软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(Suppl.(2)):64−68

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

以语音出现时频相关性为基础的语音掩模估计*

战 鸽, 黄兆琼, 应冬文, 潘接林, 颜永红

(中国科学院 声学研究所,北京 100190)

通讯作者: 战鸽, E-mail: zhange@hccl.ioa.ac.cn

摘 要: 在二维的时频域网格结构中,相邻点上语音信号的存在与否是相关的,传统的马尔可夫链不能对二维的时频相关性进行自适应的建模.基于语音信号在时频域中的相关性,提出了一种利用二维的相关模型估计语音掩模的方法.该方法将时频域中带噪语音信号的对数功率谱划分为语音和非语音类,利用时域中的状态转移概率和前向因子描述语音信号的时域相关性,同时利用频域中的状态转移概率和邻域因子描述语音信号的频域相关性.通过全局的统计最优化,该模型将时域相关性和频域相关性相结合.给出了该模型的序贯化更新方法,逐帧更新模型并估计语音出现概率.在当前已知对数功率谱和模型参数的条件下,通过最大化后验概率得到的语音信号状态矩阵可以作为语音掩模的最优估计.将该方法与几种现有的语音掩模在线估计方法进行比较,实验结果显示出了该方法的优越性.

关键词: 语音掩模;时频相关性;语音出现概率;邻域因子;在线估计

中文引用格式: 战鸽,黄兆琼,应冬文,潘接林,颜永红.以语音视频相关性为基础的语音掩模估计.软件学报,2016,27(Suppl.(2)):64−68. http://www.jos.org.cn/1000-9825/16020.htm

英文引用格式: Zhan G, Huang ZQ, Ying DW, Pan JL, Yan YH. Speech mask estimation using the time-frequency correlation of speech presence. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):64−68 (in Chinese). http://www.jos.org.cn/1000-9825/16020.htm

Speech Mask Estimation Using the Time-Frequency Correlation of Speech Presence

ZHAN Ge, HUANG Zhao-Qiong, YING Dong-Wen, PAN Jie-Lin, YAN Yong-Hong

(Institute of Acoustics, The Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper proposes a method to estimate the spectrographic speech mask based on a two-dimensional (2-D) correlation model. The proposed method is motivated by a fact that the time and frequency correlations of speech presence are interwoven with each other in the time-frequency domain. Conventional Markov chain is incapable of simultaneously modeling the time and frequency correlations in an adaptive way. The 2-D correlation model is presented to describe the correlation of speech presence in the TF domain, where the speech presence and absence are taken as two states of the model. The time correlation is modeled by the time state-transition probability and the forward factor, while the frequency state-transition probability and the corresponding neighbor factor are defined to describe the frequency correlation. The time and frequency correlations are incorporated into the model by maximizing the Q-function. A sequential scheme is presented to online estimate the parameter set. Given the observed spectrum and the parameter set, the state matrix that maximizes the posteriori probability is regarded as the optimal estimate of the speech mask. The proposed method was compared with some well-established methods. The experimental results confirmed its superiority.

Key words: speech mask; time-frequency correlation; speech presence probability; neighbor factor; online estimation

* 基金项目: 国家自然科学基金(11461141004, 91120001, 61271426); 中国科学院战略性先导科技专项(XDA06030100, XDA06030500); 国家高技术研究发展计划(863)(2012AA012503); 中国科学院重点部署项目(KGZD-EW-103-2)

Foundation item: National Natural Science Foundation of China (11461141004, 91120001, 61271426); Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06030100, XDA06030500); National High-Tech R&D Program of China (863) (2012AA012503); CAS Priority Deployment Project (KGZD-EW-103-2)

收稿时间: 2015-06-01; 采用时间: 2016-01-05

语音掩模在特征恢复[1,2]、语音分离[3,4]、语音感知[5,6]和噪声估计[7,8]方面得到了广泛应用.在这些情况下,语音掩模被视为一个表征语音信号在时频域内存在与否的状态矩阵,是影响系统性能的重要因素.

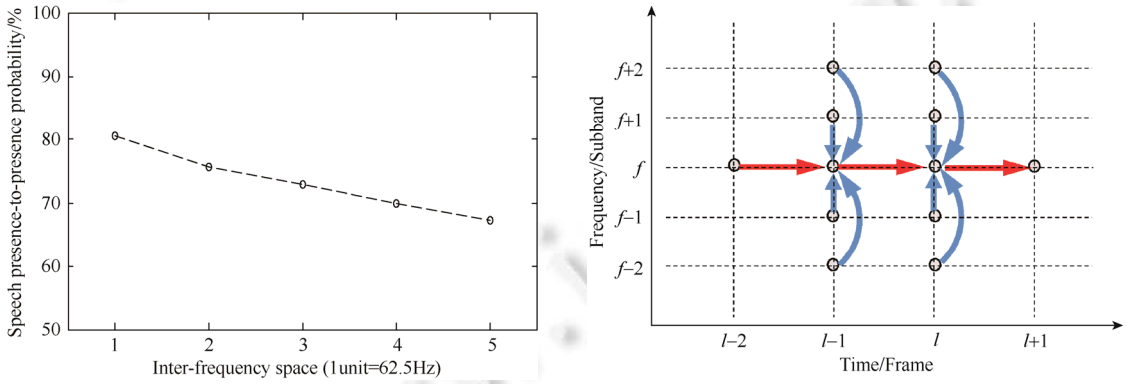
一般来说,语音掩模包含两类.一类是二值掩模,其状态矩阵中的每个元素取值为 0 或 1,表示对时频域中某个位置上语音信号存在与否的硬性判决.另一类是软掩模,其状态矩阵中的每个元素采用一个 0,1 之间的数值,表示语音出现概率(speech presence probability,简称 SPP).改进的最小控制递归平均(improved minima controlled averaging,简称 IMCRA)算法[7]采用一个二元的混合高斯模型(Gaussian mixture model,简称 GMM)对语音和非语音的信号功率谱分布建模,同步于在线估计过程得出语音出现概率.尽管 IMCRA 对状态在语音和非语音之间的转换是自适应的,其本质上的启发性体现在其中的一些参数是经验性最优地选取的,不能在统计上保证最优.约束的序贯化隐马尔可夫模型(constrained sequential hidden Markov model,简称 CSHMM)[8]将一个由语音和非语音信号组成的时域序列视为一个状态在语音和非语音缺失之间转换的动态过程进行建模.但是,CSHMM 对频域相关性的建模仅通过一个汉宁窗的平滑作用来实现[7,8].

传统的方法在一个频带上对语音信号的时间相关性建模表现得很好,但是没有给予频率相关性足够的重视.事实上,在掩模的估计过程中,频率相关性具有和时间相关性同样的重要性.本文提出的二维相关模型可以描述时频域中语音和非语音信号分布之间的时间相关性和频率相关性.该模型义了用来描述沿频率方向状态转移的频率状态转移概率,相应地提出了用来对频率相关性建模的邻域因子.同时,时间相关性通过时间状态转移概率和沿时间方向的前向因子表达.

有些方法[9,10]已被用于对语音信号中的二维的相关性进行建模.马尔可夫随机场(Markov random field,简称 MRF)[9]将时间相关性和频率相关性等同看待,这样的处理忽略了沿频率方向的功率谱不存在如同时间序列的前后顺序的实际情况.近年来,深度神经网络(deep neural network,简称 DNN)[2,3,10]也被用于掩模估计.但是,深度神经网络高度依赖大量的预训练数据.如果测试环境与预训练环境失配,系统性能将出现实质性的下降.

1 研究框架

众所周知,信号序列中的时间相关性已经得到了广泛的利用.而实际上,如同时间相关性,沿频率方向同样存在很高的相关性.图 1(a)描述了沿频率方向的相关性,其被表示为相邻位置上同时存在语音信号的概率.可以看出,即使频率间隔达到 5 个单位,语音信号的出现沿频率方向仍然存在很高的相关性.既然时间相关性与频率相关性相互交织,这样一个整体的相关性就可以由本文提出的二维的相关模型来表达.



(a) 频率相关性的统计结果 (b) 时间状态转移与频率状态转移的交织

Fig.1

图 1

该模型考虑一个处于 ℓ 时刻的对数功率谱 $\boldsymbol{X}_\ell=[\boldsymbol{x}_{\ell-L+1},...,\boldsymbol{x}_\ell]$,其中,$\boldsymbol{x}_t=[x_{1,t},...,x_{F,t}]^T$ 表示第 t 帧采样信号

对应的对数功率 ($\ell-L+1\leqslant t\leqslant \ell$).与 $\boldsymbol{X}_\ell$ 对应的状态矩阵用 $\boldsymbol{S}_\ell$ 表示,其中,在第(f,t)个时频点上的 $s_{f,t}$=1 与 $s_{f,t}$=0 分别表示语音和非语音两个状态.需要说明的是,对于带噪语音信号,语音表示语音信号与噪声信号叠加的情况,非语音表示不存在语音信号的情况.在频带 f 上,时间相关性由一个马尔可夫链建模,其中在任意时刻 t 上的似然度函数 $b\left(x_{f,t}\middle|s_{f,t}\right)$ 服从一个高斯分布.该马尔可夫链的时间状态转移概率 $\boldsymbol{a}_{f,\ell}$ 是一个 2×2 的矩阵,沿频率方向的状态转移由频率状态转移概率 $\boldsymbol{c}_{d,\ell}$ 描述,d 表示频率距离.

图 1(b)展示出该模型中时间状态转移概率和频率状态转移概率的共同作用,所有频带共享频率状态转移概率,模型的整个参数集为 Λ_ℓ 包含时间和频率状态转移概率、所有子带上高斯分布的均值和方差,全局的概率密度函数可以表示为 $p\left(\boldsymbol{X}_\ell\middle|\Lambda_\ell\right)=\sum_{\boldsymbol{s}_\ell}p\left(\boldsymbol{X}_\ell\middle|\boldsymbol{S}_\ell,\Lambda_\ell\right)p\left(\boldsymbol{S}_\ell\middle|\Lambda_\ell\right)$,其中,$p\left(\boldsymbol{X}_\ell\middle|\boldsymbol{S}_\ell,\Lambda_\ell\right)$ 表示给定状态矩阵 $\boldsymbol{S}_\ell$ 时的概率密度函数,$p\left(\boldsymbol{S}_\ell\middle|\Lambda_\ell\right)$ 是状态矩阵的概率.于是,建模问题被实现为按照极大似然(maximum likelihood,简称 ML)准则估计整体参数集.这一过程等价于以模型为基础的分类过程.

2 序贯化方法

本文采用最大化期望(expectation maximization,简称 EM)算法[11]估计模型的参数集,通过最大化 $p\left(\boldsymbol{X}_\ell\middle|\Lambda_\ell\right)$ 使参数集得到最优化估计.为了满足在线估计语音掩模的需要,本文还提出了一个逐帧更新模型的序贯化方案.该方案首先利用最初的 M 帧观察值,通过一种离线的 EM 算法初始化模型,然后逐帧更新模型参数集,同时估计语音掩模.

序贯化方案同样基于ML准则 $\Lambda_\ell=\max_{\Lambda}\log Q_{\ell|\Lambda_{\ell-1}}(\Lambda)$,其中Q-函数被定义为 $Q_{\ell|\Lambda_{\ell-1}}(\Lambda)=E\left\{\log p\left(\boldsymbol{X}_\ell,\boldsymbol{s}_\ell\middle|\Lambda_{\ell-1}\right)\right\}$.该 Q-函数被最大化的过程即按照牛顿迭代法[12,13]逐帧地搜索使似然度函数最大化的参数集,将参数集中的每一个元素带入迭代方程可以得出用于更新参数的递归过程.

用于更新均值和方差的递归过程是两个形式相近的线性递归公式,用于更新时间状态转移概率的递归过程是一个非线性递归公式,这 3 个递归过程类似于文献[8]中的递归过程.频率状态转移概率通过一个非线性递归过程更新,形式上类似于时间状态转移概率的更新.参数集的更新受到 3 个条件概率的控制.条件 SPP $\gamma_{f,\ell}(j)$、条件时间状态转移概率 $\xi_{f,\ell}(i,j)$ 和条件频率状态转移概率 $\phi_{d,\ell}(h,j)$,反映给定参数集时模型对语音和非语音状态出现的描述,它们的计算通过两个因子实现.假设模型随着时间缓慢变化,用于描述时间相关性的前向因子被定义为 $F_{f,\ell}(j)=\sum_i F_{f,\ell-1}(i)a_{f,\ell-1}(i,j)b_{f,\ell}(j)$,其中,$b_{f,\ell}(j)$ 是似然度函数的简写;用于对频率相关性建模的邻域因子被定义为 $\Psi_{f,\ell}(j)=\sum_{d=1}^{D}\sum_h b_{f+d,\ell}(h)c_{d,\ell-1}(h,j)$,其中,$b_{f+d,\ell}(j)$ 也是似然度函数的简写.经过建模,语音信号的软掩模由条件 SPP 表示,同时,作为二值掩模最优表达的 $\boldsymbol{S}_\ell$ 可以经过经典的维特比译码算法[14]给出.这两种掩模在本文提出的模型中得到了有机的结合.

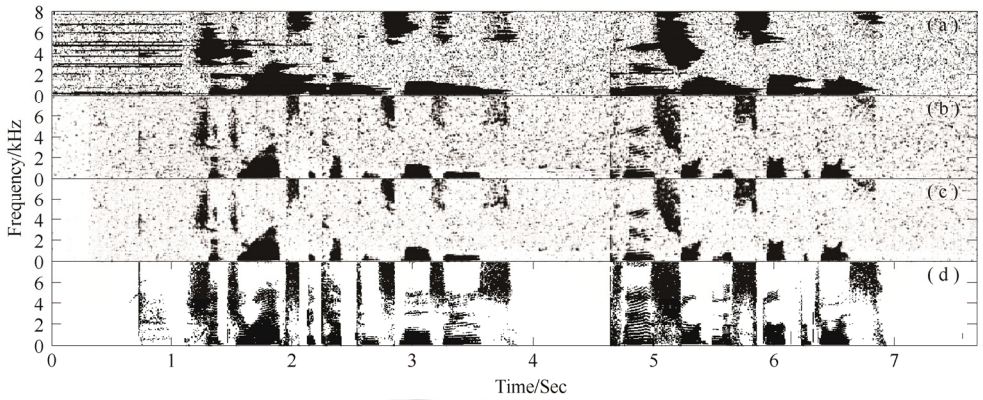
3 算法性能评价

实验选取 Noisex-92 噪声数据库[15]中的 white 噪声、F16 噪声以及 babble 噪声,纯净语音选自 TIMIT 数据库[16],所有的信号采样率均为 16kHz.测试数据集共 120 组(10 组语音×3 类噪声×4 个级别噪声),分别由经典的 IMCRA 算法[7]、CSHMM 算法[8]以及本文提出的算法进行处理.

首先采用一个带噪长句进行一项非正式的测试,将 3 种算法在 5dB 信噪比 white 噪声环境下得到的软掩模进行对比,如图 2 所示.软掩模中从白到黑的过渡对应于 SPP 从 0 到 1 的变化.SPP 的谱图可以反映出模型在保留可能的语音信号和剔除可能的噪声信号之间的权衡.可以看出,本文提出的方法有效地抑制了伪语音信号.

然后测量客观评价指标.利用模型估计的二值掩模提取出可被算法检测到的语音信号,对比被提取出的语音信号频谱和纯净语音信号的频谱得到的对数谱失真(logarithmic spectral distortion,简称 LSD)[17].LSD 结果越小,表示谱失真越小,算法的表现就越好.针对不同噪声类型和噪声水平,表 1 列出了 3 种算法处理后的评价结果.

算法性能的差异源于对语音信号在时频域内存在的时间相关性和频率相关性建模方法的区别.IMCRA 与 CSHMM 通过在相邻频带上的一个基于汉宁窗的平滑作用对频率相关性建模.因此,一旦相邻频率上信号幅度出现异常,模型对当前频带上语音信号状态的判断就会受到严重的影响.不同的是,本文提出的方法可以对频率相关性进行自适应的建模,相邻频率上信号幅度异常对建模的影响被控制在更低的水平上.因此,该建模方法给出的语音掩模更加清晰,实验结果较之 IMCRA 和 CSHMM 更优.



(a) IMCRA (b) CSHMM (c) 本文提出的算法 (d) 手工标记的纯净语音图谱

Fig.2 Comparison among speech masks estimated by different algorithms

图 2 不同算法估计的语音掩模对比

Table 1 Evaluation results under various conditions

表 1 系统计算评价结果

SNR (dB)	white 噪声			F16 噪声			babble 噪声		
	IMCRA	CSHMM	本文	IMCRA	CSHMM	本文	IMCRA	CSHMM	本文
–5	–0.41	3.08	5.42	–1.77	1.65	3.48	–0.36	–1.04	0.55
0	4.14	7.31	8.99	3.02	5.99	7.38	1.56	3.63	4.92
5	8.66	11.39	12.55	7.56	10.22	11.08	6.24	7.92	9.13
10	12.96	15.48	16.24	12.00	14.43	14.86	11.02	12.64	13.32

4 结束语

本文提出了一种用于估计语音掩模的二维相关模型.该模型采用序贯化方案逐帧更新,模型产生条件 SPP 作为软掩模,经过解码得到状态矩阵 $\boldsymbol{S}_\ell$ 作为最佳二值掩模.该模型的实验结果表明,其性能优于 IMCRA 与 CSHMM,原因在于其充分考虑了语音信号在时频域内出现与否的时频相关性.

References:

[1] Barker J, Josifovski L, Cooke M, Green P. Soft decisions in missing data techniques for robust automatic speech recognition. In: Guo DH, eds. Proc. of the 1st Annual Conf. of the Int'l Speech Communication Association (INTERSPEECH). 2000. 373−376.

[2] Narayanan A, Wang DL. Investigation of speech separation as a front-end for noise robust speech recognition. IEEE/ACM Trans. on Speech, Audio, and Language Processing, 2014,22(4):826−835.

[3] Cobos M, Lopez JJ. Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors. IEEE Trans. on Speech, Audio, and Language Processing, 2012,20(7):2059−2064.

[4] Huang PS. Deep learning for monaural speech separation. In: Proc. of the 39th IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). 2014. 1562−1566.

[5] Kjems U, Boldt JB, Pedersen MS. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. Journal of the Acoustical Society of America, 2009,126(3):1415−1426.

[6] Narayanan A, Wang DL. The role of binary mask patterns in automatic speech in background noise. Journal of the Acoustical Society of America, 2013,133(5):3083−3093.

[7] Cohen I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. on Speech Audio Processing, 2003,11(5):466−475.

[8] Ying D, Yan Y. Noise estimation using a constrained sequential hidden Markov model in the log-spectral domain. IEEE Trans. on Speech, Audio, and Language Processing, 2013,21(6):1145−1157.

[9] Andrianakis Y, White PR. A speech enhancement algorithm based on a chi MRF model of the speech STFT amplitudes. IEEE Trans. on Speech, Audio, and Language Processing, 2009,17(8):1508−1517.

[10] Li B, Sim KC. A spectral masking approach to noise-robust speech recognition using deep neural networks. IEEE/ACM Trans. on Speech, Audio, and Language Processing, 2014,22(8):1296−1305.

[11] Baum L, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics, 1970,41(1):164−171.

[12] Weinstein E, Feder M, Oppenheim A. Sequential algorithm for parameter estimation based on the Kullback-Leibler information measure. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1990,38(9):1652−1654.

[13] Krishnamurthy V, Moore J. On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure. IEEE Trans. on Signal Processing, 1993,41(8):2557−2573.

[14] Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. on Information Theory, 1967,13(2):260−269.

[15] Varga A, Steeneken H. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication, 1993,12(3):247−251.

[16] Garofolo JS. Getting started with the DARPA TIMIT CDROM: An acoustic phonetic continuous speech database. Gaithersburg: National Institute of Standards and Technology, 1988.

[17] Rabiner L, Juang BH. Fundamentals of Speech Recognition. Englewood Cliffs: Prentice-Hall Int'l, Inc., 1993.

战鸽(1990－),男,辽宁沈阳人,博士生,主要研究领域为语音信号处理,机器学习.

黄兆琼(1993－),女,博士生,主要研究领域为语音信号处理.

应冬文(1975－),男,博士,研究员,主要研究领域为语音信号处理.

潘接林(1965－),男,研究员,博士生导师,主要研究领域为语音信号处理,语音识别.

颜永红(1967－),男,博士,研究员,博士生导师,CCF 专业会员,主要研究领域为语音信号处理,语音识别,人机交互.