

文章编号:1001-9081(2009)03-0761-03

基于语音识别技术的英语口语教学系统

赵 博, 檀晓红

(上海交通大学 计算机科学与工程系, 上海 200240)

(bozhao@live.com)

摘 要:许多计算机辅助英语学习的应用欠缺口语学习的评估和反馈。描述了一个采用语音识别技术的英语口语学习系统。除了通常的发音评分外,还提供基于音素关联和音素识别的错误检测功能。结合纠正知识库的改进建议和韵律修正语音,可以及时地给学习者以帮助。实验结果表明,能够纠正有一定基础学习者的多数非故意错误。

关键词:语音识别;口语评分;错误检测;韵律修正

中图分类号: TP391.42 **文献标志码:** A

Oral English training system based on speech recognition technology

ZHAO Bo, TAN Xiao-hong

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Many applications of computer-aided language learning system lack of evaluation and feedback of learning speech. An oral English training system was described. Besides normal speech evaluation, error detection based on phoneme alignment and phoneme recognition was provided. Recommendations for improvement queried from correction knowledge base and prosody modified speech offer learners a timely help. The experimental results show that most mistakes made unconsciously by high-level 2nd language speakers can be corrected by this system.

Key words: speech recognition; pronunciation evaluation; error detection; prosody modification

0 引言

随着中国与世界的联系越来越紧密,尤其是北京奥运会的举行和上海世博会的即将召开,许多人被推到了英语学习的浪尖之上。很多学习者没有时间和机会接受全职的系统英语训练。而学习时间和地点不受限制的计算机辅助语言学习(Computer-Aided Language Learning, CALL),为他们提供了一种新的选择。最初的辅助学习主要应用于文字能力和理解能力的训练。随着语音技术的发展,越来越多的 CALL 研究和应用开始关注发音学习^[1]。

当前的计算机辅助语言学习系统,大多侧重单词、语法的学习。仅有的一些口语学习软件,功能比较单一,只能给学习者的发音一个整体的评分。然而自学者因为自身水平限制,很难完全靠自己发现错误、纠正不正确的发音。软件发音错误矫正的功能,可以帮助学习者及时改正发音错误,避免错误经多次重复而成为习惯。

1 系统概览

基于语音识别技术的英语口语教学系统主要利用语音识别技术计算学习者读音的评分,检测其中的发音错误。系统使用到了 CMU 开源项 SPHINX3.5 的音素识别器 allphone 和关联器 align 两个部分。

学习者发音的评分,必须有一个参考标准。训练好的语音模型可以用作评分标准。在模型完整的情况下,可以为许多句子评分。评分的主要模式,是将用户的语音和参考模型做比对。

1.1 系统框架

图 1 表明了使用 sphinx 音素关联和音素识别的口语评分工

具中的语音处理。获取单词音素的时间关联,一般有两种方法:强制关联和识别。本系统中学习者跟读例句,其音素文本是预知的,因此采用强制关联的方法。系统的主要模块有音素识别、音素关联、错误检测和韵律修正四个部分,向学习者提供的反馈结果包含音素评分、纠正意见和韵律修正的反馈语音。

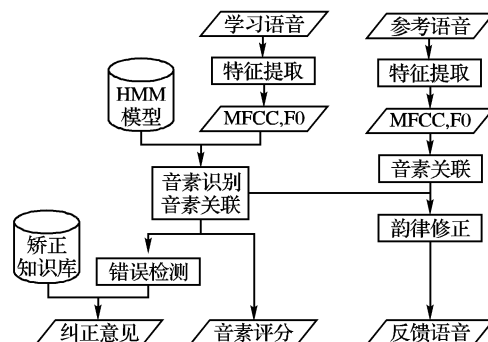


图 1 英语口语反馈纠正系统框架

系统预先对外教朗读的参考语音做自动标注。特征提取模块计算语音的 MFCC 参数以及基于短时幅度差函数的基频(F0)特征。由音素关联模块计算得到的时间关联和 F0 信息被保存下来,用于修改学习语音的韵律。

学习者朗读系统指定的英文文本,所采集学习语音的 MFCC 参数和 F0 被分别用于基于 HMM 的语音识别和韵律修正。基于 psola 的韵律修正模块合成具有参考语音韵律和学习者音色的反馈语音,韵律修正语音能够引导学习者模仿参考语音的韵律,当其习惯英语的韵律规律之后,就能在自己的发音中运用自如。

错误检测结果最终转化为纠正知识库中的纠正意见,这比仅仅给出一个低分或者指出发音有误更加有效。

收稿日期:2008-09-09;修回日期:2008-11-14。 基金项目:国家“十一五”支撑项目(2007BAH09B05)。

作者简介:赵博(1984-),男,河南周口人,硕士研究生、主要研究方向:语音识别; 檀晓红(1971-),女,安徽池州人,讲师,博士研究生,主要研究方向:模式识别、数据挖掘。

1.2 语料库

语料库内容的选取,以帮助学习者提高口语水平为目标。条件允许的话,可以分别建立参考语料库和识别语料库。例如在小语料库的情况下,建立标准口音和学习者口音 2 个语料库:学习者口音语料库应能代表用户发音特点,用于训练音素关联所用的语音模型;标准口音语料库用于评估学习者口音。参考语料库中的语音是实际上的评价标准,因此语料的内容选择和录音都必须与期望相同。比如训练说宝堂(Saybot)语音模型的语料库是由美音的说话人录制,所以英国口音的学习者可能会因此而得分不高。语料库的内容必须涵盖系统预定识别的音素和句子。而且作为学习资料,语料的选取还应考虑教学目的,涵盖基本的英语发音特点和易错点。

2 学习语音的评分

英语语音评分是整个学习系统的核心。目前绝大多数类似系统的评分都采用与参考语音模型的相似性为度量的方法。

这种评分方法应用了基于隐马尔可夫模型(Hidden Markov Model, HMM)的自动语音识别技术,一般选取音素为识别单元。在音素关联的 Viterbi 解码中,参考说话人语音训练的 HMM 被用作衡量标准,与其相似性为学习语音评分。这种相似性,与音素的发音质量呈正相关性。假设第 i 个音素关联为音素 q ,其时间段对应起点为 τ_i 。 t 时刻观测到的特征向量为 o_t ,对应的 HMM 状态为 s_t 。则基于 HMM 似然度的评分^[2]:

$$M_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg[P(s_t | s_{t-1})P(o_t | s_t)] \quad (1)$$

基于 HMM 后验概率的评分:

$$M_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg P(q | o_t) = \sum_{t=\tau_i}^{\tau_{i+1}-1} \lg \frac{P(o_t | q)P(q)}{\sum_{q_j \in Q} P(o_t | q_j)P(q_j)} \quad (2)$$

其中 Q 为所有音素集合。

通常,非英语母语语音的音素关联似然度评分会比英语母语语音低,这种评分可以反映出非英语母语者发音存在的错误^[3-4]。

在得到总的得分之后,可以以一种便于学习者理解和接受的方式展示出来。学习者最希望看到的就是在根据反馈重复跟读后,分数有所提高。而且,直观分数可能因不够精确而与发音逐步改善的事实相反,影响学习者信心,所以采用分级的评分减少这类现象。对于不同的音素,分别预先设定 2 个阈值将得分划分为 3 个区段,视学习者得分分别给予 3 类评价:积极、中性和消极。对一个验证语料库中的句子做专家人工评分和自动似然度评分,对比两者的结果可以获得这 2 个阈值。

3 发音错误矫正

音素发音的矫正包含错误检测和纠正反馈 2 部分。而对于发音的韵律矫正,因为不易对韵律的组成元素建模,所以并没有直接指出韵律中的错误和改进意见,而是将学习语音的韵律修正为参考语音韵律,播放给学习者。

3.1 音素发音错误检测

错误检测依据音素识别和音素关联的结果,依赖学习语音的文本。

发音错误的分类有多种方法。这里依据音素识别器的识别结果,可以将音素朗读错误简单地分为 3 类。1)漏读音素。音素 Q_i 关联段的音素识别结果中不存在期望音素 Q_i ,而是为无声、 Q_{i-1} 或者 Q_{i+1} 。2)误读音素。音素 Q_i 关联段的音素

识别结果中不存在期望音素 Q_i ,且不是漏读音素。3)添加音素。音素关联段的识别结果中存在多余的音素。

基于一个强制关联—音素识别的 2 遍过程,可以设计检测上述 3 类错误的方法。一个例句“Good to see you again”的关联和识别结果(前三个单词部分)及错误检测如表 1 和 2 所示。表中开始和结束时间用短时分析窗序号表示,窗长度 25 ms,步长 10 ms。其中 Viterbi 强制关联用到了 sphinx3_align,输入为语音特征及其文本。音素识别用 sphinx3_allphone 完成。

表 1 音素关联结果

时间	音素							
	g	u	d	SIL	t	u	s	i:
开始时间	22	28	35	38	46	51	65	76
结束时间	27	34	37	45	50	64	75	89

表 2 音素识别结果

音素	音素									
	g	a	d	SIL	t	a	n	t	θ	i:
开始时间	21	29	32	37	45	52	59	64	69	74
结束时间	28	31	36	44	51	58	63	68	73	92

其中误读错误有:“u”误读成了“a”2 次,“s”误读成了“θ”1 次;从正确音素“tu”到“tən”,存在误读“a”和插入“n”;从“s i:”到“tθi:”存在插入“t”和误读“θ”。

3.2 纠正反馈

纠正反馈是系统提供的核心功能,包括音素朗读错误的纠正和韵律的反馈。

3.2.1 音素发音的纠正反馈

依据错误检测,可以得到音素发音错误的上下文。CMU Phone Set 有 39 个音素,在此情况下最多会有 392 种发音不准错误。而对于漏读音素和添加音素,在考虑前后各一个音素的情况下,共有 393 种组合。但实际上并不是所有的组合都会出现。在实验结果部分将给出错误的统计。

对于每种发音错误的组合,从纠正知识库中查找对应的反馈意见,指正学习者朗读口舌位置等的不当,示以口型动画。

3.2.2 句子韵律的纠正反馈

一种直观的矫正韵律的矫正方法,就是将学习者发音音频的韵律修改为参考语音的韵律,播放出来给学习者以直观感受。使得学习者可以听到用自己嗓音说出来的正确韵律的英语。这是学习者“自己的”最佳读音,对自己是一种鼓励,也更容易与原始读音比较而发现问题。

文献[5]提到了一种方法,用于对一个单词的韵律加以修改,涉及到基频、时长等韵律特征。但参考语音和学习语音都需要人工标注。对于自动修改算法,可以将时间关联的结果当作一种自动标注。修改过程包括关联标注的修正和语音合成 2 部分。

1) 关联标注的修正。

在人工对参考语音和学习语音都做了精确标注的情况下,基音同步叠加结果的基频与参考语音十分相似;但音素时间关联结果的准确率达不到人工标注的水平,标注的偏差将导致合成语音与期望相差很大,听起来不够自然。尤其当学习语音与关联文本有出入时,错误更明显。所以在作为标注之前对关联结果做适当的修正,是十分必要的。

学习语音漏掉音素 在 3.1 节描述的检测算法可以有效

地检测出学习者漏掉一个音素的情况。此时,音素关联的文本中此音素将被删去而重新做一次关联,音素关联的结果得到修正。

去除噪声段 相比较单词的韵律修正,句子的韵律修正面临新的问题:英语初学者语速较慢,往往在单词之间插入空白。但基于音素文本的音素关联无法预知这种情况,于是会将空白语音关联到某个音素段尾部,这样关联的结果中可能包含空白语音的噪声段。因此用端点检测和语音检测来消除错分噪声段带来的影响,修正关联结果是十分必要的。

添加音素错误 学习语音中添加的音素,无法在参考语音中找到对应。最终在语音合成步骤中,这些语音段将会被丢弃。

2) 语音合成。

语音合成领域中,基音同步叠加算法^[6]被广泛用于韵律修改。这种方法计算效率较高,通过时域中对波形的拼接,同时修改基音和时长,可尽量保持学习者的音色信息。为了保持学习语音原有的基频水平,不能直接将参考语音的基频作为合成语音的基频,需要对其做一个缩放。基频缩放 P_{scale} 为参考语音和学习语音平均基频的商,在每一个已关联的音素段时长缩放 T_{scale} 为参考语音段和学习语音段时长的商。音素段内每 10 ms 利用参考语音的基频 $F0_{ref}$,计算新的 $F0 = F0_{ref}/P_{scale}$,然后以学习语音和重新计算的时长和基频合成韵律修正语音。

当在合成语音中去除音频段,例如学习者误插入的空白段时,需要通过拼接来连接前后 2 个语音单元。由于前后拼接单元端点都是衰减的,存在基频和幅度包络的跳变,应当选择合适的拼接点避免基频曲线的中断跳变。

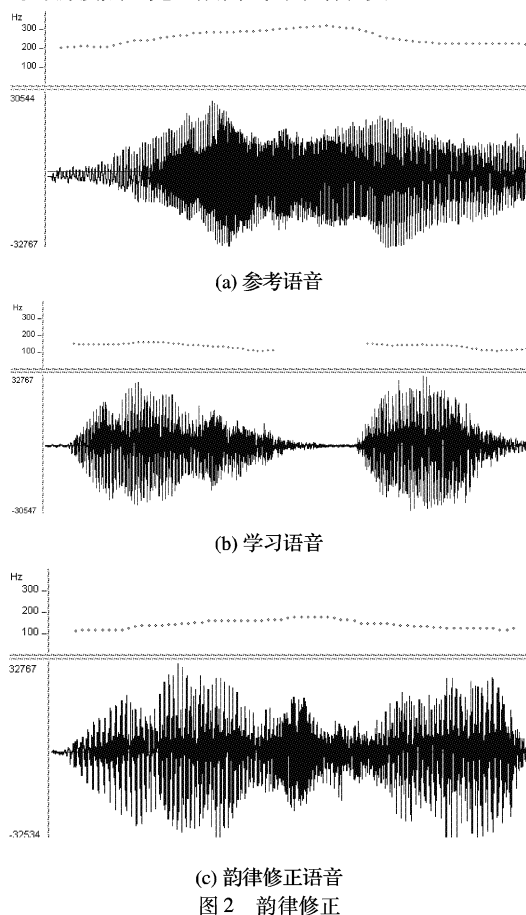


图2中(a~c)分别是朗读“Where are”的女声参考语音、男声学习语音和韵律修正语音的波形基频图。基频采用短时

幅度差(ADMF)方法计算。在合成时学习语音“where”和“are”之间的空白语音段被丢弃,另外并没有做幅度包络的修改,因为对听感的影响不大。

语音信号是由声道滤波器调制激励信号而得。上述拼接针对基频,而不是声道谱,会出现基频连续而声道谱不连续。合成语音在拼接点听起来过渡不很自然,似是在不同的语境。但合成语音的时长、基频等韵律元素听起来完全模仿参考语音,并保留了学习语音的音色。

4 错误矫正实验

4.1 数据收集

用于实验的语音文本由 100 句常用英语组成,分别由 9 位非母语发音者朗读。发音者都具有大学英语水平,并在朗读时尽量不犯错误;参考标准语音采用一位美音母语发音者的读音。声学模型使用开源的 WSJ1,此模型是听写语音,发音适合作为评分依据。

4.2 实验结果

错误检测对于误读或漏读音素,具有较好的效果。但对于插入音素,检测错误率较高。在此验证语料库的错误检测结果中,共出现误读音素错误 712 种,漏读音素错误 653 种。最不易读错和最易读错音素的统计结果,如表 3 所示。每行列出一个被误读的音素和最易被误读为哪三个音素。

表3 最不易读错和最易读错的音素

被误读音素	正确率/%	误读为音素(误读率%)
k	90.34	t(2.36), p(1.63), g(1.27)
i:	88.58	l(2.75), ʌ(1.47), eɪ(1.28)
n	81.49	ŋ(2.77), d(1.96), m(1.73)
tʃ	38.55	t(20.48), dʒ(12.04), d(6.02)
ʃ	34.85	z(12.44), s(12.03), dʒ(9.12)
ɔɪ	31.69	əʊ(23.07), ʌ(11.53), l(7.69)
aʊ	30.34	ɔ:(13.79), ʌ(13.79), ʌ(7.58)

从表 3 统计结果中可以看出,浊辅音和复合元音容易被读错。这一方面是因为汉语中浊音清化的发音习惯,另一方面是因为浊辅音本身的时长一般较短。因此在给出纠正反馈时,对不同的音素采用不同的容错阈值是必要的。

表4 最易略过音素

被漏读音素	漏读率/%	前导/后随音素(漏读率%)
d	5.75	n/r(1.15), u/j(1.08)
aʊ	4.86	h/z(2.07), h/ɛ(0.94)
v	4.74	ð/ʌ(0.88), əʊ/m(0.35)
ɛ	4.58	ð/r(0.72), w/r(0.57)

表 4 中每行列出一个被漏读的音素和最易在哪 2 种前导/后随音素的组合情况下被漏读。所统计的最易略过音素,并不一定是因错误朗读而被记录,如“d”就是因为有几处读成不完全爆破。因此在错误检测中允许这类发音规则,可以增强其可信度。

5 结语

利用音素关联和音素识别结果的错误检测,可以有效检出误读音素和漏读音素,但对于插入音素的检测效果较差。人对发音的评价是直观而迅速的,然而对于计算机,如何将多方面的特征综合,最后给出一个符合专家意见的反馈,仍是一个难题。

(下转第 773 页)

$$\mu_y^{(1)} = (l'_1 \mu^{(1)}, l'_2 \mu^{(1)})' = (3.4576 - 3.2829)'$$

$$\mu_y^{(2)} = (l'_1 \mu^{(2)}, l'_2 \mu^{(2)})' = (2.3404 - 5.1489)'$$

$$\mu_y^{(3)} = (l'_1 \mu^{(3)}, l'_2 \mu^{(3)})' = (1.3036 - 3.2349)'$$

$$\mu_y^{(4)} = (l'_1 \mu^{(4)}, l'_2 \mu^{(4)})' = (2.9720 - 3.8867)'$$

5) 计算未知工件 X_0 的判别函数向量

$$y_0 = (2.8216 - 3.8767)'$$

6) 由式(6)计算 y_0 分别到 $\mu_y^{(1)}, \dots, \mu_y^{(4)}$ 的 Euclidean 距离平方:

$$D^2(y_0, \mu_y^{(1)}) = 0.7571, D^2(y_0, \mu_y^{(2)}) = 1.8501$$

$$D^2(y_0, \mu_y^{(3)}) = 2.7161, D^2(y_0, \mu_y^{(4)}) = 0.0227$$

7) 根据式(10)判定该未知工件为第4类工件。这与文献[8,12]的识别结果一致。

很显然,若不采用主成分法对判别函数进行融合,当目标个数较多时,则上述步骤4)~6)的工作量大大加重。将上述 Euclidean 距离平方归一化,得到归一化的 Euclidean 距离平方分别为 0.1416、0.3461、0.5081、0.0042。

若将文献[8]可变模糊法的综合隶属度以及本文归一化的 Euclidean 距离平方加以比较,则尽管这两种方法的识别结果都为第4类工件,但是它们对各类目标识别的区分程度却不同。表3反映了这两种方法融合结果的差异。表中的差距是指对第4类工件与对其他各类工件的 Euclidean 距离平方(综合隶属度)之差。

表3 本文和文献[8]融合结果的比较

工件	本文方法		可变模糊法	
	归一化的 Euclidean 平方	差距	综合隶属度	差距
1	0.1416	0.1374	0.366	0.424
2	0.3461	0.3419	0.581	0.209
3	0.5081	0.5039	0.556	0.234
4	0.0042		0.790	

若将表3中的差距求和,则可得本文方法的识别差距总和为 0.9832,大于可变模糊法的识别差距总和 0.867。显然,差距越大,说明对目标识别的区分程度越高,可信度也越高。同时,本文方法对目标类别2、3的识别差距为 0.3419 和 0.5039,分别大于可变模糊法相应的差距 0.209 和 0.234。上述分析在一定程度上表明,本文方法对目标识别的区分程度优于可变模糊法。

4 结语

本文利用主成分法融合 Fisher 判别函数的个数,大大减少了工作量。该方法特别适用于具有多个特征的多目标识别。不需要知道目标的分布类型和先验概率,无需定义基本概率指派和隶属函数,不同于极大似然、Bayes 法、基于证据理

论、模糊理论的方法,对目标识别的区分程度较高。另外,尽管文献[14,15]也是采用主成分分析和 Fisher 线性判别,但是文献[14]还是要知道各类模式的先验概率,再对主成分分析法的数学公式进行改进,使其具有灰度归一化操作能力,克服光照对目标的影响,从而对人脸进行识别;文献[15]则基于 Fisher 准则对目标进行分类,完全不同于本文研究的目标级信息融合问题。

参考文献:

- [1] GIRJIA G, RAOL J R, APPAVU RAJ R. Tracking filter and multi-sensor data fusion[J]. *Sādhanā*, 2000, 25(2): 159-167.
- [2] BEGLER P. Shafer - dempster reasoning with application to multi-sensor target identification system[J]. *IEEE Transactions on Systems, Man and Cybernetics*, 1997, 17(6): 968-977.
- [3] YANG YAN, JING ZHANRONG, GAO TAN, *et al.* Multi-sources information fusion algorithm in airborne detection systems[J]. *Journal of Systems Engineering and Electronics*, 2007, 18(1): 171-176.
- [4] CHEN TIANLU, QUE PEIWEN. Target recognition based on modified combination rule[J]. *Journal of Systems Engineering and Electronics*, 2006, 17(2): 279-283.
- [5] 惠增宏. 基于加权 D-S 证据理论的分布式多传感器目标识别[J]. *计算机应用*, 2007, 27(1): 56-58.
- [6] ODEBERG H. Fusion sensor information using fuzzy measures[J]. *Robotica*, 1989, 31: 217-242.
- [7] GLOSSAS N I, ASPRAGATHOS N A. Fuzzy logic grasp control using tactile sensors[J]. *Mechatronics*, 2001, 11(7): 899-920.
- [8] 陈守煜, 胡吉敏. 可变模糊方法及其在工件识别中的应用[J]. *系统工程与电子技术*, 2006, 28(9): 1325-1328.
- [9] 万树平. 多传感器数据的 Fisher 判别法[J]. *传感器与微系统*, 2006, 25(8): 61-63.
- [10] YUAN SHAO, HE FA-CHANG, PENG JIAN. An approach of robot non-vision multi-sensor fusion [J]. *Acta Electronic Sinica*, 1996, 24(8): 94-97.
- [11] CAMERON A, DURRANT-WHYTE H. A Bayesian approach to optimal sensor placement[J]. *International Journal of Robotics Research*, 1992, 12(2): 87-111.
- [12] 车录锋, 周晓军, 徐志农, 等. 可拓方法在多传感器信息融合工件识别中的应用[J]. *系统工程理论与实践*, 2000, 20(8): 91-94.
- [13] 孙文爽, 陈兰祥. 多元统计分析[M]. 北京: 高等教育出版社, 1994: 319-332.
- [14] 石跃祥, 蔡自兴, 王学武. 基于改进的 PCA 算法和 Fisher 线性判别的人脸识别技术[J]. *小型微型计算机系统*, 2006, 27(9): 1731-1736.
- [15] 芮挺, 王金岩, 沈春林, 等. 基于线性分析的特征不变性目标识别[J]. *计算机工程*, 2005, 31(15): 4-6.

(上接第763页)

参考文献:

- [1] 岳东剑, 季洪飞. 语音处理技术在语言学习中的应用[J]. *计算机工程与应用*, 2001, 37(4): 112-114.
- [2] 王昌辉, 谢湘, 赵胜辉. 基于语音识别的汉语发音教学系统[J]. *计算机应用研究*, 2005, 22(11): 11-18.
- [3] ESKENAZI M. Detection of foreign speakers' pronunciation errors for second language training - preliminary results[C]// *Proceedings of the Fourth International Conference on Spoken Language Processing*. Philadelphia, PA: ICSLP, 1996, 3: 1465-1468.
- [4] ZHAO TIANLI, LIU JIA, LU YANFENG, *et al.* An automatic pronunciation teaching system for Chinese to learn English[C]// *Proceedings of IEEE on Robotics, Intelligent Systems and Signal Processing*. Changsha: IEEE, 2003, 2: 1157-1162.
- [5] SUNDRÖM A. Automatic prosody modification as a means for foreign language pronunciation training[C]// *Proceedings of Speech Technology in Language Learning*, 1998. Marholmen, Sweden: STILL, 1998: 49-52.
- [6] MOULINES E, LAROCHE J. Non-parametric techniques for pitch scaling and time-scale modification of speech[J]. *Speech Communication*, 1995, 16(2): 175-207.