连续话语语料库的语音切分和标记

陈 肖 霞

提要 对连续话语语料库进行切分和标记是一项新的课题,它对语料库的充分利用有重要作用,如何做好这项工作是一个值得探讨的问题。本文通过对一个语料库的切分和标记,得出了一些初步看法和认识,在这里跟同行们切磋,以使这项工作做得更完善。

A segmentation and labeling work based on continuous speech database

Chen Xiaoxia

Abstract segmentation and labeling for continuous speech database is important to the better use of database. The question of how to improve segmentation and labeling calls for further discussion. This paper shows the labeling and segmentation work we have done in standard Chinese. We have concluded with some labeling rules and segmentation units according to the database. We hope to got sayyestions and to do the work further.

一、前言

建立包含比较全面的语音现象的连续话语语料库,对进行识别、合成有重要作用,对进行大规模语音研究,也非常重要。语料库的建立,除在语料选取、录音等方面做大量工作外,对已建成的语料库做语音的标记,也是语料库建设的一个重要步骤。目前,国际上已建成的语料库,如美国的 TIMIT,除包含语料和录音外,还有语音标记。著名语音学家 KEATING 等利用 CSL 语音分析仪器,对 TIMIT 进行标记,从时间点上确定音段的实际发音和时长,将连续话语中语音的变化体现出来[4]。德国的德语声学资料库,是德语口语资料库,其中也带有语音学标音[1]。这样的语料库,由于语料丰富,又带有语音标记,可以进行统计并做声学分析,能够更定量地反映出语音的变化,从而将语音研究带入新的发展阶段。对语音现象进行标记,也为言语识别建立基本单元和进行测试提供了可资借鉴的语料。国内目前有声学所做的普通话单音节的语音标记,主要是利用声门波来判断清浊,进行声、韵切分。而对连续话语做的不多。这一方面是因为连续话语语料库是近年刚刚建立,另一方面也是对连续话语的研究刚刚开始。过去,语言所对单音节、两音节做的研究,为连续话语的研究提供了坚实的基础。自然连续话语由于受到语法、韵律的影响,其语音变化与单音节相比,有很大不同语音的变化往往更

大,例如出现语音上的弱化、浊化、甚至脱落等音段的变化。因此,对连续话语进行标记更困难。

由中国科技大学、声学所、语言所等参加建立的国家“863”语音识别语料库,是一个考虑各种语音现象的连续话语语料库,其中第一阶段共有1560个句子。这些句子分为三组,由三位发音人完成。对这些语料进行标记,一方面为识别建立基本单元提供了基本的语音信息,同时也为语音研究提供了新的研究资料。在介绍具体做法之前,先谈一谈对连续话语的语音认识。

二、对本语料库语料的语音认识

连续话语并不是单个音节的简单组合,由于协同发音、语义、韵律等的影响,连续话语与单念的音节相比,变化较大。因此,作为切分和标记的前提,首先要对语料库的语音情况有个全面的认识。

目前的连续话语语料库是说话人对照文本念出来的,速度是每秒3-4个音节,比较慢。每句最长不超过20个音节,大都是平叙句。发音人基本可以做到发音清楚。但是,即使这样,与单音节相比,语音变化也很大。同时,由于发音人的说话风格不同,也表现出一些个人差异,但基本的语音现象还是一致的。共性的东西更多些。

(一)共性特征

1. 从三位发音人的语料看,发音比较清晰,音节和音节间的分界大部分比较清楚,少数的音节间的分界不太明显,主要发生在后音节声母为零声母和浊声母的情况下。

2. 声母的浊化和送气浊化:有些塞音声母浊化,如b,d,g。塞音送气部分常常浊化,如p,t,k。见图二。“今天”中“天”的声母,送气段出现浊横杠。这是声母浊化的典型表现。

3. 唇音对舌尖鼻韵尾音的影响:后音节声母为唇音,前音节韵尾为舌尖鼻音时,舌尖鼻音几乎全部变为双唇鼻音。这种现象见图三“贸易联盟”中,“联”的韵尾受“盟”的声母影响逐渐变为唇鼻音,表现为舌尖鼻韵尾与后接双唇鼻声母在频谱上的相似性。

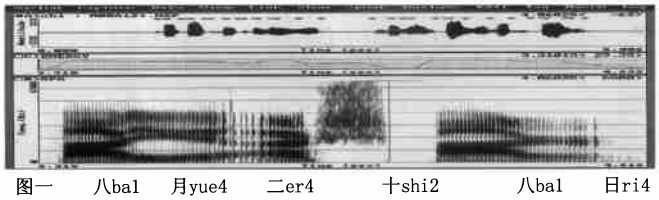
4. 闭塞段丢失:在没有明显停顿的地方,塞音尤其是塞擦音前面的持阻段常常丢失,其表现是声学上的静止段不存在,一种是直接出现冲直条,另一种是出现嗓音,对后一种情况,按照过去的研究,是VOT<0,应归于声母部分,本次切分单独区分为嗓音和声母辅音,希望今后对此做进一步的研究。在标记时,用sild表示在闭塞段出现的嗓音,dv表示冲直条部分出现的浊化嗓音。

(二)个性特征

通过对几位发音人的语料分析,发现在这些发音人中,个体之间语音现象的差异是存在的,归纳起来,主要有以下几种:

1. 个别音的丢失:例如,第一位发音人发元音(i2),就常常丢失。如图一“八月二十八日”中“十”的韵母丢失。而另外两位发音人没有这种现象。

2. 停顿的位置和次数:三位发音人停顿的时间和具体位置略有不同,有时与语法上不一致。



图一　八ba1　月yue4　二er4　十shi2　八ba1　日ri4

三、语音切分和标记的原则及标记形式

(一)切分的原则

1. 切分包括五部分:声母段　过渡段　韵母段　闭塞段　停顿段

(1)声母段:在塞音声母、塞擦音声母中,不包括闭塞段。

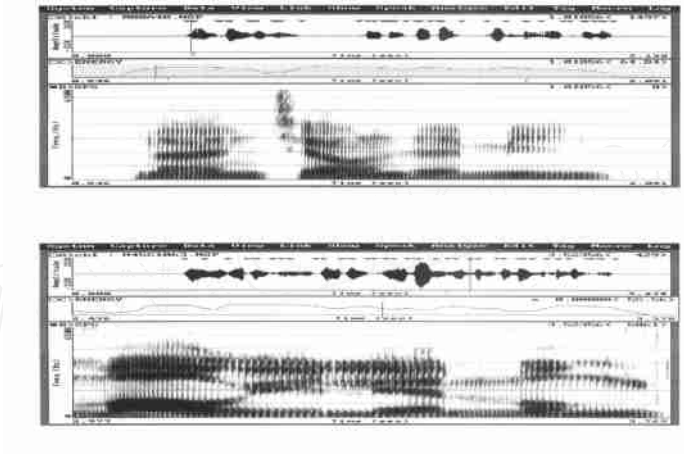
(2)过渡段: 存在于声韵之间,但本次切分只将韵母部分的过渡切出。这主要是基于这样的考虑,声母部分的过渡与韵母部分的过渡应区别对待。而声母一般较短,其过渡更短,不如韵母部分的过渡相对明显些,这需要进一步研究。

(3)韵母段:过渡段切出后剩下的就是韵母部分。包括单元音韵母、复合元音韵母和鼻韵母。

(4)闭塞段:指塞音、塞擦音前的无声段。

(5)停顿段:特指发音人在发音时出现的语音中断,表现为无声段,它与闭塞段的差别主要在时间的长短和音段的协同发音上。

2. 切分方法:利用 KAY CSL 4300B 语音分析仪器,结合宽带语图、波形及听辨来进行,兹以图二为例略加说明。



图二是"南疆人民",在上面的波形中,倒三角是波形切分标记的时间点,其中带有说明。中间是振幅,最下面是宽带语图。根据宽带语图、波形和振幅,可将各段切出。声母的切分根据声韵之间频谱的明显变化,过渡段根据元音共振峰转折点确定起止点过渡。声母切分中,比较困难的是浊擦音如"人"中的声母,从频谱可以看出,由于带有嗓音,与前面音节韵母的共振峰相连接,没有明显的分界,但在高频部分其能量很弱,所以可以将这个信息作为分界的标志。对零声母的情况,也是这样处理,如图三的"贸易联盟"中,"易"的切分,也是根据其能量和共振峰的改变来界定的。而"联盟"中,舌尖鼻韵尾与后接双唇鼻声母在频谱上很相似,两者之间除鼻音峰有动程外,作为声母的鼻音,在高频区出现明显的共振峰,在划分时,根据这些不同将鼻韵尾与鼻声母分开。

(二)标记的原则

1. 原则一:可还原性。在声学平面标记语音现象,要考虑语音的底层形式,使人能清楚地看出底层与表层的语音的变化和对应。使之可以还原。例如,闭塞段的丢失和元音或辅音的丢失,就需要用 < > 符号表示,而不是置之不理,因为这种丢失与在底层本来就不存在的情况是不同的。起码应对这种情况做深入研究,才能决定。同样,舌尖鼻音受唇音影响变为双唇鼻

音和舌尖鼻音受舌根音影响变为舌根鼻音的情况,也应该标记出来。更明显的例子是 r 的丢失,比如"例如"的"如",在声母丢失后,看起来跟"里屋"的"屋"相同了,但实际上是不同的。所以这个声母的丢失应该设法标记出来。

2. 原则二:一致性。由于在连续话语中语音的变化复杂,对同类现象的标记要保持前后一致。举例来说,sil 是指塞音或塞擦音前的闭塞,在连续话语中,有时丢掉,在标记时,以 < sil > 表示本来存在而在语流中丢掉的部分,仍以 sil 表示未丢掉的情况,这样既表示了它跟 sil 的联系,又表示了它跟 sil 未丢掉的区别。

3. 原则三:现实性。以实际发音标记。发音人发音时有与文本不符时,以实际发音来标。如"这",应为 zhe,实际发音时常常是 zhei,则以 zhei 标记。类似的音还有"谁"等。对声母浊化的标记等也贯彻这个原则。

(三)标记形式和说明

1. 标记符号:

* 语音信号的起始

& 语音信号的结束

零声母

< > 语音段的丢失

v 鱼类的韵母

i1 资类的韵母

i2 知类的韵母

2. 标记说明:

(1) 除零声母前面加 # 标记外,其余声母分别用拼音字母标记。

(2) 声母的浊化现象是在相应的声母后加 v,区别于非浊化。

(3) 声母段在标记的最后写 1。

(4) 过渡段用声韵母联接,在标记的最后写(2)。

(5) 韵母段用拼音字母表示,在标记的最后写 3。

(6) 声调用 1、2、3、4、0 表示四声和轻声。

(7) 声调标在韵母上,韵母之后为调号。标出的变调包括上上相连变调,"一"、"不"变调等,由于语流中前后语音影响较复杂,此处标的声调仅供参考。

(8) 塞音、塞擦音前的闭塞段标为 sil。明显的停顿标为 silgap。

(9) 在标出以上各段的同时,给出每段的前后语境。如下例中"ang4(sh,h)3 ?,括号里的"sh,h"分别代表"ang"前面和后面的声母辅音,4 代表声调,3 表示韵母段。

例如:shang4 hai3 de0 ……kun4 nan0(上海的……困难)

1. *sh(*,ang)1 第一音节的声母段

2. sh-ang(2) 第一音节的过渡段

3. ang4(sh,h)3 第一音节的韵母段

4. h(ang,ai)1 第二音节的声母段

5. h-ai(2) 第二音节的过渡段

6. ai3(h,sil)3 第二音节的韵母段

7. sil(ai,d) 第三音节的声母的闭塞段

8. d(sil ,e)1　第三音节的声母段
9. d-e(2)　第三音节的过渡段
10. e0(d ,sil)　第三音节的韵母段
……
n-an(2)　最后音节的过渡段
an0(n , &)3　最后音节的韵母段
an &　结束

四、讨论

对语料库进行语音标记,是建立完善的语料库的一项重要工作,也是当今国际趋势。目前英语、德语等都建立了带有语音标记的语料库。但是对汉语连续话语进行标记,这还是一次尝试,通过这次标记,不但对切分标记本身提高了认识,而且还从中发现了许多语音现象,并对许多语音现象进行了统计,以后将陆续给出。对研究而言,这种带有标记的语料库是十分有用的。一方面,其中的语料是一种相对自然的话语,它们包含了各种语音单元的组合,更接近自然语言,有利于研究者对自然语音的把握,同时由于带有切分和标记,更有利于定量研究,使语音研究更科学化,对言语处理更有参考价值。不过,在实际操作中,也发现了一些问题,需要在今后不断改进。

(一)目前的切分不能完全反映语音的实际变化,宜更细一些,利用国际音标来做出标记,以便反映出语音的细微变化。例如:声母送气段,常常带有过渡,并伴随浊化,应该单独标记出来。弱读引起的音色央化、时长改变等,也应设法标出。

(二)停顿与闭塞的差别,有时不易区分。需要结合听辨和韵律结构进行区分。应该带有韵律标记。

(三)目前的切分是纯手工来做的,下面应考虑手工与机器自动切分的结合,避免手工操作的人为误差和机器切分的系统误差,这需要一个好的识别系统的支持。

致谢:本工作是国家"863"对连续话语语音识别语料库的设计和语音标记"课题的任务。标记形式和数据的得出由祖漪清提供;标记原则的确定等是与林茂灿先生等语音室同事讨论的结果,其中标记不当之处应由本人负责;曹剑芬先生为本文提出了中肯的意见。在此一并致谢。

参考文献

[1]曹剑芬译《互补音系学——对一个声学资料库进行标记的理论框架》,《国外语言学》1995 年第 1 期。

[2]马大猷、沈豪等《声学手册》,科学出版社,1983。

[3]朱维彬、张家录《汉语语音数据库的标记》,《第四届全国人机语音通讯学术会议》,1996。

[4] P. Keating, B. Blankenship, D. Byrd, E. Flemming, Y. Todaka *Phonetic analyses of the TIMIT corpus of American English*, UCLA Num. 81, 1992.

[5] P. Keating, Peggy MacEachern, Aaron Shryock, Sylvia Dominguez A manual for phonetic *transcription segmentation and labeling of words in spontaneous speech*, UCLA Num. 88, 1994.

[6] P. Keating, Dani Byrd, Edward Flemming, Yuichi Todaka *Phonetic analyses of word and segnent variation using the TIMIT corpus of Anerican English*, Speech Communication Vol. 14 pp. 131～142, 1994.

[7] Zu, Yingqing Sentence design for speech synthesis and speech recognition database by phonetic rules, EuroSpeech '97 Vol. 2 pp. 743～746, 1997.

(陈肖霞　中国社会科学院语言研究所,邮编:100732)