類化石經過聲道復原，最終合成出語音已經成為可能。因而，怎樣從語言學的角度利用現代活的語言來研究語言進化和演變的基本性質則是語言學家面對的重要課題。

鄭錦全先生的「詞涯八千」（鄭錦全 1999，2006）很巧妙地證明了人類掌握一種語言的基本能力。如果排除掉5000年文字的影響，現在比較封閉的社會中，語音的音位通常有幾十個，基本語素在800到1,000左右，常用詞彙在3,000左右，基本句法結構在200左右。然而人類的發音器官能發出1,000至2,000個能用於語言音位的音素，但人類每種語言只選了幾十個能區別語義的音位，這説明了人類大腦目前處理語言音位的能力和水準。因此怎樣從音位數量、音位結構、音位負擔量研究人類大腦中語言的本質是本文的出發點。

在語音功能負擔量的研究方面，可以追溯到早期的布拉格學派時期，當時主要注重於音位學的二元對立。50年代的研究主要有霍凱特(Hockett 1951，1955)的研究和格林柏格(Greenberg 1959)的研究。霍凱特認為，功能負擔的重要性在於它對描寫音韻系統有重要的價值，從而使人們可以有一個尺度來認識語言資訊、語言冗餘度和言語識別。格林柏格認為，功能負擔以通用的方式反映了一組音位或一組對立特徵各成員之間的對有區別意義信號的貢獻。在60年代，主要有赫厄希斯瓦爾德（Hoenigswald 1960）關於功能負擔和音變的研究，他認為功能負擔和語言的音變有關，並提出了一個假說，即是在一種語言裏，如果一種對立用的很少，它的消失對系統造成的危害要小於功能負擔大的對立。京•羅伯特（King 1965）將音變和功能負擔一同進行研究，並着重研究了音位功能和語音音變的關係，但他發現在日爾曼語中功能負擔和歷史音變的關係不大。

在60年代，王士元教授有兩篇重要的論文（王士元 1960，1967）。第一篇針對美國英語輔音出現頻率的統計差異進行了研究，研究結果表明美國英語輔音的頻率受文獻風格、方言差異和樣本數量的影響不大。其差異主要是來自於不同的統計詞表和版本。第二篇是關於「音位功能信息量」研究的經典文章，論文首先討論了音位功能負擔的概念，在前人研究的基礎上，王士元先生首次實現了功能負擔的計算和指出了計量功能負擔的困難，並給出了解決這些困難的方法。首先他討論了音位系統中常見的三種分佈、霍凱特與格林柏格的測量方法以及這些方法和香農（Shannon 1948，1951）的通信理論及各種語言學概念的關係。其次在這些背景知識的基礎上，王士元教授討論了功能負擔計量必須滿足的五個條件。最後他系統地發展了四種計量功能負擔的方法。另外，王士元先生還指出：「音變嚴重受到其他許多因素的影響，如音位之間語音的相似度和語言的接觸等，但正如許多歷史語言學家相信的那樣，如果功能負擔在音變中確實起作用的話，那麼用量化的解釋至少可以從一個方面闡明音變這一難題」。關於音位系統的分佈，王士元教授指出：「對任意一個語音序列，有三種相互關聯的分佈，即相似性分佈、交叉性分佈和互補性分佈。當語音序列中每一個音位的排列都共有一組相同的環境時，它就處於相似性分佈中；當音位的排列共有一些相同的環境，而不是所有的環境時，該序列處於交叉性分佈中；當音位排列沒有任何共有環境時，該序列處於互補性分佈中」。王士元教授的研究為後來功能負擔的研究建立了一個理論上的基本框架。實際上現代語音辨識技術中常用的雙音子和三音子的概念就起源於音位功能負擔量的研究。

2. 音位負擔量的研究方法

過去對音位負擔量的研究主要是通過大文本的計算，由於世界上大多數語言沒有文字，因此，研究只能在個別有文字的語言中進行，這無疑限制了這一領域的發展。根據多年的研究，發展出一種基於基本語素計算詞音位負擔量的方法，從而可以對每一種語言進行音位信息量和音位結構的計算。本文以藏緬語為例，介紹基於小詞彙量的音位負擔量、信息量和結構的基本計算方法，同時從音位負擔量角度討論藏緬語語位元結構、音位負擔量及其演變的本質。

「音位負擔」定義為：一種語言音位系統內部音位結構、音位分佈和音位功能負擔，稱為「音位結構功能負擔（phoneme structural functional load）」，簡稱「音位負擔」（phoneme load）。語言音位系統內部音位結構、音位分佈和音位功能負擔研究稱為「音位結構功能負擔研究」，簡稱「音位負擔研究」。「音位負擔研究」主要是以一種語言基本語素的內部音位結構、音位分佈和音位功能負擔」為研究物件。而「言語功能負擔（functional load）研究」是以一種語言的實際文本為基礎的音位功能負擔研究，兩者有本質的區分。在本研究中只用單音節語素，主要是單音節詞。計算採用聲、韻和調為基本音位單位。

在音位負擔的定義上，筆者將兩個單音節語素定位為一個語言的資訊單位，一個資訊單位有幾個不同的音位來負擔。如果由一個音位來區別這兩個語素，這個資訊單位就負載在這個音位上，如果是由多個音位來區別，詞信息量就負載在多個音位上。

為了能對計算方法進行清楚地說明，筆者對計算中提出的概念進行詳細解釋。一，字目：每一個字稱為一個字目；二，對立：不同音節稱為音節對立，可體現在兩個音節的聲母對立、韻母對立和聲調對立。三，音節資訊單位：在封閉空間內，一個音節同音字的個數為這個音節的資訊單位。四，語言資訊單位：一個字目承載一個語言資訊單位。

在漢藏語系語言中，基本語素大多是單音節，每個音節均由聲母和韻母或聲母、韻母和聲調組成。以漢語為例，音節和音節之間的對立總共有八種類型，其中「三項對立」一種，「二項對立」三種，「單項對立」三種，「無對立」一種，這形成了漢語普通話音位元系統結構和分佈的基本形式和框架，見表9.1。

「三項對立」是指兩個單音節語素之間聲韻調都不同；「兩項對立」是指兩個單音節語素之間只有兩個音位單位不同，即聲韻不同或聲調不同或韻調不同；「單項對立」是指兩個單音節語素之間只有一個音位單位不同，即聲母不同或韻

表 9.1　漢語普通話音位對立類型

對立類型	對立方式		
三項對立	聲/韻/調對立		
兩項對立	聲/韻對立	聲/調對立	韻/調對立
單項對立	聲母對立	韻母對立	聲調對立
無對立	聲/韻/調相同		

母不同或聲調不同；「無對立」是指兩個單音節語素之間沒有音位單位不同，即同音詞。根據以上的基本定義，以彝語喜德話為例，選了 6 個單音節詞和其相關資料，見表 9.2。

表 9.2 中第一列是序號，第二列是漢義，第三列是國際音標，第四列是聲母的國際音標，第五列是韻母的國際音標，第六列是聲調的調值。從表 9.2 可以看出，聲調有 3 種，聲母有 5 種，韻母有 3 種。

表 9.2　喜德彝語例詞表

序號	漢義	彝 (喜德)	聲母	韻母	聲調
1	天	mu33	m	u	33
2	山	bo33	b	o	33
3	肝	si21	s	i	21
4	胃	hi55	h	i	55
5	汗	ku21	k	u	21
6	士兵	mo55	m	o	55

在計算時取此表中的第一個詞，將其聲韻調分別和其他所有詞對比，如果是最小對立對就得 1 分，具體方法是：如果是聲母對立就給這個詞的聲母加 1 分，如果是韻母對立就給韻母加 1 分，如果是聲調對立就給聲調加 1 分，為了方便計算，1 分的數值是 6，這樣可以避免小數。表 9.3 是這 6 個詞聲韻調的分數，本文將這種對立稱為「單項對立」。從表 9.3 可以看出，在「單項對立」上，聲母的得分較高，韻母和聲調較低。

表 9.3　單項對立表

聲母對立	韻母對立	聲調對立
192	90	12
408	30	24
246	30	12
246	18	42
120	24	24
216	18	36

眾所周知，在語言的音位系統中冗餘度是普遍存在的，在詞彙這一層面往往體現為「兩項對立」和「三項對立」。「兩項對立」本文定義為兩個詞之間，對立是由聲母、韻母和聲調中的兩項來完成，如兩個詞是靠聲韻來完成對立就分別給聲母和韻母各加二分之一分，其數值為 3；如果是由聲調兩項完成對立，就分別給聲母和聲調各加二分之一分，其數值為 3；如果是由韻調兩項完成對立，就分別給韻母和聲調各加二分之一分，其數值為 3；見表 9.4。從表 9.4 可以看出，聲韻對立的數值最大，聲調對立的數值比較小，韻調對立的數值最小。

表 9.4　聲韻調「二項對立」表

聲韻 (聲)	聲韻 (韻)	聲調 (聲)	聲調 (調)	韻調 (韻)	韻調 (調)
1161	1161	171	171	15	15
1077	1077	195	195	33	33
297	297	330	330	54	54
459	459	300	300	42	42
357	357	201	201	66	66
489	489	297	297	45	45

第三種情況是「三項對立」，即兩個詞之間聲韻調都不同，對比詞的聲韻調各加三分之一分，其數值是 2。因此在「三項對立」中，聲韻調的得分完全相同，見表 9.5。從表 9.5 可以看出，6 個詞條聲母、韻母和聲調的得分數值雖然有差別，但每一詞條的數值是相同的。

表 9.5　聲韻調「三項對立」表

聲韻調 (聲)	聲韻調 (韻)	聲韻調 (調)
594	594	594
562	562	562
1044	1044	1044
948	948	948
1118	1118	1118
950	950	950

與表 9.4 相比可以看出，「二項對立」的情況比較複雜，「單項對立」和「三項對立」的情況比較簡單，為了更加清楚的查看結果，「二項對立」的六種情況合為聲韻調三種情況，見表 9.6。從表 9.6 可以看出，表中只有聲、韻和調 3 列數值。從具體的數值來看，聲母的得分數值最大，韻母次之，而聲調的得分數值最小。因此可以得知，在「兩項對立」中，聲母的音位負擔最大，韻母的音位負擔次之，而聲調的音位負擔最小。

表 9.6　聲韻調「二項對立」綜合表

二聲母總值	二韻母總值	二聲調總值
1332	1176	186
1272	1110	228
627	351	384
759	501	342
558	423	267
786	534	342

從彝語六個例詞八種類型的資料來看，「三項對立」的數值最大，其次是「兩項對立」，最小是「單項對立」。從本文使用的藏緬語所有資料來看，情況也是如此。從這分析結果可以看出，音位學中的對立原則在漢藏語以單音節和聲韻調為基本語音和音位單位的語言中，實際上是以「三項對立」和「二項對立」為主，「單項對立」的作用很小，並不是音位對立的主體。

表 9.7　聲韻調音位負擔量總數表

聲母總值	韻母總值	聲調總值
2118	1860	792
2242	1702	814
1917	1425	1440
1953	1467	1332
1796	1565	1409
1952	1502	1328

為了方便分析，表9.7給出了聲母、韻母和聲調音位負擔量總數值，聲母為11978、韻母為9521和聲調為7115。從這些數值可以看出彝語聲韻調的音位負擔量是有差別的，而且聲母最大，聲調最小。這結果使筆者可以來研究整個藏緬語聲韻調的音位負擔量，因而為定量評價藏緬語聲韻調各自的功能和類型研究奠定了基礎，也開闢了定量分析不同語言音位系統和語言演化的新領域。[2]

3. 藏緬語的音位負擔量

根據以上對音位負擔量的定義和計算方法，筆者計算部分藏緬語的音位負擔量，藏緬語的資料庫是根據《藏緬語族語言詞彙》（黃布凡 1992）建立，在本項研究中只用了其中的單音節詞，由於每種語言的單音節數量不同，本文的計算結果除了實際的數值外，還計算了聲韻調比值，這樣就可以對所有語言進行對比研究，見表9.8。

表9.8中第一列是序號，序號是根據聲調音位負擔量比值的大小排序而成，為遞增排序；第二列是語言或某語言的方言名稱；第三列是聲母音位負擔量總值；第四列是韻母音位負擔量總值；第五列是聲調音位負擔量總值；第六列是聲韻調音位負擔量總值；第七列是聲母音位負擔量比值；第八列是韻母音位負擔量比值；第九列式聲調音位負擔量比值。其中第三列至第六列的資料要處理詞彙的總數才可使用在比較研究方面，不能簡單和單獨地使用。每種語言具體聲韻調音位負擔量的研究結果將另文討論，本文主要討論藏緬語音位元負擔量的基本性質。

表 9.8　藏緬語聲韻調音位負擔量及聲韻調比值表

序號	方言點	聲母總值	韻母總值	聲調總值	聲韻調總值	聲母比例	韻母比例	聲調比例
1	嘉戎	122112	112248	0	234360	0.52	0.48	0.00
2	藏 (夏河)	2157414	2043690	0	4201104	0.51	0.49	0.00
3	藏 (書面語)	2576946	2491050	0	5067996	0.51	0.49	0.00
4	羌	1686158	1668194	3020	3357372	0.50	0.50	0.00
5	藏 (阿力克)	1059990	1006410	2376	2068776	0.51	0.49	0.00
6	博嘎爾珞巴	848218	904144	2278	1754640	0.48	0.52	0.00
7	墨脱門巴	2565864	2576694	1024866	6167424	0.42	0.42	0.17
8	獨龍	1150068	1160004	462984	2773056	0.41	0.42	0.17
9	義都珞巴	101120	89210	42794	233124	0.43	0.38	0.18
10	邵域	2460136	2368204	1208860	6037200	0.41	0.39	0.20
11	阿儂怒	87334	84742	45316	217392	0.40	0.39	0.21
12	彝 (南華)	1630160	1564388	847544	4042092	0.40	0.39	0.21
13	彝 (喜德)	1618072	1371046	827170	3816288	0.42	0.36	0.22
14	克倫	474610	465448	263122	1203180	0.39	0.39	0.22
15	貴瓊	371858	352190	202964	927012	0.40	0.38	0.22
16	仙島	1994042	2039090	1133312	5166444	0.39	0.39	0.22
17	阿昌	2517026	2583032	1440146	6540204	0.38	0.39	0.22
18	格曼	435402	440184	250686	1126272	0.39	0.39	0.22
19	浪速	2297168	2384534	1353014	6034716	0.38	0.40	0.22
20	波拉	2088520	2173318	1231678	5493516	0.38	0.40	0.22
21	錯那門巴	665470	676474	396448	1738392	0.38	0.39	0.23
22	載瓦	2242938	2331534	1353168	5927640	0.38	0.39	0.23
23	達讓	198802	193300	116794	508896	0.39	0.38	0.23
24	哈尼 (綠春)	1058744	998612	617960	2675316	0.40	0.37	0.23
25	景頗	1000370	1043324	618854	2662548	0.38	0.39	0.23

表 9.8（續）

序號	方言點	聲母總值	韻母總值	聲調總值	聲韻調總值	聲母比例	韻母比例	聲調比例
26	藏 (巴塘)	1490360	1434542	895682	3820584	0.39	0.38	0.23
27	彝 (巍山)	1713606	1579020	1009098	4301724	0.40	0.37	0.23
28	史興	495886	466402	296812	1259100	0.39	0.37	0.24
29	勒期	1677070	1772518	1072228	4521816	0.37	0.39	0.24
30	哈尼 (墨江)	1021232	975956	622436	2619624	0.39	0.37	0.24
31	傈僳	1479748	1387264	951208	3818220	0.39	0.36	0.25
32	白	1185304	1231198	833434	3249936	0.36	0.38	0.26
33	納木茲	4148886	4108596	2855610	11113092	0.37	0.37	0.26
34	緬 (書面語)	2345622	2272662	1599816	6218100	0.38	0.37	0.26
35	藏 (拉薩)	1428336	1421790	1017570	3867696	0.37	0.37	0.26
36	彝 (武定)	2809544	2553182	1921742	7284468	0.39	0.35	0.26
37	土家	664624	620026	470794	1755444	0.38	0.35	0.27
38	紮壩	4729342	4709830	3527248	12966420	0.36	0.36	0.27
39	呂蘇	2481422	2414234	1885064	6780720	0.37	0.36	0.28
40	彝 (撒尼)	1695308	1447904	1219292	4362504	0.39	0.33	0.28
41	基諾	767482	694054	571300	2032836	0.38	0.34	0.28
42	緬 (仰光)	4555552	4465258	3541774	12562584	0.36	0.36	0.28
43	普米 (蘭坪)	6272854	6186160	4963510	17422524	0.36	0.36	0.28
44	納西	6410460	6170514	5016138	17597112	0.36	0.35	0.29
45	怒蘇怒	6144024	5974662	4955754	17074440	0.36	0.35	0.29
46	普米 (九龍)	6123764	6051278	5104298	17279340	0.35	0.35	0.30
47	拉祜	1420182	1253196	1202202	3875580	0.37	0.32	0.31
48	嘎卓	886158	804126	765120	2455404	0.36	0.33	0.31
49	木雅	3599292	3606120	4308216	11513628	0.31	0.31	0.37

從資料可以看出，前六個語言或方言是：一、嘉戎；二、藏（夏河）；三、藏（書面語）；四、羌；五、藏（阿力克）；六、博嘎爾珞巴。這六個語言或方言沒有聲調，因此聲調的音位負擔量為零，排在表的最前面。這六種語言或方言聲母和韻母的音位負擔量比較大，其和等於1。聲調音位負擔量最大的六個語言是：一、納西；二、怒蘇怒；三、普米九龍話；四、拉祜；五、嘎卓；六、木雅，其比值分別是0.29、0.29、0.30、0.31、0.31和0.37。從這些基本資料可以看出，藏緬語聲韻調的音位負擔量是不同的，它們反映了各自語言的音位系統的差別和發展的不同程度，這為我們研究藏緬語音位元體系、音位結構和音位負擔量及語言資訊之間的關係奠定了基礎。本項研究的計算結果也可以按聲母或韻母來排序，從而進一步研究聲母和韻母的特性。

4. 音位負擔量的特性

根據本文對詞彙層面音位負擔量的定義和對藏緬語49個語言和方言的計算，可以得到許多結果和結論，限於篇幅本文在此先從宏觀的角度就一些主要的結果進行討論。有關聲母、韻母和聲調音位負擔量的內部結構、分佈以及和語言係數的關係將放在另外的文章中討論。

首先，從圖9.1可以看出：一、聲調從無到有，體現為聲母和韻母音位負擔量的下降。二、在聲調的音位負擔量為0時，聲母和韻母的負擔量比有聲調的聲韻母負擔量要高許多，兩者之間有一個跳躍。三、在有聲調的語言和方言中，聲調的信息量和聲韻母的信息量成反比關係，也就說一種語言裏，如果聲調承載的信息量大，聲母和韻母能承載的信息量就是小。四、大多數語言聲母和韻母的負擔量在一個數量級上，相對來說比較大，而聲調的音位負擔量要比聲韻母小很多。從這49個語言和方言上來看，聲調音位負擔量在有些語言中已經很接近聲母和韻母的數值，這體現了藏緬語聲調音位負擔量發展程度的不同。見圖9.1。

顯然在藏緬語中，聲調的音位負擔量是很不同的，這反映了藏緬語聲調在不同語言中發展和演化的不同階段。最小的只有0.17，而最大的為0.37，這個數值大過該語言的聲母和韻母的音位負擔量。

第二，宏觀上聲母和韻母的信息量相對於聲調來說基本相等，也就是說，一種語言裏，聲母能承載的信息量和韻母基本相同，但和聲調有很大的差別。因此從音位結構、功能和層次上可以看出，在藏緬語的發展過程中，聲母和韻母在結構和功能上是一個層次，隨着聲調的發展和演化，聲調的功能逐漸增強，而聲母和韻母的功能相對減弱。

圖9.1　藏緬語聲韻調音位負擔量比值表

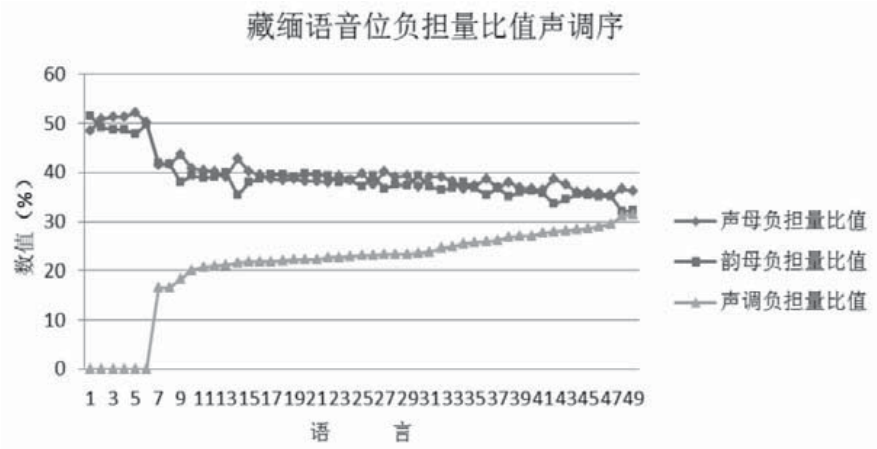


圖9.2　音位位負擔量比值聲調排序

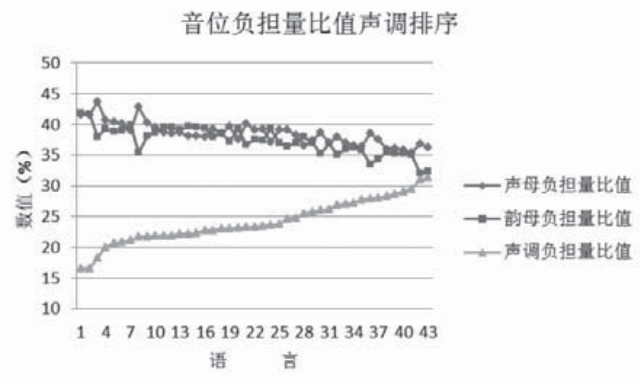


圖9.3　音位負擔量比值聲母排序

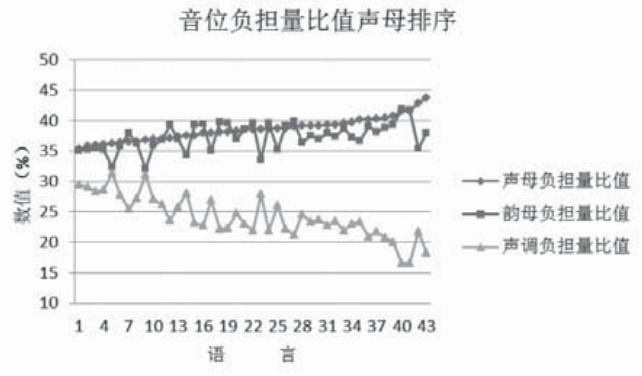


圖9.4　音位負擔量比值韻母排序

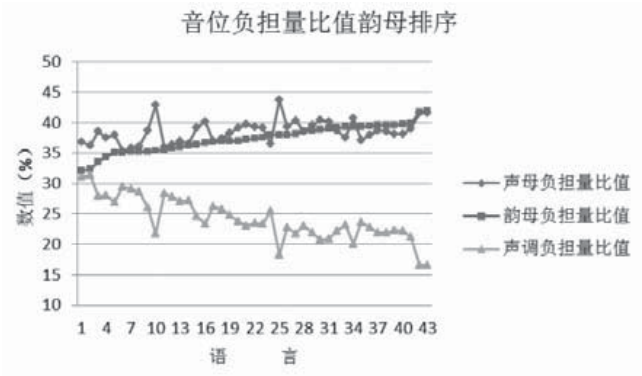
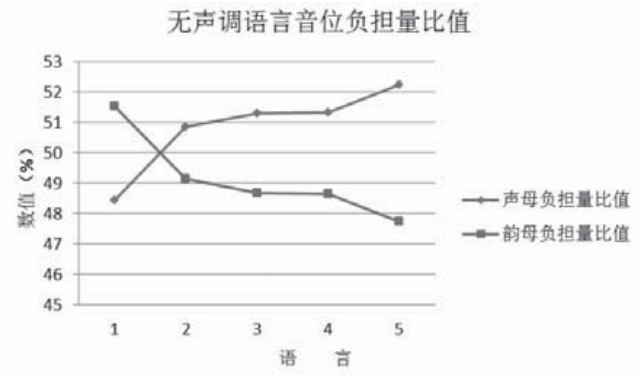


圖9.5　無聲調語言聲音位元負擔量比值



根據本文的計算方法，按聲調排序可以發現聲母和韻母的音位負擔量成鏡像分佈，見圖9.2；按聲母排序可以發現聲調和韻母的音位負擔量成鏡像分佈，見圖9.3；按韻母排序可以發現聲調和聲母的音位負擔量成鏡像分佈，見圖9.4，但3者在數值上有較大差別。

第三，在無聲調的語言裏，聲母和韻母在信息量上成反比關係，也就說聲母承載的信息量大，韻母承載的信息量就小，否則相反，見圖9.5和表9.8。總的來說，在這些語言呈現出鏡像的分佈中，聲母和韻母的數值差別不太大，基本在50%左右。這一點和採用了聲母和韻母做基本單位有關。可以看出前面六種語言是無聲調的語言，其聲調的音位負擔量為0，所以音位負擔量由聲母和韻母承載，聲韻母的負擔量比有聲調語言的聲韻母要高許多，因為其總和等於100%。

第四，對於聲母和韻母來說，音位數量的增減不太影響聲韻母總的音位負擔量。以藏語為例，從古藏（9世紀）語到現在的大多數方言，複輔音聲母大量脫落，但聲母總的音位負擔量並沒有減少，而是由剩餘的聲母承載，直到聲調產

生，聲母整體音位負擔量下降。這無疑從一個語言內部的音位負擔量為研究語言的演化提供了線索。

5. 討論

本文的研究表明，語言音位負擔的語言學研究能解釋許多語言信息量的結構、承載和演變的性質和規律。從49個藏緬語語言和方言的情況看，有聲調語言和無聲調語言之間存在一個跳躍，聲韻調音位負擔量的曲線不是平滑的，主要原因是：在調查語言的過程中，聲調產生的過程沒有調查出來。如，藏語有許多方言正處於聲調的產生過程中，而目前採用的結構主義音位學的調查方法不能很好地或者說精確地描寫聲調產生的過程。另外，在計算方法上，由於不同音位對立類型的數量級差別較大，也有可能掩蓋了一些內部的規律，因此在計算上做進一步的加權演算法研究是很必要的。最後希望這項研究能成為研究語言學研究和信息量研究的一個很好的切合點，形成科學地研究語言演化的一個新的方向。

注釋

1. 本項研究得到了國家社會科學重大項目「中國有聲語言及口傳文化保護與傳承的數字化方法研究和基礎理論研究」的支持，批准號：10&ZD125。
2. 音位負擔量計算的數學和程序細節將發表在〈音位負擔量的計算方法〉一文中。

參考文獻

黃布凡主編 1992。《藏緬語族語言詞彙》。北京：中央民族學院出版社。

鄭錦全 2006。〈從詞語八千到學海無涯〉為講座題目。高雄：國立中山大學中文系。

Greenberg, H. H. 1959. A method of measuring functional yield as applied to tone in Mrican languages. *Georgetown University Monograph Series on Languages and Linguistics* 12:7–16.

Hockett, C. F. 1955. *A manual of phonology*, 218. Baltimore Waverly Press.

Hockett, C. F. 1961. The quantification of functional load. *Rand report,* 2338. Santa Monica.

Hoenigswald, H. M. 1960. *Language change and linguistic reconstruction*, 79-80. University of Chicago Press, Chicago.

King, R. D. 1955. Functional load: Its measure and its role in sound change. University of Wisconsin PhD dissertation. A version of this will appear in *Language*.

Shannon, C. E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30:50–64.

Shannon, C. E., and W. Weaver. 1949. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Wang, W. S-.Y. 1967. Phonemic Theory A (with Application to Midwestern English), Doctoral dissertation. The University of Michigan, Stress in English. Language

Wang, W. S.-Y., and J. Crawford. 1960. Frequency studies of English consonants. *Language and Speech* 3:131–139.

Cheng, C. C. 1999. 詞涯八千 (Activevocabulary upper limit 8000), Graduate Institute of Linguistics, National Taiwan University, Taipei, May 10. Invited lecture.

Quantitative High-Speed Laryngoscopic Analysis of Vocal Fold Vibration in Fatigued Voice of Young Karaoke Singers

*Edwin M.-L. Yiu, *Gaowu Wang, *Andy C.Y. Lo, *Karen M.-K. Chan, *Estella P.-M. Ma, †Jiangping Kong, and *Elizabeth Ann Barrett, *Pokfulam, Hong Kong, and †Beijing, China*

Summary: Purpose. The present study aimed to determine whether there were physiological differences in the vocal fold vibration between nonfatigued and fatigued voices using high-speed laryngoscopic imaging and quantitative analysis.

Methods. Twenty participants aged from 18 to 23 years (mean, 21.2 years; standard deviation, 1.3 years) with normal voice were recruited to participate in an extended singing task. Vocal fatigue was induced using a singing task. High-speed laryngoscopic image recordings of /i/ phonation were taken before and after the singing task. The laryngoscopic images were semiautomatically analyzed with the quantitative high-speed video processing program to extract indices related to the anteroposterior dimension (length), transverse dimension (width), and the speed of opening and closing.

Results. Significant reduction in the glottal length-to-width ratio index was found after vocal fatigue. Physiologically, this indicated either a significantly shorter (anteroposteriorly) or a wider (transversely) glottis after vocal fatigue.

Conclusion. The high-speed imaging technique using quantitative analysis has the potential for early identification of vocally fatigued voice.

Key Words: Vocal fatigue–High-speed imaging–Amateur singing.

INTRODUCTION

Vocal fatigue

Vocal fatigue is a common complaint found in teachers, sales professionals, singers, and individuals who constantly use their voice for a prolonged period. It is often described as an increased effort in voicing, harshness, strained voice quality, dryness, and sensation of pain in the throat.[1,2] Some authors considered vocal fatigue as one of the symptoms of voice disorders.[3] Others considered it as an isolated phenomenon.[2,4,5] Stemple et al[2] reported that subjects complained about a dry sensation in the throat and effortful speaking after reading aloud for 2 hours. Teachers with vocal fatigue rated their voice with increased harshness, breathiness, and strain after a day of teaching.[1]

There is yet a consensus on the definition for vocal fatigue. It is generally regarded as vocal tiredness after voice overuse, misuse, or abuse.[2,6] It could happen in speakers with or without any voice problem. Chronic vocal fatigue, however, could be an indicator of subsequent voice disorder.[2,3]

Researchers have attempted to investigate vocal fatigue using aerodynamic, acoustic, and laryngoscopic evaluations. For example, phonation threshold pressure was found to increase after prolonged reading in both women[4] and men.[5] Solomon and DiMattia[4] found that the phonation threshold pressure increased under low-hydration condition. They argued that the viscosity of the vocal folds increases when fatigued. Without sufficient hydration to the vocal folds during prolonged voice use, the stiffness of vocal folds increased, and more effort was required to initiate the vibration of vocal folds. This would lead to an increase in phonation threshold pressure. Stemple et al[2] reported a significant increase in the fundamental frequency after 2 hours of reading. Gelfer et al[7] reported that untrained singers demonstrated an increase in aperiodicity (jitter) and noise levels. Inefficient vocal functions, as indicated by increased airflow rate and reduced maximum phonation time, were found in subjects with chronic laryngeal fatigue, but the fundamental frequency and jitter values remained within normal limit.[8] Indeed, the literature showed mixed results when acoustic measures were used to examine vocal fatigue. The inconsistent results could have been attributed to the different methodologies used in different studies.

With the use of videolaryngostroboscopic examination, anterior glottal chinks[2,8] and abnormal spindle-shaped closure[4,8] have been reported in speakers with vocal fatigue. The presence of chinks or incomplete closure correlated with the perceptual finding of increased breathiness and airflow. Stemple et al[2] hypothesized that the thyroarytenoid muscles would become strained and weak in a fatigue state. Such weakness would cause the bowing of the vocal folds and would lead to an incomplete closure. Prolonged strained contraction of the muscles would give rise to a sensation of pain and increased effort in voicing. In a study by Mann et al,[9] a significant increase in vocal fold edema was observed after a 5-day vocally demanding training. The investigators contended that vocal fold tissues were damaged after an extended vocally demanding task. Gelfer et al[10] found an increase in the amplitude of glottal opening after 1 hour of loud reading. They[10] suggested that the participants might have adopted the loud speaking mode even during the endoscopy task.

The studies reviewed the aforementioned laryngostroboscopic technique, which provided only a pseudo slow motion analysis of the vibration. Intracycle vocal fold vibration pattern

Accepted for publication June 13, 2013.

From the *Voice Research Laboratory, Division of Speech & Hearing Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China; and the †Department of Chinese, Peking University, Beijing, China.

Address correspondence and reprint requests to Edwin M.-L. Yiu, Voice Research Laboratory, Division of Speech & Hearing Sciences, The University of Hong Kong, 5/F Prince Philip Dental Hospital, 34 Hospital Road, Pokfulam, Hong Kong, China. E-mail: eyiu@hku.hk

Journal of Voice, Vol. ■, No. ■, pp. 1-9
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2013.06.010

would not be possible with this technique because of the limited camera recording rate. The camera records around 24–30 frames per second. This recording rate could not have captured even one complete vibratory cycle, with the vocal folds vibrating at 80–300 cycles per second. The stroboscopic technique produces images of vibratory motion by sequencing different image frames from different phases across many glottal periods,[11] and hence, videolaryngostroboscopic images do not show a complete full glottal cycle.

High-speed laryngoscopy

High-speed laryngoscopic imaging system emerged in the last 10 years. It can capture up to 8000 frames per second, compared with the ordinary camera system that captures around 24–30 frames per second. High-speed imaging technique has gained popularity over the past decades because of reduced cost and improved resolution of the equipment. With the high-speed system using digital technology, it is possible to examine complete cycle-by-cycle vocal fold vibratory patterns.[11] Cycle-to-cycle visualization allows any irregular vibratory cycle or phase asymmetry to be identified. In a study that compared the usefulness of high-speed imaging and videolaryngostroboscopy in identifying the vibratory features in dysphonic voice,[12] it was found that the use of the high-speed imaging system achieved a 100% identification rate for dysphonic voices compared with the use of videostroboscopy that could only identify 37% of the dysphonic voices correctly. This finding indicates that the videolaryngostroboscopy would not be effective in capturing pathological voices with aperiodic signals. It is, therefore, reasonable to expect that the use of high-speed laryngoscopic imaging system would provide more relevant intracycle information for investigating the physiological changes of fatigued vocal folds, as intracycle aperiodic vibration has been reported as one of the features of vocal fatigue.[7]

Both qualitative and quantitative data of normal and pathological high-speed laryngoscopic images have been reported in the literature. Qualitative methods have been found to be useful in identifying specific vibratory patterns of the vocal folds. For example, specific patterns of glottal closure were reported in diplophonic phonation using kymography, which is a high-speed line scanning technique.[13] A more recent qualitative high-speed laryngoscopic study by Inwald et al[14] reported the glottal closure, degree of mucosal wave, asymmetry, and the amount of mucus deposit in individuals with voice disorders. These qualitative studies were based on perceptual evaluations. Reliability is always an issue in perceptual measurement as it is a subjective rating process. Interrater reliability in perceptual evaluation of laryngoscopic images is usually no better than 70% (eg, Patel et al[12]).

Quantitative measurement (eg, quantifying glottal area in pixels), on the other hand, is a more objective and less variable method that facilitates data summarization and interpretation. Quantitative measures of the glottal area, glottal width, and glottal length have been reported to be useful for studying normal and dysphonic phonations (eg, Yiu et al[15]). The study by Inwald et al[14] also reported the use of asymmetry and perturbation measures extracted quantitatively from the high-speed images. These investigators[14] contended that the combined method of qualitative and quantitative evaluations could best differentiate between dysphonic and normal voices. Temporal measure using frame-to-frame analysis of the high-speed images has also been suggested to be useful in describing phonation. For example, fundamental frequency and open quotient (ie, ratio of the open phase to the glottal period) can be determined by examining the number of frames with open and nonopen glottis.[16] Whether these measures are useful for describing vocal fatigue voice is still open to investigation.

It should be noted that quantitative methods are not always useful to describe different types of phonations. For example, Mehta et al[17] quantified the amount of left-right displacement waveforms of normal and pressed phonations using kymography. They found no significant difference in the asymmetry between these two types of phonations. In another study, Koster et al[18] quantified and analyzed the change in glottal area and glottal width during different modes of voice onset. They were also not able to find any significant differences among the different modes of voice onset.

The present study used the high-speed video processing (HSVP) program developed at the University of Hong Kong[15,19] to investigate vocal fold vibration in vocally fatigued voices. As reviewed previously, the observation of stiffness of vocal folds,[4] glottal chinks, abnormal spindle-shaped glottic closure,[4,8] and strained thyroarytenoid muscles that might cause bowing of the vocal fold[2] in fatigue voices suggested that quantitative analysis of the glottal configuration and vocal fold vibratory pattern would be a logical choice of assessment direction. The present study aimed to examine the glottal configuration and vocal fold vibratory pattern of fatigued voice, induced by prolonged singing,[20] using quantitative analysis of high-speed laryngoscopic images. It was hypothesized that fatigued and nonfatigued voices would demonstrate different glottal configurations and vibratory patterns because of changes in vocal fold physiology.

METHOD

Participants

Ten males and 10 females were recruited from the University of Hong Kong through the social circle of the second (G.W.) and third (A.C.Y.L.) authors. The participants were 18–23 years old (mean, 21.2 years; standard deviation [SD], 1.3 years). All participants reported to be free of any voice or general health problems, nonsmokers, nonalcoholic drinkers, and had no prior voice training. All the participants were Cantonese speakers who, at the time of the study, were attending or had completed their tertiary education. All the participants were further evaluated perceptually by the third author (A.C.Y.L.) to have normal voice quality at the time of the study. Participants were excluded if they reported to have respiratory disease, such as sore throat or flu, 1 day before the examination.

Procedure

Singing task inducing vocal fatigue. All participants were asked to undertake karaoke singing for a minimum of

95 minutes without rest and without drinking water to induce vocal fatigue. The singing time used in the present study was based on the mean plus two SDs singing time that resulted in vocal fatigue, as reported in a separate study by Yiu and Chan.[20] The loudness level of the background music was set at around 60 dB sound pressure level (SPL) for all participants, and the participants were required to sing at least above 80 dB SPL measured with a sound pressure meter (TES 1530A) at a distance of 30 cm from their mouth.

After singing for 95 minutes, all the participants reported feeling tired. They were further asked to continue singing until they felt they could not sing anymore. The purpose of this procedure was to ensure that all the participants had indeed achieved a vocal fatigue condition after the minimum recommended time (at least 95 minutes) of singing. The final mean singing time among the participants was 103.8 minutes (SD, 7.2 minutes; range, 95–115 minutes).

Participants' self-ratings of vocal conditions. Before the singing task, each participant was asked to rate his or her own vocal conditions on the level of discomfort, dryness in the throat, and effort used in voicing, using an 11-point rating scale (0 = normal and 10 = most severely affected). After the singing task, each participant also rated his or her vocal conditions again.

High-speed laryngoscopic and voice recordings. High-Speed camera 5562 digital high-speed imaging system (Richard Wolf GmbH, Knittlingen, Germany) was used to record the laryngoscopic images before and after the singing task. This was performed by the fourth author (K.M.-K.C.), a qualified speech pathologist who had more than 10 years of experience in conducting laryngoscopy. Synchronized voice signals (in WAV format) were also recorded by a microphone attached to the endoscope at approximately 10 cm from the participant's mouth. The participants were asked to sustain an /i/ phonation for as long as possible at their most comfortable pitch and loudness with their tongue protruded. Two seconds of the sustained /i/ phonation with the onset and offset excluded were captured by the digital high-speed imaging system. Synchronized voice signals were recorded by a microphone attached to the endoscope, which was approximately 10 cm from the mouth. A total of 8192 frames (in AVI format) were recorded for the 2-second time span in each recording. Each participant produced three /i/ phonations for recording before and also after the singing task. Therefore, a total of six recordings were collected for each participant.

Preparation of high-speed laryngoscopic images and synchronized voice samples

The high-speed image recordings were analyzed using the HSVP program developed by the Voice Research Laboratory, The University of Hong Kong.[15,19] A number of ratio indices and temporal measurements could be extracted using the HSVP program. The present study, however, focused on four measures only: fundamental frequency, length-to-width ratio index of the glottis (based on the maximum opening of 100 vibratory cycles), open quotient, and speed quotient.

The third author (A.C.Y.L.) first carried out five manual steps to select and prepare the images for the automatic analysis by the HSVP program:

1. *Selection of a comparable pair of presinging and postsinging recordings*. From among each of the three recordings produced by each participant before and after the singing task, a pair of presinging and postsinging recordings with comparable quality and pitch level was selected for each participant. Each pair of presinging and postsinging recordings was perceptually judged to be similar in pitch and loudness by the third author (A.C.Y.L.). Subsequent analysis of the fundamental frequency and intensity of these selected presinging recordings (mean, 240.3 Hz; SPL, 90.4 dB, respectively) and postsinging phonations (mean, 241.2 Hz; SPL, 88.8 dB) showed no significant difference between them (Wilcoxon signed ranked test—fundamental frequency: $Z = -0.09$, $P = 0.93$; intensity: $Z = -1.36$; $P = 0.17$).
2. *Extraction of image frames for analysis*. A minimum of 100 vibratory cycles is considered to be necessary[19] for the HSVP analysis. Because the fundamental frequency for each participant was different, approximately between 1000 and 2000 stable frames of images that contained 100 cycles were extracted from the 8192 frames of each raw video recording according to the frequency produced by each of the participant. Frames that did not contain view of a full glottis were eliminated from the extraction process.
3. *Postextraction processing of images—resizing and gray-scale conversion of the images*. The HSVP program, at the time of the study, was designed to analyze gray-scale images with a resolution of 120 × 256 pixels (although the most current version of HSVP is now capable of analyzing color images of 256 × 256 pixels). Therefore, the size of the raw video was cropped into 120 × 256 pixels and converted into gray-scale image.
4. *Fine adjustment of image quality.* The processed images were rotated manually so that the longitudinal axis of the glottis aligned with the vertical axis (Figure 1). Manual zooming feature and brightness and contrast controls were also available for clearer visualization. The HSVP program has a build-in motion compensation function, which allows tracking the dynamic movement of the images because of the movement of the endoscope. Corresponding adjustment was made to keep the glottis to remain at the relative position across the frames using the automatic motion compensation function if there was endoscopic movement during the recordings.
5. *Delineation of glottis for analysis*. Once the structures in the image were clearly visualized using the fine adjustment described in step 4 previously, an analysis window (Figure 1) was added to the image to enclose the glottis. The window was placed on the left, right, anterior and posterior edges of the glottis.

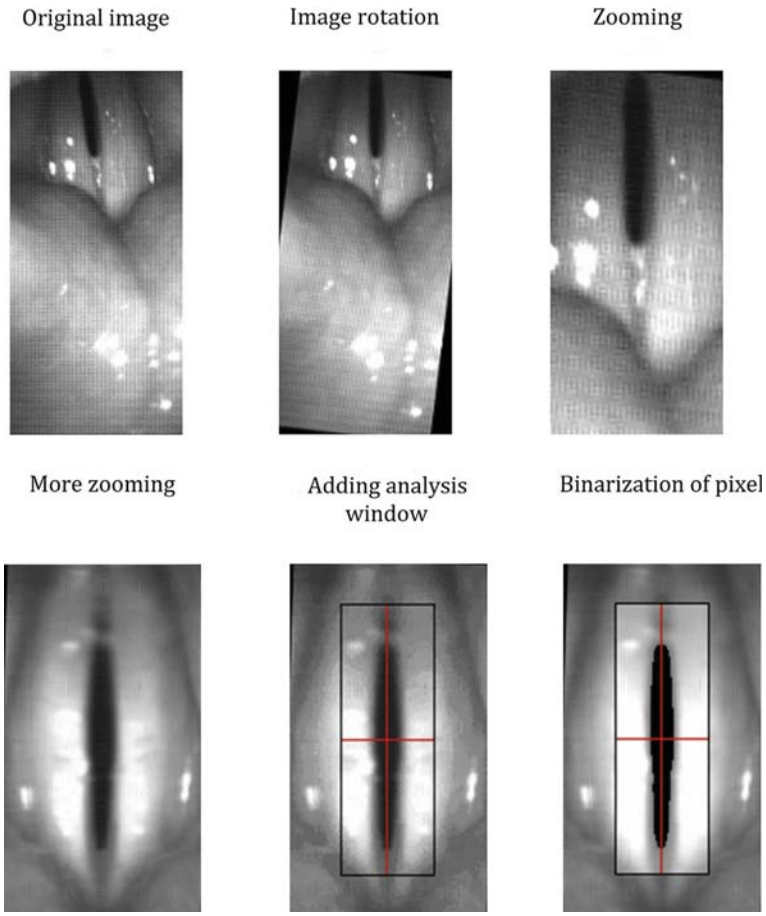


FIGURE 1. Procedures in analyzing high-speed laryngoscopic images.

Data analysis

Interrater and intrarater reliability. Because the preparation of the high-speed images required some subjective judgment, it was therefore necessary to determine the reliability in preparing these images. Eight video samples, which represented 25% of the total analyzed samples, were randomly selected for reanalysis to determine the reliability of frames selection and the effect on the calculation of the glottal measures. Intrarater reliability was carried out with the third author (A.C.Y.L.) analyzing the images 2 weeks later. Interrater reliability was carried out by comparing the analyses undertaken by the third author with those of the second author (G.W.), who worked as an independent rater.

Acoustic analyses. Fundamental frequency and the intensity level of the synchronized voice signals were extracted using the *Praat* software.[21] This was carried out by the third author (A.C.Y.L.).

Extraction of the glottal measures. A procedure to determine the vibratory (fundamental) frequency was carried out on each extracted video image using the HSVP program by the third (A.C.Y.L.) author. The HSVP program calculated the vibratory (fundamental) frequency by transforming the number of frames into a time function based on the sampling rate at 4000 frames per second. For example, if 200 completed vibratory cycles were identified within a 4000-frame video sample, it will be equivalent to 200 cycles per second (ie, fundamental frequency is equal to 200 Hz). The glottal length-to-width ratio index, open quotient, and speed quotient were extracted by analyzing the pixels of the glottis. The HSVP program automatically binarized the pixels into black and white within the analysis window. The black pixels represented the size of the glottis. As it was impossible to determine the actual size of the glottis because of the unknown magnification factors (ie, the distance between the vocal folds and the laryngoscope), a length-to-width ratio index of the glottis based on the maximum opening of 100 vibratory cycles was calculated. A higher index indicated the shape of the glottis to be longer (anteroposteriorly) and/or narrower (transversely) during the maximum opening. The open quotient (the ratio of the glottal opening over one vibratory cycle—calculated by dividing the duration of the open phase by the glottal period) and the speed quotient (the symmetry between the open phase and the close phase—calculated by dividing the duration of opening by the duration of the closing within the open phase) were extracted automatically by the HSVP program. A high open quotient indicated a longer glottal opening in a given cycle, whereas a high speed quotient suggested a longer glottal opening and shorter glottal closing in a given cycle.

TABLE 1.
Mean (SD) for Self-Ratings of Vocal Conditions on an 11-Point Rating Scale (0–10)

	Mean (SD)	
Vocal Conditions	Presinging	Postsinging
Discomfort in the throat*	0.90 (0.91)	6.80 (1.74)
Dryness in the throat*	0.95 (1.00)	7.70 (1.78)
Effort in voicing*	0.55 (1.05)	6.95 (1.99)

* The asterisk symbol indicates significant difference ($P < 0.05$) between singing.

RESULTS

Participants' self-ratings on vocal conditions

Table 1 lists the participants' mean self-rating of vocal conditions before and after the singing task. Bonferroni adjustment was used because three tests were carried out. The alpha was set at .017 (.05/3). The subjects complained of significantly more discomfort and dryness in the throat as well as more effortful in voicing (Wilcoxon signed ranked test for each of the conditions: $Z = -3.93$, $P = 0.001$).

Intrarater and interrater reliability measures

Table 2 lists the interrater and intrarater agreement measures in extracting image frames for the final analysis. For the intrarater agreement measure, 75% of the reanalyzed videos were within ±500 frames of the first analysis. The interrater agreement was lower, with 50% of the videos agreed within ±1000 frames between the two raters.

To determine whether two different samples selected by the same rater or the different raters produced significantly different results in the glottal measures, Wilcoxon signed rank tests were used to determine whether there were significant differences in the glottal measures between the analyses. No significant differences ($P > 0.09$) were found in the glottal measures between the two samples in the interrater and intrarater procedures.

Acoustic and high-speed glottal measures

Table 3 lists the mean acoustic measures (fundamental frequency and intensity) and the mean glottal frequency, mean glottal length-to-width ratio index, mean open and speed quotients before and after the singing tasks from the extracted high-speed images. Analyses were carried out using the combined data from the two gender groups and data from each gender group separately. Bonferroni adjustment with the alpha set at .017 (.05/3) was also used because three tests were carried out for each measure.

Gender differences. No significant differences were found between the two gender groups in the mean voice intensity, glottal length-to-width ratio index, and speed quotient in both presinging and postsinging conditions ($P > 0.05$). The female group, however, demonstrated a significantly higher open quotient when compared with the male group before singing ($Z = -3.78$, $P < 0.0001$) but not after singing ($P > 0.05$).

Changes after vocal fatigue. None of the fundamental frequency, intensity, open quotient and speed quotient measures showed any significant changes after singing ($P > 0.05$, Table 3). The glottal length-to-width ratio index, however, showed a significant reduction after the singing task both in the male and female groups ($P \leq 0.01$, Table 3). Although the mean reduction in the glottal length-to-width ratio index in the male group was bigger in magnitude (−1.67) than that in the female group (−0.56), no statistical significance was found ($Z = -1.96$, $P = 0.052$). The variability of the glottal length-to-width ratio index in the male subjects was indeed rather large, with an SD (1.61) larger than the mean (1.22, Table 3). Therefore, a closer examination of the individual data was conducted.

It was noted that nine males and six females (total = 15) showed a reduction in the ratio index after singing, whereas one male and four females (total = five) demonstrated an increase in the ratio index after singing. For those that demonstrated a reduction in the ratio index after vocal fatigue, the change in the ratio indices ranged from −0.28 to −4.33 and −0.03 to −3.30 in the male and female groups, respectively. For those who demonstrated an increase in the ratio index, the change in the ratio index was 0.63 in the male and ranged from 0.09 to 0.90 in the female groups.

DISCUSSION

The aim of this study was to examine the glottal configuration and vocal fold vibratory pattern after vocal fatigue using high-speed laryngoscopic imaging. Vocal fatigue was induced using a prolonged karaoke singing task. The subjects' self-perception of the vocal conditions was analyzed, and the glottal configuration and vibratory patterns were quantified using three primary glottal measures: glottal length-to-width ratio index, open quotient, and speed quotient.

Self-ratings on vocal condition before and after singing task

After a mean singing time of 103.8 minutes (SD, 7.2 minutes; range, 95–115 minutes), all the participants reported to have

TABLE 2.
The Interrater and Intrarater Agreement in the Extraction of Image Frames for Analysis

Agreement	±250 Frames	±500 Frames	±1000 Frames	±2000 Frames	±4000 Frames
Intrarater	37.5% (3/8)	75% (6/8)	87.5% (7/8)	87.5% (7/8)	100% (8/8)
Interrater	0% (0/8)	37.5% (3/8)	50% (4/8)	50% (4/8)	100% (8/8)

TABLE 3.
Acoustic Measures and Extracted High-Speed Glottal Measures for the Combined Group and Separate Gender Groups

	Mean (SD)		
Measures	Presinging	Postsinging	Wilcoxon Signed Rank Test
Frequency (Hz) measured acoustically	240.25 (71.13)	241.20 (71.97)	$Z = -0.09$, $P = 0.93$
Male	179.80 (17.66)	181.20 (23.31)	$Z = -0.10$, $P = 0.9$
Female	300.70 (47.41)	301.20 (48.91)	$Z = -0.15$, $P = 0.88$
Intensity (dB)	90.40 (6.75)	88.80 (6.46)	$Z = -1.36$, $P = 0.17$
Male	89.80 (7.36)	87.50 (8.41)	$Z = -1.30$, $P = 0.2$
Female	91.00 (6.43)	90.10 (3.69)	$Z = -0.54$, $P = 0.59$
High-speed extracted glottal frequency (Hz)	239.85 (69.67)	242.90 (72.13)	$Z = -1.09$, $P = 0.27$
Male	181.00 (19.25)	183.60 (25.02)	$Z = -0.46$, $P = 0.65$
Female	298.70 (46.69)	302.20 (50.44)	$Z = -0.89$, $P = 0.37$
Glottal length-to-width ratio index*	2.79 (1.12)	1.68 (1.52)	$Z = -2.65$, $P = 0.006$
Male	2.89 (1.39)	1.22 (1.61)	$Z = -2.49$, $P = 0.01$
Female	2.69 (0.81)	2.13 (1.35)	$Z = -2.70$, $P = 0.007$
Open quotient (%)	69.50 (9.6)	67.10 (11.9)	$Z = -1.07$, $P = 0.28$
Male	63.40 (9.24)	61.80 (12.22)	$Z = -0.59$, $P = 0.55$
Female	75.60 (5.10)	72.40 (9.35)	$Z = -0.76$, $P = 0.44$
Speed quotient (%)	106.40 (20.5)	106.60 (29.4)	$Z = -0.56$, $P = 0.57$
Male	98.60 (13.51)	103.30 (21.94)	$Z = -0.53$, $P = 0.59$
Female	114.10 (23.95)	109.90 (36.27)	$Z = -0.15$, $P = 0.88$

* The asterisk symbol indicates significant difference at 0.017 level.

vocal fatigue. The participants self-rated their level of discomfort in the throat, dryness, and voicing effort to be significantly worse after the singing task (Table 1). These three features are common signs of vocal fatigue (eg, Stemple et al[2]; Hunter and Titze[22]) and could be considered as cardinal perceptual signs for identifying vocal fatigue in speakers.

Reliability of the data processing

The reliability in extracting the high-speed glottal measures is dependent on the precision of the manual extraction procedures carried out by the investigator. The intrarater agreement within 1000 frames was more than 87%, whereas the interrater agreement was only 50%. Nevertheless, statistical results showed no significant impact on the final extracted measurements despite the discrepancy in the frame selection. The use of a large number of frames and hence averaging out the measurements could have possibly reduced the impact on the final extracted data, despite the relatively moderate interrater agreement. It is essential that these reliability data should be considered carefully in any imaging processing studies, and investigators should ensure that every effort has been made to obtain the highest reliability or agreement.

High-speed glottal measures

The mean glottal length-to-width ratio index decreased significantly from 2.79 to 1.68 after the prolonged singing task (Table 3). This suggested that the glottis in the fatigued voice demonstrated a relative shortening of the vocal folds anteroposteriorly or widening of the glottis transversely. Whether it was a shortening in the anteroposterior dimension or widening in the transverse dimension was not possible because of the limitation of the derivation of the ratio index. Figure 2 illustrates the shape of the glottis of one of the subjects who showed a typical reduced glottal length-to-width ratio. The magnification of the two images was adjusted accordingly based on the reference landmarks using the anterior and posterior ends of the vocal folds identified in the two images. It can be seen that the glottis after singing (fatigue) became relatively wider when compared with that of before singing. Such widening of the glottis could be interpreted as an increase in vibratory amplitude. Gelfer et al,[10] who also found an increased amplitude in fatigued vocal fold vibration, argued that this might have been an adaptation or compensatory effect to the already fatigue condition. Indeed, the compensatory hypothesis has also been proposed by Linville,[23] who found in a study that the amount of glottal closure increased after 15 minutes of loud reading. Both Linville[23] and Gelfer et al[10] contended that their participants generalized the loud reading phonatory mode to the posttask evaluations, resulting in the endoscopic observation of increase in vocal fold contact and greater amplitude of vocal fold excursion.

On the other hand, Stemple et al[2] reported an increase in the presence of incomplete glottal closure after vocal fatigue. They hypothesized that the incomplete glottal closure was because of thyroarytenoid muscle weakness, causing bowing at the edge of the vocal folds. The present study, however, found no change in the glottal closure pattern before and after the singing task.

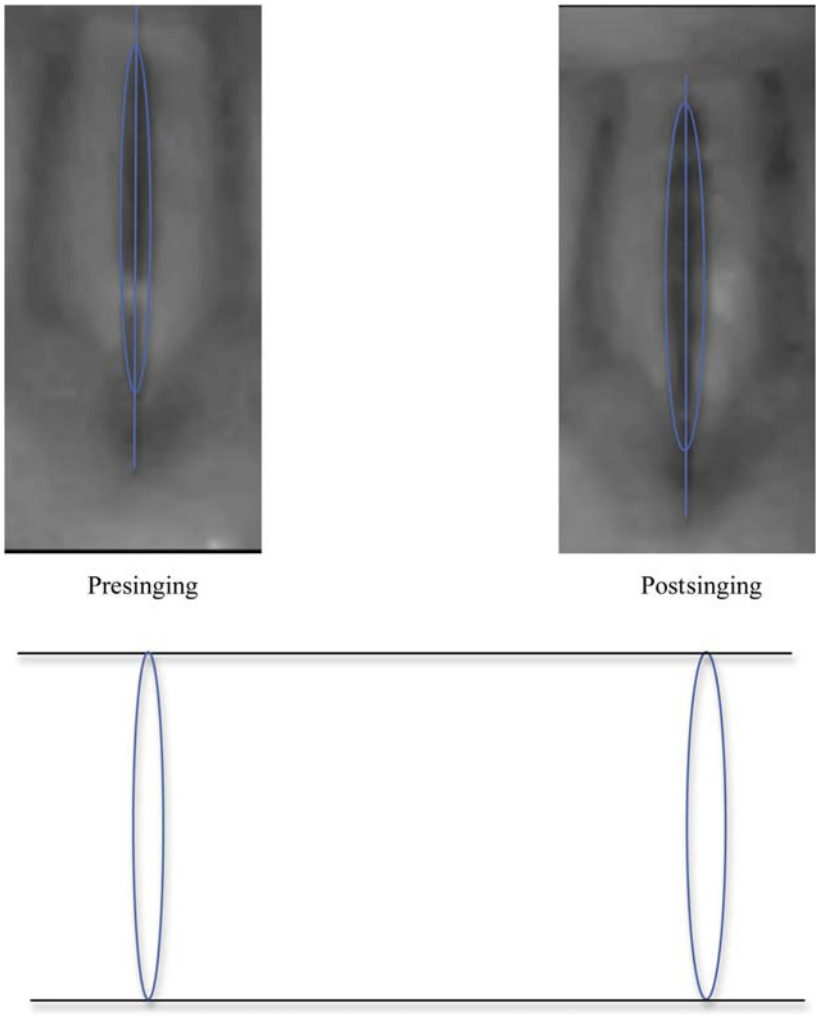


FIGURE 2. An illustration of the shape of glottis in a participant who showed a typical reduced glottal length-to-width ratio after prolonged singing task.

Indeed, 17 of the subjects showed a posterior glottal chink before and after the singing task. One subject showed an incomplete glottal closure before and after the singing task, whereas only two subjects showed complete glottal closure before and after the singing task. Hence, the present study did not have adequate evidence to support the thyroarytenoid weakness hypothesis. It was likely that the participants in the present study might have also adopted a hypertensive mode so that the medial compression increased after the singing task resulted in a decrease in the mean glottal length-to-width ratio index.

It should be noted, however, that not all subjects showed similar changes in the glottal shape. A closer examination of individual data revealed that five (one male and four females) of the 20 subjects showed a reversed pattern, that is, an increase in the glottal length-to-width ratio after singing. Nevertheless, the magnitude of increase (mean, 0.56; range, 0.09–0.90) was relatively smaller than that of the magnitude of reduction (mean, −1.67; range, −0.03 to −4.33). Such small magnitude changes that did not conform to the general trend could have been attributed to individual differences in response to vocal fatigue. This observation, however, appeared to be similar to a single case study reported by Boucher and Ayad.[24] They found reduced lateral cricoarytenoid muscle activities in their subject during vocal fatigue. At the same time, the muscular activities of thyroarytenoid and cricothyroid muscles increased to compensate for the decrease in activity in the lateral cricoarytenoid muscles. The increase in thyroarytenoid muscle activities served to tense and stretch the vocal folds to stabilize the adduction force.

No significant differences were found in the open and speed quotients before and after the singing task. This result contradicted with that reported by Lauri et al,[25] who found a higher speed quotient and a lower closing quotient using electroglottography (EGG) after vocal fatigue. They hypothesized that these changes were because of an increase in adductory force, reflecting a hyperfunctional vocal adjustment. It should be noted that the temporal parameters like open quotient generated from high-speed imaging did not necessarily correlate with that obtained using EGG.[26] There are two possible explanations for such a discrepancy in the findings. First, it might have been because of the technical limitation of the high-speed imaging. The sampling rate of the high-speed imaging in the present study

was only 4000 frames per second (4 kHz), whereas EGGs are analog signals that are usually sampled at above 20 kHz or even up to 44 kHz. Second, it might well be that vocal fatigue does not necessarily cause any changes in the vibratory pattern in the temporal dimension, as measured by open and speed quotients, whereas, the spatial configuration of the glottis could be affected. To conceptualize this, one might like to consider an analogy in the physics of a swinging pendulum. In a free-swinging pendulum, the amplitude of the swings (spatial configuration) gradually reduces, whereas the frequency (temporal dimension) remains the same. Although this second hypothesis seems more plausible, further studies are required to determine whether the temporal dimensions (open and speed quotients) are preserved in vocally fatigued voice.

The results in this study should not be interpreted without some cautions in mind. First, the participants who took part in this study were relatively young (mean age, 21.2 years). Therefore, the physiological reactions to vocal fatigue might not be the same in individuals who are older. Second, the vocal fatigue-inducing task used in the present study was a singing task, which might have produced different vocal fatigue pattern that was induced by an extended talking task. The pitch level during the recordings was controlled in this study, and the voice onset and offset were excluded from the analysis. Therefore, the effect of vocal fatigue on one's pitch, and the pattern of voice onset and offset could not be determined. Further high-speed imaging investigations should take into consideration the voice onset and offset. In addition, high-speed imaging should be accompanied with other instrumental measurements, such as phonation threshold pressure, electroglottograph, and stroboscopy.

Vocal fatigue is often treated as one of the symptoms of voice disorders.[3] Early identification on vocal fatigue, especially for those occupational voice users, who are prone to developing chronic fatigue, would help preventing from the development of a chronic condition that could result in voice disorders. The present study demonstrated the quantitative analysis of high-speed images using the HSVP program. The program was able to detect the difference on vocal vibration pattern between nonfatigue and fatigue vocal conditions. The findings suggested that the participants might have developed a compensatory behavior after vocal fatigue. More studies using multiple measurements will be needed for a better understanding of the physiological changes in vocal fold under vocal fatigue condition.

Methods in studying vocal fatigue

A final concluding remark on the methodology is warranted here for consideration by investigators on vocal fatigue in the future. It should be noted that studies that investigated vocal fatigue have used different procedural methods to induce vocal fatigue. The present study employed amateurs using a singing task. Whether the use of a reading or talking aloud task would make any difference in the effect on vocal fatigue is not known. Furthermore, it would be interesting to determine if amateur and professional singers would have similar or different response to vocal fatigue in terms of the glottal configuration.

Acknowledgment

This study was supported in part by a grant from the Hong Kong Research Grant Council General Research Fund (#HKU757811). The authors also acknowledge Professor William Wei, Dr Elaine Kwong, Dr Cate Madill, and Samantha Wurhurst, who have contributed to the discussion of the research topic in the initial stage of the project planning.

REFERENCES

1. Gotaas C, Starr CD. Vocal fatigue among teachers. *Folia Phoniatr (Basel)*. 1993;45:120–129.
2. Stemple J, Stanley J, Lee L. Objective measures of voice production in normal subjects following prolonged voice use. *J Voice*. 1995;9:127–133.
3. Colton RH, Casper JK, Leonard RL. *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2006.
4. Solomon NP, DiMattia MS. Effects of a vocally fatiguing task and systemic hydration on phonation threshold pressure. *J Voice*. 2000;14:341–362.
5. Solomon NP, Glaze LE, Arnold RR, van Mersbergen M. Effects of a vocally fatiguing task and systemic hydration on men's voice. *J Voice*. 2003;17:31–46.
6. Welham NV, Maclagan MA. Vocal fatigue: current knowledge and future directions. *J Voice*. 2003;17:21–30.
7. Gelfer MP, Andrews ML, Schmidt CP. Effects of prolonged loud reading on selected measures of vocal function in trained and untrained singers. *J Voice*. 1991;5:158–167.
8. Eustace CS, Stemple JC, Lee L. Objective measures of voice production in patients complaining of laryngeal fatigue. *J Voice*. 1996;10:146–154.
9. Mann EA, McClean MD, Gurevich-Uvena J, Barkmeier J, McKenzie-Garner P, Paffrath J, Patow C. The effects of excessive vocalization on acoustic and videostroboscopic measures of vocal fold condition. *J Voice*. 1999;13:294–302.
10. Gelfer MP, Andrews ML, Schmidt CP. Documenting laryngeal change following prolonged loud reading. A videostroboscopic study. *J Voice*. 1996;10:368–377.
11. Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. *Folia Phoniatr Logop*. 2008;60:33–44.
12. Patel R, Dailey S, Bless D. Comparison of high-speed digital imaging with stroboscopy for laryngeal of glottal disorders. *Ann Otol Rhinol Laryngol*. 2008;117:413–424.
13. Larsson H, Hertegard S, Lindestad P-A, Hammarberg B. Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. *Laryngoscope*. 2000;110:2117–2122.
14. Inwald EC, Dollonger M, Schuster M, Eysholdt U, Bohr C. Multiparametric analysis of vocal fold vibrations in healthy and disordered voices in high-speed imaging. *J Voice*. 2011;25:576–590.
15. Yiu EM-L, Kong J, Fong R, Chan KMK. A preliminary study of a quantitative analysis method for high speed laryngoscopic images. *Int J Speech Lang Pathol*. 2010;12:520–528.
16. Hess M, Gross M. High-speed imaging of vocal fold vibrations and larynx movements within vocalizations of different vowels. *Ann Otol Rhinol Laryngol*. 1996;105:975–981.
17. Mehta DD, Deliyski DD, Quatieri TF, Hillman RE. Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. *J Speech Lang Hear Res*. 2011;54:47–54.
18. Koster O, Marx B, Gemmar P, Hess MM, Kunzel HJ. Qualitative and quantitative analysis of voice onset by means of a multidimensional voice analysis system (MVAS) using high-speed imaging. *J Voice*. 1999;13:355–374.
19. Kong JP, Yiu EM-L. Quantitative analysis of high-speed laryngoscopic images. In: Ma EP-M, Yiu EM-L, eds. *Handbook of Voice Assessments*. San Diego, CA: Plural Publishing; 2011.

20. Yiu EM-L, Chan RMM. Effect of hydration and vocal rest on the vocal fatigue in amateur karaoke singers. *J Voice*. 2003;17:216–227.
21. Boersma P, Weenink D. Praat: doing phonetics by computer [computer program]. Version 5.3.39. 2013. Available at: http://www.praat.org/.
22. Hunter EJ, Titze IR. Quantifying vocal fatigue recovery: dynamic vocal recovery trajectories after a vocal loading exercise. *Ann Otol Rhinol Laryngol*. 2009;118:449–460.
23. Linville SE. Changes in glottal configuration in women after loud talking. *J Voice*. 1995;9:57–65.
24. Boucher VJ, Ayad T. Physiological attributes of vocal fatigue and their acoustic effects: a synthesis of findings for a criterion-based prevention of acquired voice disorders. *J Voice*. 2010;24:324–336.
25. Lauri ER, Alku P, Vilkman E, Sala E, Sihvo M. Effects of prolonged oral reading on time-based glottal flow waveform parameters with special reference to gender differences. *Folia Phoniatr Logop*. 1997;49:234–246.
26. Echternach M, Dippold S, Sundberg J, Arndt S, Zander MF, Richter B. High-speed imaging and electroglottography measurements of the open quotient in untrained male voices' register transitions. *J Voice*. 2010;24:644–650.

本期执行编辑：张锐锋
北京大学中文系语言学实验室
电话：86-10-62753016
邮箱：ell2pku.edu.cn
网址:www.phonetics.ac.cn