时间反转语音掩蔽的语音信号可懂度的客观评价方法*

王 玥[1] Philip Leistner [德][2] 李 平[1]

([1] 中国科学院声学研究所 北京 100190 [2]Fraunhofer Institute of Building Physics Stuttgart 70569)

摘要: 对于开放型办公室语音掩蔽系统性能的评价,语言可懂度是很重要的一个方面,目前通常采取的客观评价方法是 STI 。将语音信号按一定时间帧长反转后得到的信号我们称为时间反转语音,时间反转语音已被作为有效掩蔽信号之一。虽然对于由平稳噪声掩蔽的语音信号,STI 与主观理解的语言可懂度相关性很好。但研究发现 STI 不适用于估计由时间反转语音掩蔽的语音信号的语言可懂度。文章分析了 STI、PESQ 及 mNCM 客观评价方法并进行了实验,实验结果表明,PESQ 及 mNCM 对于由反转语音掩蔽的语音信号仍能较好估计语言可懂度。文章根据客观评价结果,进一步比较了反转语音掩蔽算法的不同参数(反转帧长与信噪比) 对于语言可懂度的影响。发现反转帧长的增加和信噪比的降低会导致较低的语言可懂度。

关键词: 语言可懂度,客观评价,时间反转语音,掩蔽

Objective Measurements of Speech Intelligibility for Speech Masked by Time Reversed Speech

WANG Yue[1,2], Philip Leistner [DE][2], LI Ping[1]

([1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China,
[2]Fraunhofer Institute of Building Physics, Stuttgart 70569, Germany)

Abstract: Speech intelligibility is an important aspect for evaluating speech masking system in open – plan offices. STI is a common objective measurement so far. Time reversed speech, which is speech reversed according to certain frame length in time domain, became one of effective maskers. Although STI has good correlation with subjective speech intelligibility when masker is steady noise, it is shown in this paper that STI cannot used to predict speech intelligibility for speech masked by time reversed speech. We analyzed STI, PESQ and mNCM. Results showed that PESQ and mNCM can predict speech intelligibility well. We also compared the effects to speech intelligibility of different parameters (reversed frame length and SNR) for speech with time reversed masker, and found increase of reversed frame length and decrease of SNR would lead to poorer speech intelligibility.

Keywords: speech intelligibility; objective measurements; time reversed speech; masking

1 介绍

开放型办公室中语音信号的语言可懂度越低,声音环境的满意度越高[1] 。近年来,研究者尝试了多种方法及信号掩蔽语音,期望降低语言可懂度。研究表明,掩蔽信号越接近语音信号的频谱越能有效的掩蔽[1-3] 。因此,将被掩蔽语音信号在时域上以一定帧长进行反转得到的时间反转语音由于具有与语音类似的频谱且较易被实时获取而被认为是有效掩蔽信号之一[4] 。

评价语言可懂度通常采用主观实验的方法,这也被认为是最可信赖和最准确的方法,但它相较客观评价方法费时费力且不方便实施。因此,为了评价语音掩蔽的效果我们期望可以采用客观评价方法估计语言可懂度。Hongisto 等人[5,6] 采用了语言传输指数(Speech Transmission Index STI) 进行客观评价, 并提出了一

本文于 2012 – 02 – 08 收到。
* 中国科学院“科技助残行动计划”项目,高性能数字助听器,项目编号 KGCX – YW – 616。

个计算在开放型办公室中计算 STI 的有效模型。但是,他们采用的是类似于粉红噪声的平稳噪声(能量集中在低频,每倍频程能量以 -5dB 衰减的噪声) 作为掩蔽信号。而时间反转语音不同于平稳噪声,它是具有波动性的。我们需要验证 STI 是否对于采用时间反转语音作为掩蔽信号的语音信号仍能有效估计语言可懂度。如果不能,是否有其他客观评价方法可以较好的估计时间反转语音掩蔽的语音信号的语言可懂度。

时间反转语音是被掩蔽语音信号在时域上以一定帧长进行反转而获得。然后调整掩蔽信号,即时间反转语音的幅度,按一定信噪比(原语音信号与掩蔽信号能量之比) 与原语音信号相加,就获得了时间反转语音掩蔽的语音信号。因此,时间反转语音掩蔽算法主要涉及到两个参数: 反转帧长及信噪比。参数的改变会影响时间反转语音掩蔽的语音信号的语言可懂度。本文也研究分析了这两个参数对于语音信号语言可懂度的影响。

文章的主要内容包括两个方面: ①验证是否可以采用客观评价方法估计时间反转语音掩蔽的语音信号的语言可懂度; ②分析两个参数(反转帧长及信噪比) 对于时间反转语音掩蔽的语音信号语言可懂度的影响。文章一共介绍了三种客观评价方法,包括传统的 STI,分别给出了这三种客观评价方法的结果与主观语言可懂度的相关性。然后通过采用与主观语言可懂度相关性较高的客观评价方法,分析了反转帧长及信噪比对于时间反转语音掩蔽的语音信号语言可懂度的影响。

2　方法与实验

语音传输系数(STI) [7] 是目前广为应用的一种客观评价语言可懂度的方法。Hongisto 等人提出了一种预测开放型办公室中 STI 的模型 [8],即可以根据已知的室内及语音信号信息,不需实际测量而获得 STI 值。改进的协方差归一方法(Modified Normalized covariance method, mNCM) 是基于语音的 STI 方法中的一种,与传统的人工测试信号相比,基于语音的 STI 采用语音信号获得语言可懂度。mNCM 基于一个新的频带权重函数,在噪声环境下评价语言可懂度方面效果突出 [7]。语音质量的感知评价方法(Perceptual evaluation of speech quality, PESQ) 在用于预测电话网络和语音编码质量方面非常可靠。虽然它最初是为了评价语音质量而设计的,PESQ 在噪声环境下也能很好的评价语言可懂度 [8]。对于时间反转语音掩蔽的语音信号,为了获得这些客观评价方法与主观可懂度的关系,我们分别研究了这些客观评价方法,并将它们的结果与主观实验结果进行了对比。

2.1　语言传输指数(STI)

STI [7] 是一种通过计算语言可懂度测量语音传输通道质量的物理方法。它通过信号调制深度的减少预测混响以及加性噪声的影响。文献 [6] 假设语音信号的信噪比及混响时间已知,提出了一种有效预测开放型办公室中 STI 的模型。本文在接下来的实验中采用 [8] 中的模型预测 STI,忽略了说话人与听众的距离和开放型办公室的混响。

2.2　改进的协方差归一方法(mNCM)

mNCM 是基于语音的 STI 的方法之一,它基于原信号与输出信号包络的协方差,并且采用了新的频带权重函数 [8]。首先,原信号与输出信号通过带通滤波器组分为 k 个频带。每个频带中原信号与输出信号分别为 $x_k(t)$ 和 $y_k(t)$。在每个频带中,表面信噪比(SNR) 如下:

$$aSNR = 10\log_{10}\left(\frac{r^2}{1-r^2}\right) \tag{1}$$

其中,r 是 $x(t)$ 和 $y(t)$ 的标准协方差,即 $r^2 = \dfrac{\lambda_{xy}^2}{\lambda_x \lambda_y}$,

$$\lambda_{xy} = E\{(x(t)-\mu_x)(y(t)-\mu_y)\} \tag{2}$$

$$\lambda_x = E\{(x(t)-\mu_x)^2\} \tag{3}$$

其中,$\mu_x = E\{x(t)\}$,$\mu_y = E\{y(t)\}$,$E\{\cdot\}$ 表示期望值。然后与 STI 方法 [8] 的计算一样,$aSNR$ 的取值被限制在范围 $[-15,15]$ dB。每个频带中的传输系数是表面 SNR 的线性函数,定义在 0 到 1 之间。

$$TI_k = \frac{aSNR_k + 15}{30} \tag{4}$$

最后，通过计算各频带中传输函数 TI_k 的加权平均得到 STI

$$STI = \sum_k w_k TI_k \tag{5}$$

在 mNCM 中，$w_k = \left(\sum_t x_k^2(t)\right)^p$。本文中我们选择 p = 1.5。

2.3 语音质量的感知评价方法(PESQ)

PESQ 是一种复杂的评价语音质量的客观方法。但是在所有语音质量客观评价方法中它表现出了与主观语言可懂度最好的相关性[8]。PESQ[9]方法中，信号首先统一到标准听力水平并在时间上一致。然后，通过听觉转换将信号变为响度谱。在认知模型中计算两个误差参数: 平均扰动值和平均非对称扰动值，并结合二者得到客观听觉质量的平均分数(MOS)。PESQ 取值在 1.0 到 4.5 之间，越大的值意味着语音质量越好。

2.4 主观实验

采用 25 个德语语音信号作为原信号。因为在开放型办公室中常见的噪声不是单词或单音节，我们仅采用了包含句子的语音信号。每个信号分别含有 20 个短句。全部的句子都只包括简单的主谓宾成份，并由专业播音员录音。之后原信号加上自身的实践反转信号形成了时间反转语音掩蔽的语音信号。时间反转语音的参数如下: 反转帧长分别取 50ms，100ms，150ms，200ms，500ms; 信噪比在每个反转帧长情况下分别为 -15dB，-10dB，-5dB，0dB，5dB。本次实验没有考虑时间延迟。

总共 10 位母语为德语的成年人作为听众参与了实验。首先进行一共 10 句话的测试练习，测试句中不含有掩蔽信号。练习中要求每个听众在电脑上利用键盘复述出刚听到的句子。成功完成测试练习后开始正式的主观实验。实验素材包括 25 个语音信号，分别对应 25 种情况(5 种反转帧长 ×5 种信噪比)，每个语音信号包括 20 个短句。因此每位听众听到一共 500 个短句(20 个短句 ×25 种情况)。在实验过程中要求每个听众在电脑上复述出所有他听懂了的单词，复述的正确率即为实验的主观语言可懂度。每位听众的测试信号顺序随机，每个短句后跟随一定的间隔时间以便于复述，测试时间大约 90 分钟。

2.5 实验结果

客观评价指标与主观语言可懂度的关系如图 1 和图 2 所示。图 1 显示了通过模型[6]预测的 STI 与主观可懂度的关系。图 2(a)，图 2(b) 分别显示了 PESQ 和 mNCM 与主观可懂度的关系。

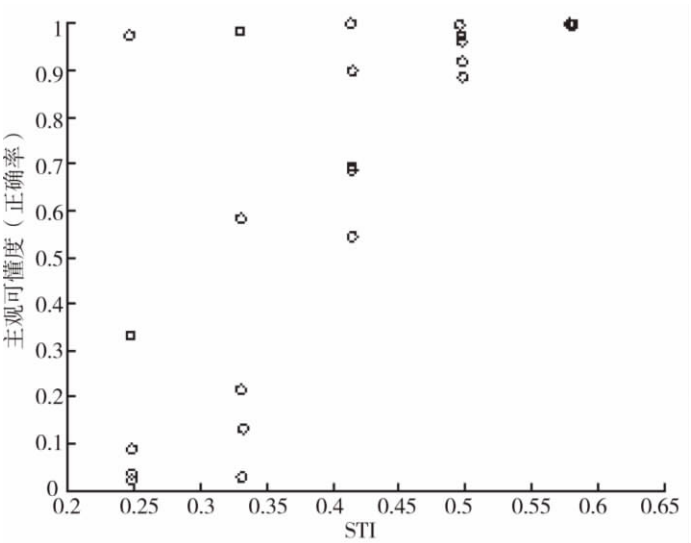

图 1 主观可懂度与 STI 指标的关系

图 1 显示不同的主观语言可懂度可能对应相同的 STI，也就是说 STI 对于时间反转语音掩蔽的语音信号无法预测语言可懂度。图 2 中 PESQ、mNCM 与主观可懂度的关系可通过二次曲线进行拟合，获得最小的标准偏差。图 2(a) 显示了 PESQ 与主观可懂度的关系，拟合后标准偏差为 $\sigma = 0.14$ 。图 2(b) 显示了 mNCM 与主观可懂度的关系，拟合后标准偏差为 $\sigma = 0.13$ 。

本文还采用了两个值来评价以上客观评价指标与主观可懂度的关系。第一个为相关系数，r，第二个为估计误差的标准偏差 $\sigma_e = \sigma_d \sqrt{1 - r^2}$ ，其中 σ_d 是给定情况下语音识别的标准偏差，$\sigma_d = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$，其中$\bar{x}$表示矢量中各元素 X_i 的平均值。σ_e 是计算的误差的标准偏差。较小的 σ_e 代表客观评价能较好的估计语言可懂度。如前所述，本文中的情况是语音信号由时间反转语音所掩蔽。计算的相关系数与预测误差如表 1 中所示。

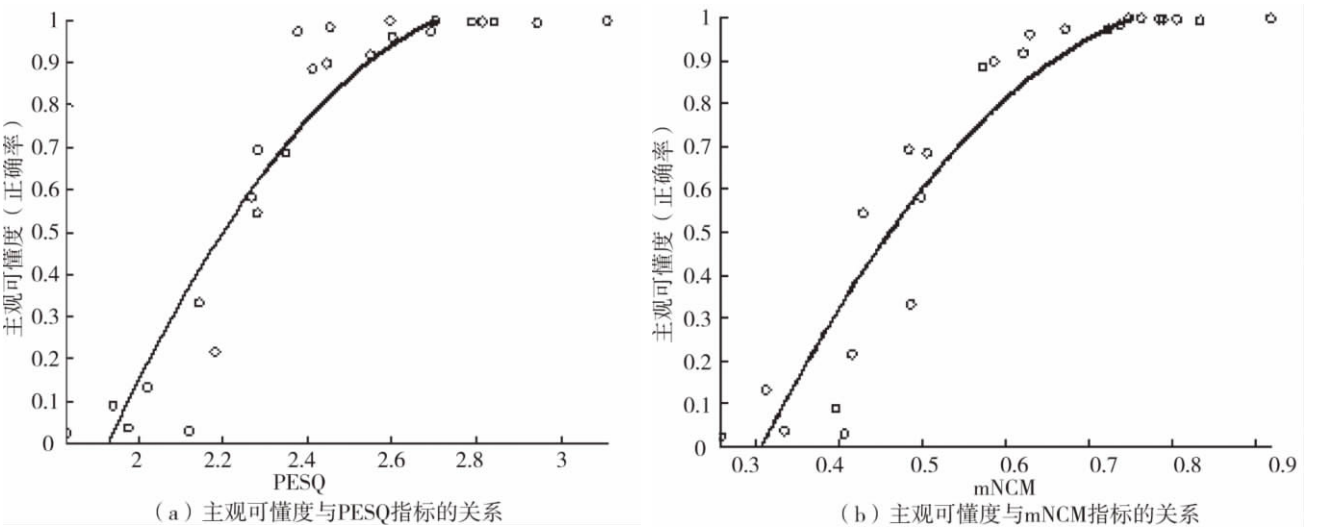

图 2　客观评价指标与主观语言可懂度关系图

表 1　客观评价指标与主观可懂度的相关性

Objective measure	r	σ_e
STI	0.7442	0.2541
PESQ	0.8780	0.1601
mNCM	0.9019	0.0796

注: 其中，r 代表相关系数，σ_e 代表标准偏差。

2.6　实验讨论

实验发现 STI 无法准确估计时间反转语音掩蔽的语音信号语言可懂度，这与之前研究者的结论一致[3]。可能的原因是时间反转语音与通常的平稳噪声有很大区别，它具有明显的波动性并且类似于语音。所以 STI 可能将时间反转语音也认为是语音信号而得到错误的估计值。

图 2 说明针对时间反转语音掩蔽的语音信号，PESQ 和 mNCM 与主观可懂度之间存在较好的相关性。根据表 1，也可见 STI 与主观可懂度的相关性最低，偏差最大; mNCM 与主观可懂度的相关性最好，偏差最小; PESQ 与主观可懂度的相关性较好，但低于 mNCM。

语言可懂度可由客观评价指标值推断得到。通常认为语言理解正确率低于 50% 为几乎不可懂，根据以上的实验结果，我们可以得到对于时间反转语音掩蔽的语音信号，PESQ 和 mNCM 与主观听觉感受的对应关系。如果 PESQ < 2.3，语音信号几乎完全不可懂; 如果 2.3 < PESQ < 2.4，信号可懂度很差; 如果 2.4 < PESQ < 2.8，信号可懂度中等; 如果 PESQ > 3，信号可懂。对于 mNCM 而言，如果 mNCM < 0.45，语音信号完全不

可懂; 如果 0.45 < mNCM < 0.55,信号可懂度较差; 如果 0.55 < mNCM < 0.75,信号可懂度中等; 如果 NCM > 0.75,信号可懂度好。

3 时间反转参数对语言可懂度的影响

尽管时间反转语音是一种较为有效的可以降低语言可懂度的掩蔽信号[4],算法中参数的选取仍会对掩蔽的效果产生影响。时间反转语音掩蔽语音信号主要有两个参数: 反转帧长和信噪比。根据研究结果[10],最佳的反转帧长介于 120 至 240ms 之间。为了了解不同参数对于掩蔽效果的影响,我们以 20ms 为步长,在 120 至 240ms 内取反转帧长; 信噪比以 5dB 为步长,在 -15 至 10dB 间取值。为了比较和了解极端情况下的掩蔽效果,我们也选择了反转帧长为 5ms 和 500ms,对文中 2.4 提到的一共 33 个信号进行仿真。采用客观评价指标 PESQ 和 mNCM 对掩蔽后信号的语言可懂度进行评价。结果如图 3 所示。

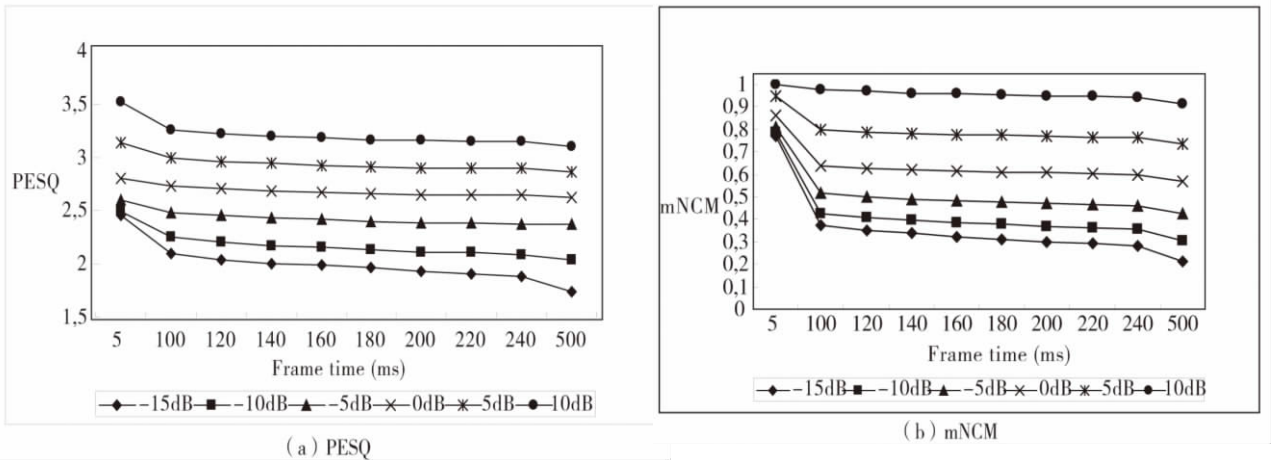

图 3 时间反转语言掩蔽的语言信号的客观评价指标

图 3 中显示,随着信噪比的减小,PESQ、mNCM 都逐渐减小; 随着反转帧长的增加,PESQ、mNCM 也逐渐减小。这就是说,掩蔽后信号的语言可懂度随着反转帧长的增加和信噪比的减小而减小。

文献[10]中认为最佳的反转帧长介于 120 至 240ms 之间,而本节实验结果表明,反转帧长越长,掩蔽后信号的语言可懂度越低。造成二者之间差别的原因可能有以下两点: ①文献[10]的反转帧长范围是 40 - 240ms,即反转帧长上限只取到了 240ms,没有给出反转帧长超过 240ms 的掩蔽信号语言可懂度结果; ②测试语音信号的不同。文献[10]主要针对三个辅音/p/、/t/、/k/进行实验,并以日语为载体; 本文采用德语语音信号,测试样本符合日常语流情况,不针对某一单个辅音进行测试。

在仿真中,我们设定掩蔽信号加入的延迟时间为一个反转帧长,并且每个信号都归一化到相同的声压级。在实际应用中,越长的反转帧长会给系统带来越长的延迟时间。因此实际中为了取得较好的掩蔽效果,即达到较低的语言可懂度,最佳帧长应该考虑系统延迟及语言可懂度取一个折中值。

4 结束语

文章研究了对平稳噪声掩蔽的语音信号语言可懂度能够较好评价的客观评价指标 STI 以及另外两个客观评价指标: PESQ, mNCM 与时间反转语音掩蔽的语音信号语言可懂度的关系。结果显示 STI 无法有效的估计时间反转语音掩蔽的语音信号的语言可懂度。虽然 PESQ 和 mNCM 均不是专门设计用于估计波动性噪声影响的语音信号的可懂度的客观评价方法,但实验结果表明,对于时间反转语音掩蔽的语音信号,二者与主观语言可懂度的相关性较好(r > 0.87) 。这个结果可以帮助人们更易获得时间反转语音掩蔽的语音信号

可懂度,不必一定采用费时费力的主观评价方法。文章接着采用这两个客观评价指标,对时间反转语音掩蔽算法的参数对掩蔽效果的影响进行了研究。研究发现反转帧长的增加会导致较低的语言可懂度,而信噪比的增加会带来较高的语言可懂度。在这个的基础上,我们可以结合实际情况,选择最佳参数利用时间反转语音进行掩蔽,以获取期望的掩蔽效果。

致谢:

本文的研究受到了中科院中欧联合培养博士生项目以及德国弗劳恩霍夫建筑物理研究所的支持。

参 考 文 献

[1] Veitch, J. A. , Bradley, J. S. , Legault, L. M. , Norcross, S. , Svec, J. M. Masking speech in open – plan offices with simulated ventilation noise: noise level and spectral composition effects on acoustic satisfaction [J]. Institute for Research in Construction, Internal report No. 2002, IRC – IR – 846

[2] Hongisto, V. Effects of sound masking on workers – A case study in a landscape office [C]. In 9th International Congress on noise as a Public Health Problem (ICBEN) Mashantucket, Connecticut, USA, 2008, 21 – 25

[3] Haapakangas A, Kankkunen E, Hongisto V, Virjonen P, Oliva D, Keskinen E. Effects of five speech masking sounds on performance and acoustic satisfaction – implications for open – plan offices [J]. acta acustica united with acustica, 2011, 97(4) 641 – 655.

[4] Koenraad S. Rhebergen, Niek J. Versfeld, Wouter. A. Dreschler. Release from informational masking by time reversal of native and non – native interfering speech [J], J. Acoust. Soc. Am. 2005, 118 , 1274 – 1277

[5] Valtteri Hongisto, Annu Haapakangas, Esko Keskinen, Miia Haka. Effects of office noise on work performance and acoustic comfort – laboratory experiment simulating three different office types [C], the 39th International Congress and Exposition on Noise Control Engineering, Lisbon, 2010.

[6] Valtteri Hongisto, Jukka Keraenen, Petra Larm. Prediction of speech transmission index in open – plan offices [C], Joint Baltic – Nordic Acoustics Meeting, 2004.

[7] CEI – IEC standard n. 60268 – 16, March 1998, Sound system equipments – objective rating of speech intelligibility by speech transmission index [S], Second Edition.

[8] Jianfen Ma, Yi Hu and Philipos C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band – importance functions [J]. J. Acoust. Soc. Am. 125, 2009, 3387 – 3405

[9] Rix, A. , Beerends, J. , Hollier, M. , and Hekstra, A. Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs [C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2001, 2, 749 – 752

[10] Takayuki Arai. Making speech with its time – reversed signal [J], Acoust. Sci. & Tech. 31, 2010, 2, 188 – 190