第 28 卷第 1 期　　声 学 学 报　　Vol. 28, No. 1
2003 年 1 月　　ACTA ACUSTICA　　Jan., 2003

一种用于强噪声环境下语音识别的含噪 Lombard 及 Loud 语音补偿方法*

田 斌　易克初
(西安电子科技大学综合业务网国家重点实验室　西安　710071)

2001 年 6 月 26 日收到
2001 年 8 月 28 日定稿

摘要　针对语音识别中由于强噪声的影响而引起的 Lombard 和 Loud 效应进行研究，提出了基于训练数据的加性噪声和 Lombard 及 Loud 效应的联合补偿法。对于加性噪声是从谱减法的逆向角度对训练数据在频谱域采用谱加法；对于 Lombard 和 Loud 语音，则采用基于隐马尔可夫模型 (HMM) 状态标注的训练数据补偿，该方法同时考虑 Lombard 和 Loud 语音不同声学单元的不同状态在倒谱域的多种变化和多种变异情况下不同声学单元的音长及相对音长的变化。这种基于数据的多模式补偿使模型自动适应多种噪声和语音变异情况，在强噪声环境下具有很强的鲁棒性，并且不影响识别系统在正常环境或正常发音时的识别性能. 同时，由于补偿是在训练过程中得到，不增加识别时的计算复杂度。

PACS 数：　43.70

A noisy Lombard and Loud speech compensation approach for speech recognition in extremely adverse environment

TIAN Bin　YI Kechu
(*National Key Lab. On ISN, Xidian University* Xi'an 710071)

Received Jun. 26, 2001
Revised Aug. 28, 2001

Abstract　This paper proposes a unified approach for the noisy Lombard and Loud speech recognition based on training data compensation. A spectral addition to the training data is applied to the additive noise which is derived from the reversed point of spectral subtraction, while the compensation in Mel frequency cepstrum (MFC) domain for the Lombard and loud speech is based on HMM state labeling of the training data which take jointly the Mel frequency cepstrum coefficient (MFCC) variance and duration of different states in different acoustic units into account. The new approach is of great robustness in extremely noise and does not worsen the performance under normal environment and normal style. Meanwhile, since the compensation is made in the training phase, it does not increase the complexity of recognition.

引言

在实际环境中，由于语音信号受到各种噪声的影响，会使语音识别系统的性能严重下降[1]。去除加性噪声最常用的方法为谱减法[2,3]。通常，谱减法做为语音识别的预处理技术，对于加性噪声而言，是一种简单而有效的方法。但是在只考虑加性噪声的情况下，语音识别系统性能下降的根本原因在于训练与测试噪声条件的失配，即模板与测试样本的失配。从这个意义上讲，估计再好的噪声模板用于谱减也达不到训练与测试一致的情况。

同时，在高噪环境下，语音信号要受发音人语音变异的影响。发音人在高噪声环境下由于生理、心理受到影响而导致的发音声学特征的变异，称为 Lombard 效应[4]。另外，在噪声环境下，人会不由自

* 国家自然科学基金资助项目 (69872027)

主地大声发音，即所谓的Loud效应。语音识别中Loud和Lombard效应的研究受到人们的广泛关注[5−9]。

本文提出从训练数据倒谱域对训练数据进行补偿，这样能够考虑不同声学单元的多种变异情况及变异语音音长和相对音长的变化。由于 Lombard 语音总是伴随着噪声出现，同时考虑加性噪声的影响，提出了基于训练数据的加性噪声和 Lombard 及 Loud 效应的联合补偿法。对于加性噪声是从谱减法的逆向角度对训练数据在频谱域进行谱加；对于 Lombard 和 Loud 语音，则采用基于 HMM 状态标注[10]的训练数据补偿，该方法同时考虑 Lombard 和 Loud 语音不同声学单元的不同状态在倒谱域的变化和多种变异情况下不同声学单元的音长及相对音长的变化。这种基于数据的多模式补偿使模型自动适应多种噪声和语音变异情况，且不影响识别系统在正常环境或正常发音时的识别性能，而且不增加识别时的时间复杂度。

1　基于训练的加性噪声补偿算法

对加性噪声部分采用的训练数据谱加补偿受启发于去除加性噪声最常用的方法 —— 谱减法和多噪音模式训练[11]的结合。算法可表达如下：

在各种不同的噪声条件下，分别得到一系列噪声频谱，通过聚类生成包含多种噪声类型的模板集：

$$\boldsymbol{N}=\{\boldsymbol{N}_1,\boldsymbol{N}_2,\cdots,\boldsymbol{N}_M\}, \tag{1}$$

其中，M 表示环境噪声的种类，$\boldsymbol{N}_i$ 为：

$$\boldsymbol{N}_i=\{N_i(\omega_1),N_i(\omega_2),\cdots,N_i(\omega_B)\}, \tag{2}$$

ω_k, $k=1,2,\cdots,B$ 为语音分析的各个频段。假定对于词表中的某个词，在安静环境下得到 K 遍训练用语音，在求 MFCC 参数时，经过预加重，短时傅里叶变换，进行 Mel 频率带通滤波，便得到与 (2) 式对应的 B 个 Mel 频带的频谱。这样对一遍 T 帧的语音便得到一个 T 帧的频谱矢量序列。

将每一帧频谱矢量分别加上 (1) 式中的某一个矢量，便得到一个新的 T 帧的频谱矢量序列，连同原来的频谱矢量序列，总共可以得到 $M+1$ 个 T 帧的频谱矢量序列。对每一个频谱矢量求 DCT 变换 (离散余弦变换)，得到 MFCC 参数，这样一遍语音便扩展成 $M+1$ 遍的 MFCC 参数序列，共得到 $K\times(M+1)$ 遍参数序列，用这 $K\times(M+1)$ 参数序列训练该语音的隐马尔可夫模型。

多模式谱加训练补偿法将噪声的多种情况补偿到训练数据中，使得 HMM 自动适应多种复杂的环境噪声，因而模型更具有鲁棒性。因为这种补偿仅需要安静环境下的发音，并且补偿是在训练时完成，不增加识别时的复杂度，因而允许对各种可能在识别时出现的噪声进行充分考虑。

2　Lombard 和 loud 变异语音的补偿方法

Lombard 语音和 Loud 语音不象噪声那样可以简化为频谱的叠加，因此常用的方法就是对语音特征直接在倒谱域进行补偿或者对倒谱域的模型进行补偿[5−9]。由 Lombard 和 Loud 语音与正常语音的区别知道[12−16]，语音的变异与声学内容有关且变化多样，因而直接对测试语音进行补偿就不易考虑各种情况，即使考虑也会把识别时的复杂度变得非常高。Lombard 和 Loud 语音不仅造成频谱的变化，还会造成音长和词语中各声学单元相对音长的变化，而体现在以 HMM 为基础的声学模型中，就是转移概率的变化，而这在模型补偿中往往也体现不出来。

为了能够在倒谱域进行充分的补偿，并能考虑各种不同的变化情况，包括对相对音长变化的影响，提出基于倒谱域的训练补偿法，描述如下：

第一步：声学状态相关的倒谱偏移和相对音长比的获得，可分为如下几个子过程：

(1) 分析词表中声学单元，找出包含在其中的基本声学单元总数，定出一个基本单元表[17]。

(2) 基本单元模型训练和标注

对基本单元表中的声学单元 $i(i=1,2,\cdots,U$, U 为基本单元的总数)，发音人在安静环境下和通过耳机对耳朵加上 M 种不同的噪声的情况下获得这些单元的孤立词发音各 K 遍，共 $(M+1)\times K$ 遍。这样，既得到发音变异的语音，又使语音没有直接叠加噪声。这里采用同一个人的发音，因为不同人在噪声下的 Lombard 和 Loud 变异是相似的，主要体现在音位差别[9,18]。将安静环境下的语音训练成一个单高斯密度 HMM$\lambda_{\text{isilence}}$，将全部环境 (含安静) 下的语音训练成一个混合高斯密度 HMM$\lambda_{\text{imul-style}}$。用 $\lambda_{\text{imul-style}}$ 对 $M+1$ 类的所有语音进行状态标注。

(3) 求状态相关 (State-dependent) 倒谱偏移和状态驻留比

对上面标注的同一基本语音单元 i 的每一类语音属于同一状态的语音求均值和平均帧数。分别求得 i 语音单元 m 类 j 状态和相应安静语音的均值偏移矢量 $\boldsymbol{C}_{imj}$ 和平均驻留比 d_{ijm}。

第二步：补偿训练

对于某个词 v 在正常环境下获得的 N 遍语音数据，假定 v 由若干个基本单元组成，由这些基本单元在安静环境下获得的孤立发音模型得到一个拼接模型 λ_{vpatch}，用 λ_{vpatch} 作为初始模型对 N 遍语音数据进行状态标注，并不断修改模型参数，进行迭代，直至收敛，形成一个新的模型 $\lambda_{\text{vsilence}}$ 。

用 $\lambda_{\text{vsilence}}$ 对任意一遍训练数据通过 Viterbi 算法 [10] 进行状态标注。假定某一遍语音特征矢量 $\boldsymbol{O} = o_1 o_2 \cdots o_{vT}$ 被标注后， $o_{s1} o_{s2} \cdots o_{se}$ 被对应为第一步中的 i 语音单元 j 状态，通过如下的两步可以得到与 $o_{s1} o_{s2} \cdots o_{se}$ 对应的第 m 类变异语音：

(1) 将 $o_{s1} o_{s2} \cdots o_{se}$ 各倒谱矢量分别加上 i 语音单元 m 类 j 状态和相应安静语音的均值偏移矢量 $\boldsymbol{C}_{imj}$ 。

(2) 令 $d_{inc} = \lceil c \times (d_{ijm} - 1) \rceil$，其中 $\lceil\ \rceil$ 为求整函数。如果 $d_{inc} > 0$，则每 $\lceil e/d_{inc} \rceil$ 帧可插入一帧，插入方式采用前后两帧的算术平均生成一帧；如果 $d_{inc} < 0$，则每 $|\lceil e/d_{inc} \rceil|$ 帧可删除一帧，删除方式采用前后两帧的算术平均生成一帧取代原先的两帧。

因而对每一遍语音针对第 m 类语音变异，可补偿生成一遍语音。这样 N 遍语音数据共可生成 $N \times (M+1)$ 遍语音 (含原语音) 。

用 $N \times (M+1)$ 语音数据训练该词的 HMM，便得到该词的补偿的 HMM 。

补偿 HMM 根据声学模型 HMM 的状态相关的倒谱偏移和持续时间的差异补偿到训练数据中，因为补偿在训练过程中，不影响识别时的时间，因而可以考虑更为复杂的情况，比如在相同噪声情况下，可以考虑一般的大声、较快、较慢以及类似喊的发声方式。

3 噪声和语音变异的联合训练补偿法

结合上面加性噪声的补偿原理和 Lombard 和 Loud 变异语音的补偿原则，便得到如下的层次综合补偿策略：

第一步：采用第 1 节中的方法，得到多种噪声类型的模板集，

$$\boldsymbol{N} = \{\boldsymbol{N}_1, \boldsymbol{N}_2, \cdots, \boldsymbol{N}_{M_N}\}, \tag{3}$$

其中， M_N 表示环境噪声的种类， $\boldsymbol{N}_i$ 为：

$$\boldsymbol{N}_i = \{N_i(\omega_1), N_i(\omega_2), ..., N_i(\omega_B)\}. \tag{4}$$

第二步：采用第 2 节中第一步的方法，获得声学状态相关的倒谱均值偏移矢量 $\boldsymbol{C}_{imj}$ 和平均驻留比 d_{ijm}， $i = 1, 2, \cdots, N$， $m = 1, 2, \cdots, M_L$， $j = 1, 2, \cdots, S$， N 为基本单元数， M_L 为变异情况类数目， S 为基本单元的状态数。

第三步：噪声补偿数据的获得。假定对于词表中的某个词，在安静环境正常发音下得到 K 遍训练用语音，在求 MFCC 参数时，经过预加重，短时傅里叶变换，进行 Mel 频率带通滤波，对每一帧语音便得到与 (4) 式对应的 B 个 Mel 频带的频谱。这样对一遍 T 帧的语音便得到一个 T 帧的频谱矢量序列。

将每一帧频谱矢量分别加上 (3) 式中的某一个矢量，便得到一个新的 T 帧的频谱矢量序列，连同原来的频谱矢量序列，总共可以得到 $M_N + 1$ 个 T 帧的频谱矢量序列。对每一个频谱矢量求 DCT 变换 (离散余弦变化)，得到 MFCC 参数，这样一遍语音便扩展成 $M_N + 1$ 遍的 MFCC 参数序列，共得到 $K \times (M_N + 1)$ 遍参数序列。

第四步：用第 2 节中第二步的方法，将上面的一遍特征语音序列经过补偿扩展为 $(M_L + 1)$ 遍特征参数序列，那么经过两种分别在频谱域和倒谱域的补偿，一遍原始语音最终可以生成 $(M_N + 1) \times (M_L + 1)$ 遍特征矢量序列， K 遍训练用语音便可以生成 $K \times (M_N + 1) \times (M_L + 1)$ 遍特征语音矢量序列。用这 $K \times (M_N + 1) \times (M_L + 1)$ 遍语音进行训练，可以得到该词的 HMM 。

4 实验结果

4.1 实验条件

基本词表：语音识别词表共 200 个词，包括单字词 30 个，二字词 137 个，三字词 20 个，四字词 7 个，五字词 5 个和六字词 1 个。

语音集：由特定人在安静环境下男一人、女一人各 30 遍发音，并同时有 60~75 dB, 80 dB, 85 dB, 90 dB, 95 dB, 100 dB, 105 dB, 110 dB, 115 dB 的噪声的环境下录音各 5 遍，噪声环境下录音时没有给与特别提示，允许产生变异。其中， 85~115 dB 的噪声为歼八 – II 座舱内的噪声录音。

环境噪声： 60~75 dB, 80 dB, 85 dB, 90 dB, 95 dB, 100 dB, 105 dB, 110 dB, 115 dB 的噪声的各约 2 min，其中， 85~115 dB 的噪声为歼八 – II 座舱内的噪声录音。

语音识别特征参数：语音特征采用 14 维 MFCC (Mel 频率倒谱系数) 。语音经 16 kHz 采样， 8 比特量化，取 20 ms 为一帧，帧间隔为 10 ms，经高频预加重，将信号通过 21 个按 Mel 尺度划分的三角形带通滤波器，取对数后作 IDCT(反余弦变换) 。

语音模型：采用 12 状态自左至右的分划高斯密度函数的隐马尔可夫模型 (HMM)，混合数为 2 ～ 24，在实验中将给与说明。

4.2　实验结果

(1) 不考虑语音变异的加性噪声下的识别实验

用安静环境特定人 30 遍语音生成混合数为 2 的 HMM，作为基准系统 (BS)，这样对每个人生成一个基本系统。其对各种噪声下的平均语音识别结果如表 1，为了排除语音变异的影响，采用仿真的方法，在安静环境下的语音中加进噪声。

将噪声聚为 4, 8, 12 类，进行基于谱加的训练补偿，得到混合数为 12 的 HMM，其对仿真含噪语音的误识率如表 2 所示，其中：

$$\text{误识率下降比} = \frac{\text{基准系统误识率} - \text{当前误识率}}{\text{基准系统误识率}} \times 100\%. \tag{5}$$

(2) 噪声环境下的联合补偿识别实验

首先用基准系统 (BS) 对各种噪声下采集到的变异语音进行识别实验，结果如表 3 。我们看到表 3 的结果要比表 1 差得多，也从一个侧面说明了噪声环境下 Lombard 效应和 Loud 效应对语音识别的影响远大于加性噪声的影响。

对语音识别命令表，首先确定一个基本单元表。对于这个系统，基本单元可以有三种选择。第一种就是直接选择单词作为获得倒谱偏移的基本单元。第二种选择就是音节，该语音命令表中含有的不同音节共 155 个。第三种选择就是声韵母。对本实验而言，似乎选择单词为单元已经比较简单，但为了使本文的算法具有普适性，仍然选择音节作为基本单元。为了简单，没有考虑同一音节的不同声调的情况，即假定它们的 Lombard 和 Loud 变异相似，可以归为一类。

对基本单元表，用和训练语音之一的发音人 (设为 A) 在安静环境下和 9 种不同的噪声通过耳机送入耳朵的境况下获得这些单元的孤立词发音各 10 遍，获得补偿参数。为了降低复杂性和增加可靠性，这里没有按噪声情况分类，而是采用聚类的方法，对每一基本单元的每个状态得到 6 个变异参数类。对于噪声的不同情况，采用实验一中得到的 8 类。然后用第 3 节的方法进行补偿训练， HMM 的混合数为 24 。识别结果如表 4 和 5 所示，表 4 是识别系统为训练补偿参数的发音人，表 5 所示为由 A 的补偿参数得到的另一个发音人的补偿系统的识别结果。结果表明，联合补偿法极大地降低了误识率。同时，语音变异的差异主要体现在音位差异，不同人在同一条件下发同一个音的变异相近。

表 1　基准系统 (BS) 对仿真含噪语音的误识率

噪声量级 /dB	< 60	60~75	80	85	90	95	100	105	110	115
误识率 /%	1.72	1.90	2.12	4.44	11.24	13.91	18.77	21.45	27.72	32.09

表 2 谱加补偿系统 (SA) 对仿真含噪语音的误识率

类　数		噪声量级 /dB									
		< 60	60~75	80	85	90	95	100	105	110	115
4	误识率 (%)	1.73	1.82	1.94	2.79	3.70	4.48	4.45	5.10	5.41	6.80
	误识率下降比 (%)	−0.58	4.21	8.49	37.16	67.08	67.79	76.29	76.22	80.48	78.80
8	误识率 (%)	1.69	1.80	1.90	2.69	3.60	4.42	4.40	4.97	5.33	6.90
	误识率下降比 (%)	01.74	5.26	10.37	39.41	67.97	68.22	76.55	76.82	80.77	78.49
12	误识率 (%)	1.73	1.81	1.84	2.54	3.10	4.00	4.09	4.86	4.90	6.74
	误识率下降比 (%)	-0.58	4.73	13.20	42.79	72.41	71.24	78.20	77.34	82.32	78.99

表 3　基准系统 (BS) 对噪声环境下变异语音的误识率

噪声量级 (dB)	< 60	60~75	80	85	90	95	100	105	110	115
误识率 (%)	1.72	1.92	2.40	4.81	11.62	19.6	23.5	39.6	46.7	68.2

表 4 补偿模型对噪声环境下变异语音的识别性能 (补偿参数发音人)

噪声量级 /dB	< 60	60~75	80	85	90	95	100	105	110	115
误识率 (%)	1.74	1.86	1.90	2.41	4.69	4.67	8.54	9.61	10.86	12.82
误识率下降率 (%)	−1.16	3.12	20.83	49.89	59.63	76.17	63.65	75.73	76.74	81.20

表 5 补偿模型对噪声环境下变异语音的识别性能 (非补偿参数发音人)

噪声量级 (dB)	< 60	60~75	80	85	90	95	100	105	110	115
误识率 (%)	1.73	1.90	2.10	2.57	4.98	6.66	9.49	12.01	13.97	14.79
误识率下降率 (%)	−0.58	1.04	12.50	46.56	57.14	66.02	59.61	69.67	70.08	78.31

5 结论

本文提出了基于训练数据的加性噪声和 Lombard 及 Loud 效应的联合补偿法。对于加性噪声是从谱减法的逆向角度对训练数据在频谱域采用谱加法; 对于 Lombard 和 Loud 语音, 则采用基于 HMM 状态标注的训练数据补偿, 该方法同时考虑 Lombard 和 Loud 语音不同声学单元的不同状态在倒谱域的多种变化和多种变异情况下不同声学单元的音长及相对音长的变化。这种基于数据的多模式补偿使模型自动适应多种噪声和语音变异情况, 且不基本影响识别系统在正常环境或正常发音时的识别性能。由于实际情况下, Lombard 效应语音不能脱离噪声环境单独采集到, 且无论噪声还是发音变异都会存在多种复杂情况, 这种联合补偿使识别系统在强噪声环境下具有很高鲁棒性, 且由于补偿是在训练过程中进行的, 不增加识别时的计算复杂度。

参 考 文 献

1 Palival K K. Neural net classifiers for robust speech recognition under noise environments. Proceedings of ICASSP 90, 1990: 429—432

2 Boll S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979; **27**(2): 113—120

3 Compernole D Van. Increased noise immunity in large vocabulary speech recognition with the aid of spetral subtraction. Proceedings of ICASSP 87, 1987: 1143—1146

4 Rajasekaran P, Doddington G, Picone J. Recognition of speech under stress and in noise. Proceedings of ICASSP 86, 1986: 733—736

5 Mokbel, Chafic E, Chollet, Gerard G. Automatic word recognition in cars. *IEEE Transactions on Speech and Audio Processing*, 1995; **3**(5): 346—356

6 Milner B P, Vaseghi. Comparison of some noise-compensation methods for speech recognition in adverse environments. *IEE Proceedings: Vision, Image and Signal Processing*, 1994; **141**(5): 280—288

7 Applebaum T H, Hanson. Robust speaker-independent word recognition using spectral smoothing and temporal derivatives. EUSIPCO 90, 1990: 1183—1186

8 Anglade Y, Junqua. Acoustic-phonetic study of Lombard speech in the case of isolated-words. EUSIPCO 90, 1990: 1195—1198

9 Stanton B J, Jamieson L H. Robust recognition of Loud and lombard speech in the fighter cockpit environment. Proceedings of ICASSP 88, 1988: 675—678

10 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 1989; **77**(2/3): 257—285

11 陈景东, 徐 波, 黄泰翼. 高鲁棒性语音识别算法的研究现状. 见: 王炳锡主编, 第八届语音图像与通信信号处理论文集, 1998: 126—137

12 Afify M, Yifan Gong, Haton J P. A general joint additive and convolutive bias compensation approach applied to noisy Lombard speech recognition. IEEE Transactions on Speech & Audio Processing, 1998; **6**(6): 524—38

13 Sang-Mun Chi, Yung-Hwan Oh. Spectral magnitude normalisation and cepstral coefficient transform for noisy-Lombard speech recognition. *Electronics Letters*, 1996; **32**(19): 1761—1763

14 Laurent Barbier, Gerard Ghollbt. Robust speech parameters extraction for word recognition in noise using neural networks. Proceedings of ICASSP 91, 1991: 145—148

15 Afify M, Yifan Gong, Haton J P. A unified maximum likelihood approach to acoustic mismatch compensation: application to noisy Lombard speech recognition. Proceedings of ICASSP 97, 1997: 839—842

16 韩纪庆. 噪声环境下顽健的语音识别方法. 博士学位论文, 哈尔滨工业大学, 1998

17 易克初, 田 斌, 付 强. 语音信号处理. 北京: 国防工业出版社, 2000: 198—201

18 田 斌. 实用化汉语语音识别理论及关键技术研究. 博士学位论文, 西安电子科技大学, 2000