基于特征流融合的带噪语音检测算法

龙华，杨明亮，邵玉斌
（昆明理工大学信息工程与自动化学院，云南 昆明 650031）

摘 要：针对语音通话中语音段的起始检测性能不佳，检测语音连续性结构受到破坏的问题，提出了一种基于特征流融合的带噪语音检测算法。首先，根据语音特性分别提取时域特征流、谱图特征流和统计特征流；其次，利用不同的语音特征流分别对带噪音频中的语音段进行概率估测；最后，将各个特征流估测得到的语音估测概率进行加权融合，并利用隐马尔可夫模型对语音估测概率进行短时状态处理。通过对复合语音数据库在多类型噪声与不同信噪比条件下的性能测试表明，所提算法相对于基于贝叶斯与 DNN 分类器的基线模型相比，语音检测正确率分别提高了 21.26%与 11.01%，显著提高了目标语音的质量。

关键词：语音通话；语音检测；特征流融合；隐马尔可夫模型
中图分类号：TP391.42
文献标识码：A
doi:10.11959/j.issn.1000−436x.2020067

Noisy voice detection algorithm based on feature stream fusion

LONG Hua, YANG Mingliang, SHAO Yubin
Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650031, China

Abstract: Aiming at the problem that the initial detection performance of voice segment was poor, and the voice continuity structure was damaged in voice communication, a noisy voice detection algorithm based on feature stream fusion was proposed. Firstly, the time domain feature stream, the spectral pattern feature stream and the statistical feature stream were extracted according to the voice characteristics. Secondly, the voice segment in the noisy audio was estimated by different voice feature streams. Finally, the voice prediction probability obtained by each feature stream was weighted and fused, and the voice estimation probability was processed in short time by the hidden Markov model. The performance test of composite voice database under the condition of multi-type noise and different signal-to-noise ratio shows that compared with the baseline model based on Bayesian and DNN classifier, the voice detection accuracy of the proposed algorithm is improved by 21.26% and 11.01% respectively, and the quality of target voice is significantly improved.

Key words: voice communication, voice detection, feature stream fusion, hidden Markov model

1 引言

语音检测就是从带有背景噪声的音频中准确定位出语音的开始点和结束点，去除静音和纯噪声部分，提高语音信号的有效利用率，不同的应用场景对语音检测的要求有所不同。例如在语音通话系统中，为了提高语音分组转发的有效性（节约话路资源）与通话用户的舒适性（语句完整），从干扰环境下准确检测出语音段（即具有完整含义的句子，由若干短时语音帧组成）的起始位置且保证语音段不被检测割裂的语音检测技术就显得尤其重要，这是有别于语音检测中的关键词检测、词中的音节起始检测（英文多为多音节词）及语音特征提取前端的语音活动检测（VAD, voice activity detec-

收稿日期：2019−10−31；**修回日期**：2020−03−12
通信作者：杨明亮，yml19941122@163.com
基金项目：国家自然科学基金资助项目（No.61761025）
Foundation Item: The National Natural Science Foundation of China (No.61761025)

tion）等应用的。语音特征提取前端常使用的检测方法有双门限法[1]、谱熵法[2]等，其目的都是去除音频段中的静音段、纯背景噪声段以及表现为随机噪声特性的清音部分，进而获得音频信号中的浊音信号，而语音通话系统中完整语句的所有信息都是需要保留的。再者，语音通话系统中完整语句的检测对语音段中的加性噪声抑制处理（如谱减法[3]等）也是有利的（完整语句中保留着原始的噪声帧，便于信噪比估计）。

当前的语音检测可大致划分为基于阈值、基于分类器和基于模型的 VAD。基于阈值的 VAD 主要包括双门限法、谱熵法等，通过提取语音特征（短时能量、过零率、谱熵等）并设定判决门限，对静音段、清音浊音具有较好的效果，但在噪声较大的环境下则表现得无能为力。基于分类器的 VAD 则有基于网络框架的语音检测方法[4-5]及利用指数核函数构建语音检测模型[6]，将噪声和语音的帧特征作为分类数据进行目标训练和测试，在语音检测时，从带噪音频中检测出的语音丢掉了原始语句中应有的短时字间隔，严重降低了听众的舒适度。基于模型的 VAD 主要包括统计模型和算法模型。基于统计模型的语音检测包括文献[7]，以及在统计基础上构建谐波加噪声模型与最大后验概率相结合的语音检测模型[8]。算法模型则包括利用语音谐波检测技术用于语音检测[9]，同基于阈值的 VAD 一样，在面对复杂环境下，其检测性能也表现得不太好。

Shamma 等[10]分析并指出，语音流的形成主要取决于编码声源各种特征响应之间的时间一致性，即多条相干的特征流（一系列连贯实体/声音的内部特征，流强调了一个事实，声音特征同声音信号一样是随着时间而展开的。同一段语音的不同时刻所对应的同一特征在特性上是具有差异的，如声音的振幅是动态变化的，在数学上则表现为不同的数值序列即特征流）构成了一条与其他源非相干特征分离的流，这也是多通道语音可以利用不同通道之间的差分信息进行语音检测[11]而单通道行不通的原因。Teki 等[12]指出人类听觉系统对某些特征有显著敏感，这些特征是根据混合声音中小部分音频元素的时间重合而调整的，即当前所熟知的语音特征并不能完全表征语音的全部独特信息，这也是网络模型对反映语音特征的独特之处。

为了进一步提高语音通话中语音段起始检测的准确性以及避免语音段被检测割裂等问题，本文提出了基于特征流融合的带噪语音检测算法。利用神经网络的非线性拟合能力构建语音谱图特征与语音之间的映射关系，实现对语音的检测[13]并取得了一定的效果，而语音统计特征在工程中的成功应用已证实其表征语音信号的有效性，时域信号更是语音信号最直观的反映形式。由于利用单一特征进行语音检测性能不佳，因此首先对待检测语音提取时域特征流、谱图特征流及统计特征流并分别对带噪音频中的语音段进行估测后，然后对各个特征流估测得到的语音预测概率进行加权融合。过去语音检测从帧层级出发，忽略了语音连贯性特征，而高阶隐马尔可夫模型可以充分考虑过去的状态信息[14]，对滑动窗时长内的语音起着平滑的作用，保证了检测后语音段保持原始语句的连续性。

2　基于特征流的语音估测

2.1　时域特征流的语音估测

现实采集到的带噪语音可定义为

$$y(i)=s(i)+n(i) \tag{1}$$

其中，$y(i)$ 为观测到的音频信号，$s(i)$ 为纯语音信号，$n(i)$ 为噪声信号，i 为数据点号。每一帧信号的语音信号可分为以下 2 种状态

$$\begin{aligned} H_0&:\ \boldsymbol{y}_n=\boldsymbol{n}_n \\ H_1&:\ \boldsymbol{y}_n=\boldsymbol{s}_n+\boldsymbol{n}_n \end{aligned} \tag{2}$$

其中，下标 n 表示第 n 帧信号，每一帧信号中包含 M 个数据点，即 $\boldsymbol{y}_n=[y_{n,1},y_{n,2},\cdots,y_{n,M}]^{\mathrm{T}}$。这里假设噪声 $\boldsymbol{n}_n$ 的均值为 0，对角协方差矩阵为 $\boldsymbol{\delta}_n^2$。语音信号可以看作具有随机、周期和拟周期性质的非线性时间序列，那么依据谐波理论可将语音信号分解为 k 阶谐波[15]，即

$$s_{n,m}=\sum_{i=1}^{k_n}[\alpha_{n,i}\cos(w_n mi)+\beta_{n,i}\sin(w_n mi)] \tag{3}$$

其中，$s_{n,m}$ 表示第 n 帧第 m 个元素值，k_n 表示谐波阶数，$\alpha_{i,n}$ 与 $\beta_{i,n}$ 表示第 i 阶谐波的线性权值，$w_n=\dfrac{2\pi f_n}{f_{\mathrm{s}}}$ 为基频（弧度），f_{s} 为音频信号采样率。

进一步地，将语音信号用矩阵向量表示，并引入一个隐变量 h_n 用于表示音频帧信号中语音是否存在（其中，1 为语音，0 为非语音），可得

$$\boldsymbol{y}_n=h_n\boldsymbol{A}_n\boldsymbol{B}_n+\boldsymbol{n}_n \tag{4}$$

其中，

$$\boldsymbol{A}_n=\left[\boldsymbol{c}(w_n),\cdots,\boldsymbol{c}(w_nk_n),\boldsymbol{s}(w_n),\cdots,\boldsymbol{s}(w_nk_n)\right]$$
$$\boldsymbol{c}(w_n)=\left[\cos(w_n),\cos(w_n2),\cdots,\cos(w_nM)\right]^{\mathrm{T}}$$
$$\boldsymbol{s}(w_n)=\left[\sin(w_n),\sin(w_n2),\cdots,\sin(w_nM)\right]^{\mathrm{T}} \tag{5}$$
$$\boldsymbol{B}_n=\left[\alpha_{n1},\cdots,\alpha_{nk_n},\beta_{n1},\cdots,\beta_{nk_n}\right] \tag{6}$$

自此，一帧音频信号的语音状态可以由参数向量 $\boldsymbol{s}_n=[k_n,w_n,\boldsymbol{\delta}_n^2,\boldsymbol{A}_n,h_n]$ 表示，为了便于后述计算式的推导，此处令 $\boldsymbol{x}_n$ 的子集为 $\boldsymbol{R}_n=[w_n,k_n,h_n]$，通过观察当前及过去的音频信号 $\boldsymbol{Y}_n=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n]$ 来推断当前帧信号中为语音信号的概率

$$p(\boldsymbol{s}_n|\boldsymbol{Y}_n)=\frac{p(\boldsymbol{y}_n|\boldsymbol{s}_n,\boldsymbol{Y}_{n-1})p(\boldsymbol{s}_n|\boldsymbol{Y}_{n-1})}{p(\boldsymbol{y}_n|\boldsymbol{Y}_{n-1})},2\leqslant n\leqslant N$$
$$p(\boldsymbol{s}_1|\boldsymbol{Y}_1)=\frac{p(\boldsymbol{y}_1|\boldsymbol{s}_1)p(\boldsymbol{s}_1)}{p(\boldsymbol{y}_1)},\ n=1 \tag{7}$$

其中，N 为总帧数目。等式右边分子的第一项可由拉普拉斯近似算法表示，第二项可写成

$$p(\boldsymbol{s}_n|\boldsymbol{Y}_{n-1})=\sum_{\boldsymbol{s}_{n-1}}p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1})p(\boldsymbol{s}_{n-1}|\boldsymbol{Y}_{n-1})$$
$$p(\boldsymbol{s}_1)=0.5 \tag{8}$$

语音信号状态具有连贯性，带噪音频信号中的语音状态应满足隐马尔可夫模型，为了便于计算，这里假设时序帧的语音状态满足一阶隐马尔可夫，即 $p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1})=p(\boldsymbol{s}_n|\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_{n-1})$。第 n 帧音频信号的状态空间可进一步表示为

$$\boldsymbol{S}_0(n):[h_n=0]$$
$$\boldsymbol{S}_1(n):[w_n=w^f,k_n=k,h_n=1]^{\mathrm{T}},1\leqslant f\leqslant F,1\leqslant k\leqslant k^{\max} \tag{9}$$

语音帧推断后一帧为语音帧的条件概率为

$$p(\boldsymbol{S}_1(n)|\boldsymbol{S}_1(n-1))=$$
$$p(w_n,k_n|w_{n-1},h_{n-1}=1,h_n=1)p(h_n|h_{n-1}=1) \tag{10}$$

谐波阶数和基音频率无直接关系，进而可对等式右边第一部分进行条件联合分布分解

$$p(w_n,k_n|w_{n-1},k_{n-1},h_n=1,h_{n-1}=1)=$$
$$p(w_n|w_{n-1},h_n,h_{n-1}=1)p(k_n|k_{n-1},h_n,h_{n-1}=1) \tag{11}$$

当前一帧为非语音帧时，后一帧为语音帧的条件概率为

$$p(\boldsymbol{S}_1(n)|\boldsymbol{S}_0(n-1))=$$
$$p(w_n,k_n|h_n=1,h_{n-1}=0)p(h_n=1,h_{n-1}=0) \tag{12}$$

当前一帧为非语音帧时，选择离当前帧 n 最相近的过去语音帧 c 作为推断，根据文献[16]的式(18)对式(12)等号右边第一部分做条件联合分布分解，即

$$p(w_c,k_c|\boldsymbol{Y}_c,h_c=1)=\frac{p(w_c,k_c,h_c|\boldsymbol{Y}_c)}{1-p(h_c|\boldsymbol{Y}_c)} \tag{13}$$

联合式(9)~式(11)可得

$$p(\boldsymbol{S}_1(n)|\boldsymbol{Y}_{n-1})=$$
$$\sum_{S_1(n-1)}p(\boldsymbol{S}_1(n)|\boldsymbol{S}_1(n-1))p(\boldsymbol{S}_1(n-1)|\boldsymbol{Y}_{n-1})+$$
$$p(\boldsymbol{S}_1(n)|\boldsymbol{S}_0(n-1))p(\boldsymbol{S}_0(n-1)|\boldsymbol{Y}_{n-1}) \tag{14}$$
$$p(\boldsymbol{S}_0(n)|\boldsymbol{Y}_{n-1})=$$
$$\sum_{l=0}^{1}p(h_n=0|h_{n-1}=l)p(h_{n-1}=l|\boldsymbol{Y}_{n-1})=$$
$$p(h_n=0|h_{n-1}=0)p(\boldsymbol{S}_0(n-1)|\boldsymbol{Y}_{n-1})+$$
$$p(h_n=0|h_{n-1}=1)(1-p(\boldsymbol{S}_0(n-1)|\boldsymbol{Y}_{n-1})) \tag{15}$$

根据文献[16]的式(23)得到状态空间的后验概率为

$$p(\boldsymbol{R}_n|\boldsymbol{Y}_n)\propto\sum_{\boldsymbol{A}_n}\sum_{\boldsymbol{\delta}_n^2}p(\boldsymbol{y}_n|\boldsymbol{s}_n,\boldsymbol{Y}_{n-1})p(\boldsymbol{s}_n|\boldsymbol{Y}_{n-1})=$$
$$p(\boldsymbol{y}_n|\boldsymbol{R}_n,\boldsymbol{Y}_{n-1})p(\boldsymbol{R}_n|\boldsymbol{Y}_{n-1}) \tag{16}$$

其中，$p(\boldsymbol{y}_n|\boldsymbol{R}_n,\boldsymbol{Y}_{n-1})$ 表示边缘似然函数，采用拉普拉斯近似算法表示。

由此，按照文献[16]中的策略，根据式(10)、式(12)、式(14)和式(15)对状态空间做出概率预测，依据式(5)~式(8)和式(16)计算状态空间的后验概率，并联合文献[15-16]的快速基频估计和谐波阶数估计对 $p(\boldsymbol{S}_1(n)|\boldsymbol{Y}_n)$ 进行迭代更新，其约束条件为

$$p(\boldsymbol{S}_0(n)|\boldsymbol{Y}_n)+\sum_{\boldsymbol{S}_1(n)}p(\boldsymbol{S}_1(n)|\boldsymbol{Y}_n)=1 \tag{17}$$

2.2 谱图特征流的语音估测

因为短时功率谱是对按照时间序列展开的帧信号进行傅里叶变换得到的，所以谱图特征也是以“流”的形式出现的。时频谱图的频率分辨率为线性，而人类听觉系统对低频十分敏感，对高频就比较迟钝，为了解决频率分辨率的问题，本文利用 64 个 Gammatone 滤波器组提取 Cochleagram 特征[17-18]，并利用窗长为 32 ms、帧位移为 16 ms 的汉明窗口对其输出进行了瞬态积分。Gammatone 滤波器的脉冲响应 $h(t)$ 为

$$h(t)=gt^{a-1}\mathrm{e}^{-2\pi bt}\cos(2\pi ft),\ t>0 \tag{18}$$

其中，g 为输出增益；t 为时间；a 为滤波器阶数；b 为矩形带宽，它随中心频率 f 的增大而增大。

梅尔频谱倒谱系数（MFCC, Mel frequency cepstrum coefficient）特征是一种基于人耳对等距的音高变化的感官判断而定的非线性频率刻度，现已应用于语音识别、音乐检索等多个方面。Gabor 滤波器是一个频率和方向表达同人类视觉系统类似、用于边缘提取的线性滤波器，在图像的纹理表达和分离方面具有优异的性能。故利用 Gabor 滤波器对 MFCC 的纹理特征进行提取（详见文献[13]）。

长时信号变化特征（LTSV, long term signal variability）测量方法[19]在多个语音检测应用研究中证明，其在平稳噪声环境下具有较好的稳健性。首先，计算第 n 帧短时能量谱为

$$S_x(n,w_k)=\left|X(n,w_k)\right|^2$$

$$X(n,w_k)=\sum_{i=1}^{M}w(i)y_n(i)\mathrm{e}^{-\mathrm{j}w_k i} \tag{19}$$

其中，$X(n,w_k)$ 为第 n 帧信号 $\boldsymbol{y}_n$ 在频率 w_k 处的短时傅里叶变换，$w(\cdot)$ 为短时窗，M 为帧长。根据文献[19-20]对 LTSV 的定义，对每一帧的每个频率点进行熵的计算，有

$$L_x(m)\overset{\Delta}{=}\frac{1}{K}\sum_{k=1}^{K}\left(\xi_k^x(x)-\overline{\xi^x(m)}\right)^2 \tag{20}$$

$$\overline{\xi^x(m)}=\frac{1}{K}\sum_{k=1}^{K}\xi_k^x(m)$$

$$\xi_k^x(m)\overset{\Delta}{=}-\sum_{l=m-R+1}^{m}\frac{S_x(l,w_k)}{\sum_{i=m-R+1}^{m}S_x(i,w_k)}\log\left(\frac{S_x(l,w_k)}{\sum_{i=m-R+1}^{m}S_x(i,w_k)}\right) \tag{21}$$

其中，$m\geqslant R$，R 表示长时滑动窗包括的帧数目，这种滑动处理通过计算第 m 帧 K 个频点下熵的方程 $L_x(m)$，实现了对语音的长时分析。

语音具有一定的周期性，通过利用谐波理论估算带噪音频信号的基音周期，然后根据自相关函数计算相关图，为了克服吉布斯现象，此处在进行自相关函数计算时需要对帧信号进行加窗和预加重处理，其中窗长为 32 ms，帧移为 16 ms。随后，每次都推导出一个语音测量值，即信号能量归一化估计基音周期时的自相关值，为了考虑前后帧之间的相关性，对一维预测概率做了前后扩张，最终取五阶自相关值。

2.3　统计特征流的语音估测

语音发音通常是由清音和浊音交叉构成的，而非语音却不满足这样的构造特点，清音的过零率远高于浊音，故语音过零率的变化一般要比非语音激烈。高过零率比率（HZCRR, high zero crossing rate ratio）用于描述一段音频段过零率变化的剧烈程度，计算式为[21]

$$\mathrm{HZCRR}=\frac{1}{2N}\sum_{n=1}^{N}[\mathrm{sgn}(\mathrm{ZCR}(n)-1.5\overline{\mathrm{ZCR}})+1]$$

$$\overline{\mathrm{ZCR}}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{ZCR}(n) \tag{22}$$

其中，N 为帧数目，n 为帧索引，sgn 为符号函数，ZCR(n)为第 n 帧的过零率，$\overline{\mathrm{ZCR}}$ 为 ZCR 均值。为了避免一些非语音段能量较低，但过零率比率高于所设阈值 0.08 的误判，本文加入式(23)所示的短时能量特征共同进行判断。

$$\mathrm{STE}(n)=\sum_{i=1}^{M}y_n^2(i) \tag{23}$$

其中，符号 i 为帧内音频数据点索引，M 为帧长，y_n 为第 n 帧归一化信号。此处阈值设置为 0.05。

3　特征流的融合

基于特征流融合的带噪语音估测流程如图 1 所示。

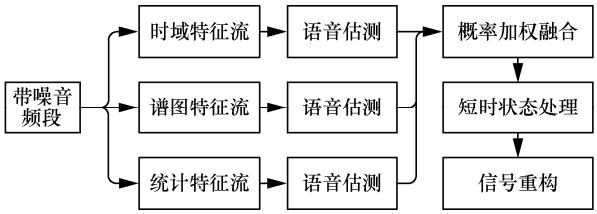

图 1　基于特征流融合的带噪语音估测流程

为了减少估测语音概率的复杂性，在 2.1 节整个初步估计中假设状态空间为一阶隐马尔可夫模型，只考虑前一帧的影响，忽略语音长时帧的相关信息，而事实上，可将基于特征流对每帧音频预测的语音概率值序列看作离散平稳有记忆信源 $X:\{a_1,a_2,\cdots,a_r\}$，在任何时刻 t_{m+1}，随机变量 X_{m+1} 所发符号 a_{im+1} 通过其前 m 个符号 $(a_{i1},a_{i2},\cdots,a_{im})$ 进而与更前面的符号发生联系[22]，即 a_{im+1} 只与它前面的 m 个符号相关，与更前面的符号无关。假设每一个短时状态由 m 个信源构成，而这 m 个符号取遍信源 X 的符号集，m-M 信源共有 r^m 种不同的消息，

令 $\boldsymbol{S}_i$ 为 $(i=1,2,\cdots,r^m)$ 某一状态，则有

$$\begin{aligned}&\boldsymbol{S}_i=(a_{i1},a_{i2},\cdots,a_{im})\\&a_{i1},a_{i2},\cdots,a_{im}\in X:\{a_1,a_2,\cdots,a_r\}\\&i1,i2,\cdots,im=1,2,\cdots,r\\&i=1,2,\cdots,r^m\end{aligned}\tag{24}$$

对应于带噪音频帧的语音检测只存在 2 种状态，即信源 X 的符号集 $X:\{0,1\}$，$i1,i2,\cdots,im$ 为符号状态序列号。因为本文分帧帧长为 32 ms，帧移为 16 ms，又考虑到正常语速 7 个音节需要 3 s，所以此处 m=39，即 39 阶马尔可夫信源，由式(24)即有

$$\begin{aligned}&\boldsymbol{S}_i=(a_{i1},a_{i2},\cdots,a_{i39})\\&a_{i1},a_{i2},\cdots,a_{i39}\in X:\{0,1\}\\&i1,i2,\cdots,i39=1,2\\&i=1,2,\cdots,2^{39}\end{aligned}\tag{25}$$

符号序列号为 1 时对应符号集中的符号 0，符号序列号为 0 时对应符号集中的符号 1，进一步即有

$$\begin{aligned}&\boldsymbol{S}_1=(\underbrace{00\cdots0}_{39}),\quad \boldsymbol{S}_2=(\underbrace{00\cdots1}_{39}),\cdots,\\&\boldsymbol{S}_{2^{39}-1}=(\underbrace{11\cdots0}_{39}),\quad \boldsymbol{S}_{2^{39}}=(\underbrace{11\cdots1}_{39})\end{aligned}\tag{26}$$

将预测语音概率值大于 0.5 当作状态 1，其余的为状态 0，为了对语音概率曲线做平滑处理，本文统计了每一个短时状态中 0 与 1 的个数，令 1 的个数大于 20 且小于 35（设定个的判决阈值）的短时状态 $\boldsymbol{S}_i$ 内的符号状态值均为 1，其他的短时状态符号值均为 0。为了保证最终重构语音舒适度和可理解性，短时状态时序步长取 11 帧。

令时域特征流的输出概率为 prob_1，经过短时状态处理的时域特征流的输出概率为 prob_1'（0-1 序列），谱图特征流输出的概率为 prob_2，统计特征流的输出概率为 prob_3。为了突出 prob_1' 的语音部分，即有

$$p_1=\text{prob}_1\text{prob}_1'\tag{27}$$

将 p_1 与谱图特征流的语音概率融合并利用统计特征流 prob_3 突出谱图特征流 prob_2 的语音部分，与此同时将 prob_1 与 prob_3 加权融合，则有

$$p_2=p_1q_1+\text{prob}_2(1-q_1)\tag{28}$$

$$p_3=\text{prob}_2\text{prob}_3\tag{29}$$

$$p_4=\text{prob}_1q_2+\text{prob}_3(1-q_2)\tag{30}$$

最后，将两两特征流融合结果再次加权融合，有

$$p_5=p_2q_3+p_4(1-q_3)\tag{31}$$

其中，q_1、q_2、q_3 分别为不同特征流语音估测的权重系数。根据人类听觉系统自身的特殊特性，以及对某些特征更加敏感的理论[13]，对权值进行猜测和检验（如同深度学习模型中卷积核大小的选择一样）。经实验测试推断，谱特征流包含有更多区分语音和噪声的有用信息，因此取 $q_1=0.35$，$q_2=0.40$，$q_3=0.35$。最后，利用隐马尔可夫短期状态对语音概率曲线进行平滑处理，得到最终的语音估测 0-1 折线 p_5（其中，1 为语音，0 为非语音）。

4 实验设计与结果分析

4.1 实验设置

1) 数据准备

实验测试检验中，本文采用的语音库为 TIMIT 语音库、THCHS30 语音库、2018 方言种类识别 AI 挑战赛（DRC, dialect recognition contest）语音库 3 种，语音为采样率 f_s=16 000 Hz、单通道的 wav 音频文件，每句语音时长在 4~7 s 左右。语音训练数据集包含 TIMIT 语音库中美国 8 个地区 462 个说话人语音，每人 10 句英语语音；THCHS30 语音库包含 11 个来自中国各地的说话人语音，每人 20 句普通话；DRC 语音库包含中国 10 个方言地区的 300 个说话人，每人 10 句中国方言。噪声训练数据集采用自建噪声库（SBNL, self-built noise library），其中包括动物鸣叫、公共场所噪声、户外活动噪声、室内活动噪声、工厂噪声、音乐、自然音效、交通噪声共 8 类噪声类型，每一类噪声又包含 10 段不同的噪声段。从语音训练数据集和噪声训练数据集中随机选取一种语音和噪声，随机选择 SNR=[−5, 0, 5, 10, 15, 20] dB 中的信噪比进行带噪语音合成，建立多条件训练集用于训练基于谱图特征流的语音概率预测深度神经网络（DNN, deep neural network）模型。

实验测试语音数据包含 TIMIT 语音库中 168 个说话人（女性 52 人，男性 114 人，每人 10 句）、THCHS30 语音库中 20 个说话人（女性 18 人，男性 2 人，每人 10 句）、DRC 语音库中 50 个说话人（女性 30 人，男性 20 人，每人 10 句）。为了更加贴近真实环境，在原始语音数据基础上的语音段前后随机补充 2~4 s 的静音段，以便与噪声混合成的带噪音频更符合现实情况，噪声库采用 Nonspeech 公开噪声库，合成语音的信噪比等级 SNR=[−5, 0, 5,

10] dB。测试语音库中的语音段与噪声库中随机选取的噪声依次按照–5 dB、0 dB、5 dB、10 dB 这 4 个信噪比等级合成 4 个不同信噪比的测试数据库。语音库的分配设置如表 1 所示。

表 1　　语音库的分配设置

数据集	语音数据/句			噪声数据/句	
	TIMIT	THCHS30	DRC	SBNL	Nonspeech
训练集	4 620	220	3 000	80	0
测试集	1 680	200	500	0	100

2) 模型设置与训练

基于谱图特征流的语音估测部分，将提取的 64 维 Cochleagram 特征、110 维 Gabor 特征、5 维 LTSV 特征以及 5 维基音周期自相关值构建成 184 维特征流，再将特征流送入 3 层网络[13]、节点数为 184→64→64→1 的 DNN 中进行训练（迭代），得到一个谱图特征流的语音概率预测模型。其余按照第 3 节特征流融合策略进行构建系统。

4.2　实验测试与分析

为了分析不同程度噪声对带噪语音估测性能的影响，用 4 组不同信噪等级的音频测试库对不同语音检测方法进行性能测试。以基于贝叶斯特征流（BFS, Bayesian feature stream）[16]和光谱特征流（SCS, spectral characteristic stream）[20]的带噪音频语音概率估测作为本文特征流融合（FSF, feature stream fusion）的基线模型。因为 HZCRR 在干净环境下能够准确地对语音段进行估测，故以 2.3 节的统计特征流作为原始干净环境的语音检测方法，其语音检测结果作为音频段语音状态的标准标注。以误检率（Pf, false-alarm probability）、漏检率（Pm, miss probability）、正确率（Pc, correct probability）作为性能评价指标（其中，Pf 和 Pm 越小越好，Pc 越大越好），则有

$$\mathrm{Pf}=\frac{T_{\text{误检}}}{T},\quad \mathrm{Pm}=\frac{T_{\text{漏检}}}{T},\quad \mathrm{Pc}=\frac{T_{\text{正确}}}{T} \tag{32}$$

其中，T 为总音频长度，$T_{\text{误检}}$ 为误检为语音的音频长度，$T_{\text{漏检}}$ 为漏检语音的音频长度。不同信噪比条件下的语音估测误检率、漏检率和正确率分别如表 2~表 4 所示。

由表 2~表 4 可知，对于 4 个信噪比等级下的带噪音频段语音检测性能，除 SNR=5 dB 和 SNR=10 dB 条件下 BFS 的 Pf 指标略优于本文所提的 FSF 方法外，其余本文所提 FSF 方法的性能指标都显著优于对比实验模型，根据表 1~表 4 可得，不同方法的平均语音检测性能如表 5 所示。

表 2　不同信噪比条件下的语音估测误检率

检测方法	SNR=–5 dB	SNR=0 dB	SNR=5 dB	SNR=10 dB
BFS	0.206 4	0.197 2	**0.162 1**	**0.118 3**
SCS	0.219 4	0.198 4	0.165 3	0.153 8
FSF	**0.194 1**	**0.185 8**	0.175 1	0.122 2

表 3　不同信噪比条件下的语音估测漏检率

检测方法	SNR=–5 dB	SNR=0 dB	SNR=5 dB	SNR=10 dB
BFS	0.194 7	0.175 9	0.186 7	0.193 6
SCS	0.127 6	0.116 9	0.110 6	0.106 1
FSF	**0.094 0**	**0.057 3**	**0.032 9**	**0.028 1**

表 4　不同信噪比条件下的语音估测正确率

检测方法	SNR=–5 dB	SNR=0 dB	SNR=5 dB	SNR=10 dB
BFS	0.598 9	0.626 9	0.651 2	0.688 1
SCS	0.653 0	0.684 7	0.724 1	0.740 1
FSF	**0.711 9**	**0.756 8**	**0.792 0**	**0.849 7**

表 5　　不同方法的平均语音检测性能

检测方法	Pf	Pm	Pc
BFS	0.171 050	0.187 725	0.641 275
SCS	0.184 225	0.115 300	0.700 475
FSF	**0.169 300**	**0.053 0475**	**0.777 600**

对表 5 分析可知，相对于 BFS，FSF 的 Pf、Pm、Pc 评价指标分别提高了 1.023%、71.73%、21.26%；相对于 SCS，FSF 的 Pf、Pm、Pc 评价指标分别提高了 8.10%、53.97%、11.01%。

为了更直观地展示不同语音检测方法在不同信噪比条件下的检测性能，从 4 个不同信噪比等级测试库中挑选不同音频进行性能可视化展示，分别如图 2~图 5 所示。

由图 2 可知，当 SNR=–5dB 时，语音完全被噪声所覆盖掉，无法从时频图中观察出语音信号的清晰脉络。在此条件下，BFS、SCS、FSF 都出现了不同程度的误检，其中 SCS 的误检率最大，其次是 SCS，BFS 的误检最少。漏检方面，FSF 的漏检最多，与总体测试结果中 FSF 的漏检率最低的结论有所差异，但 FSF 的检测正确率大于 BFS 和 SCS 的检测正确率。由于噪声环境的复杂性，出现了使 FSF 对个别音频段的语音检测性能低于 BFS 或 SCS 的

情况。第 1 节中已指出，对于语音通话中的语音检测，除了要求准确检测出完整语句的起始位置外，还需要避免完整语句被检测割裂的问题（保证听者的舒适度），另一方面也是语音增强的需要。例如，图 2 中 BFS 的语音检测结果以单个字或词的形式出现，这既不是语音通话中所期待的，也不利于检测语音段后续的语音增强（非平稳噪声很难被准确估计）。对于 SCS 语音检测，其检测结果大部分都是连贯的，但在大约 t=14 s 处也出现了检测波动（导致音乐噪声的产生），而 FSF 由于进行了高阶隐马尔可夫处理，保证了检测出的语音段的连续性。

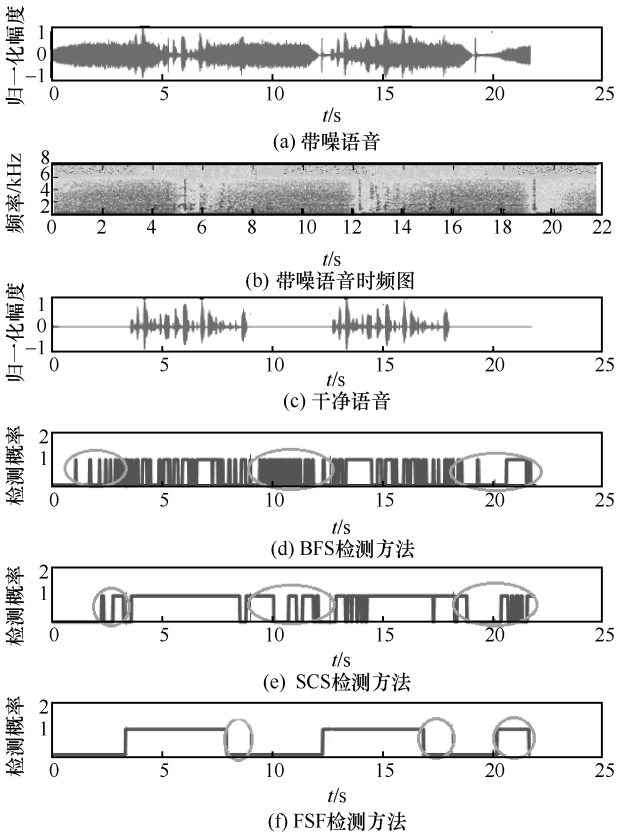

图 2　不同语音检测方法在 SNR=−5 dB 时的检测性能

将图 2 与图 3 对比可知，在信噪比提高的条件下，BFS、SCS、FSF 的检测性能都得到提高，特别是 FSF 的误检和漏检大幅度降低，但 BFS 和 SCS 仍然出现较多的误检。此外，在信噪比提高的情况下，SCS 检测的语音段连贯性得到较大改善（t=14 s 处），但对于 BFS 检测结果改善并不明显。

图 4 和图 5 展示了 BFS、SCS、FSF 在噪声干扰环境 SNR=5 dB 和 SNR=10 dB 时的语音检测性能。从图 4 可知，BFS、SCS 和 FSF 仍然存在误检

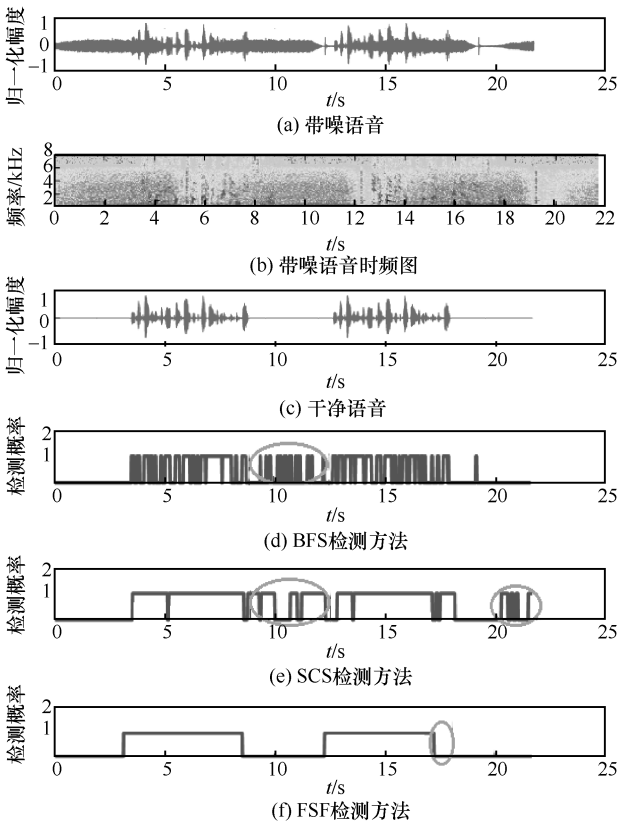

图 3　不同语音检测方法在 SNR=0 dB 时的检测性能

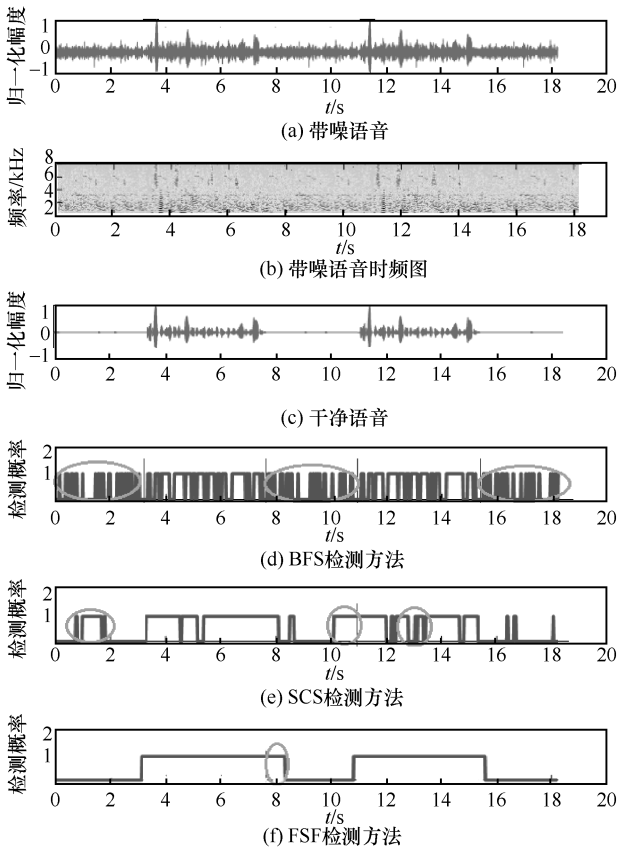

图 4　不同语音检测方法在 SNR=5 dB 时的检测性能

的情况，BFS 和 SCS 的语音检测中还存在漏检的情况，而 FSF 不存在漏检。图 4 与图 2、图 3 所展示的结果一样，BFS 和 SCS 的语音检测由于从帧层面进行语音检测判定，并未进行短时状态考虑，在语音检测时对语音段内的字间间隔进行了移除，使检测出的语音段连贯性受到破坏，严重影响着听者的舒适度，也制约着检测语音段后期的语音增强性能。

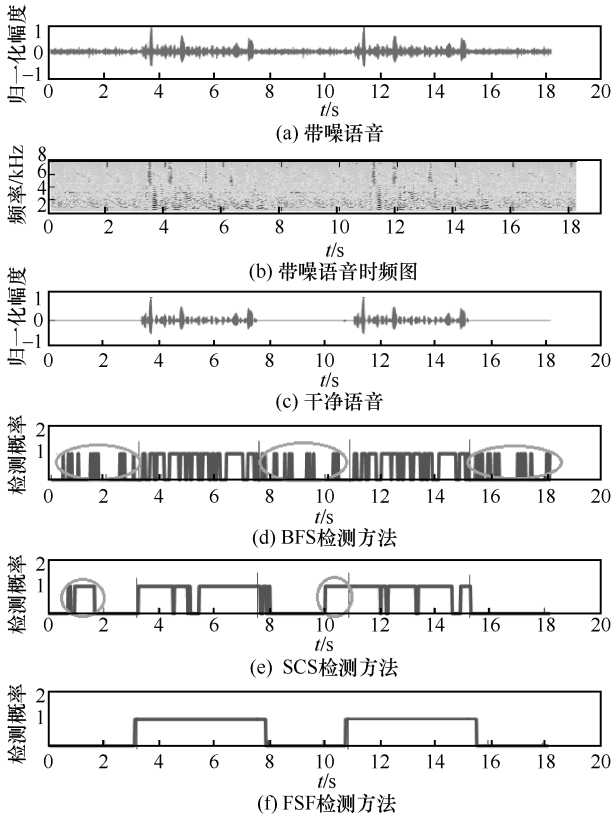

图 5　不同语音检测方法在 SNR=10 dB 时的检测性能

将图 4 与图 5 对比可知，当信噪比提高时，在图 4 中被误检或漏检的部分也得以被正确检测出，但对于某些部分依旧无法正确检出，特别是 BFS 和 SCS 的语音检测方法。将图 5 进一步与图 2、图 3 对比可知，在一定的信噪比范围内，影响语音检测性能的并非只有信噪比指标，噪声类型也是干扰语音检测性能的重要因素。从图 5 语音检测的可视化结果中可知，此时 FSF 可在尽可能去除静音段和纯噪声段的同时，避免完整语音段被检测割裂的问题。

5　结束语

为了进一步提高语音通话中语音段起始检测的准确性及避免语音段被检测割裂等问题，本文提出了基于特征流融合的带噪语音检测算法。相比于基于单纯的时域特征流或单纯的谱图特征流（DNN 模型训练检测），所提算法在不同信噪比情况下的带噪语音检测性能（误检率、漏检率和检测正确率）都有了较大的提高。这主要归功于所提算法将多种特征流进行了融合，相对于利用单特征（如 HZCRR 统计特征等）进行语音检测方法，增大了语音检测的运算力。因为利用了高阶隐马尔可夫模型的多状态考虑能力即对语音估测结果进行了短时处理，使经过 FSF 语音检测方法的语音段保持原始的连贯性（即具有完整含义的句子）。进一步提高语音检测抗噪性能仍是未来的目标（同等数值漏检率比误检率更具破坏性，少量的误检率可以通过语音增强来进一步消除，而漏检率会破坏原始语句连续性结构）。影响语音检测准确性的因素有多种，其中包括噪声类型、噪声强度、个人习惯等，后期可发展自适应隐马尔可夫或使用深度学习模型来代替隐马尔可夫的作用，进而提高检测语音段的完整性和连贯性，通过构建多条件的训练数据集提高语音检测模型的稳健性。

参考文献：

[1] WEI J, SUN X. Research on speech endpoint detection algorithm with low SNR[J]. Open Access Library Journal, 2017, 4(3): 1-8.

[2] LI Q, XIE H E, ZHENG Q J, et al. The voice activity detection algorithm based on spectral entropy and high-order statistics[J]. Applied Mechanics and Materials, 2014(624): 495-499.

[3] PALIWAL K K, WOJCICKI K, SCHWERIN B, et al. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain[J]. Speech Communication, 2010, 52(5): 450-475.

[4] MENDELEV V, PRISYACH T, PRUDNIKOV A, et al. Robust voice activity detection with deep maxout neural networks[J]. Mathematical Models and Methods in Applied Sciences, 2015, 9(8): 153-159.

[5] LENGERICH C, HANNUN A. An end-to-end architecture for keyword spotting and voice activity detection[J]. arXiv Preprint, arXiv: 1611.09405, 2016.

[6] IVRY A, BERDUGO B, COHEN I, et al. Voice activity detection for transient noisy environment based on diffusion nets[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 254-264.

[7] SOHN J, KIM N S, SUNG W, et al. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999, 6(1): 1-3.

[8] FISHER E, TABRIKIAN J, DUBNOV S, et al. Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(2): 502-510.

[9] KIM J T, JUNG S H, CHO K H, et al. Efficient harmonic peak detection of vowel sounds for enhanced voice activity detection[J]. Let Signal Processing, 2018, 12(8): 975-982.

[10] SHAMMA S A, ELHILALI M, MICHEYL C, et al. Temporal cohe-

rence and attention in auditory scene analysis[J]. Trends in Neurosciences, 2011, 34(3): 114-123.

[11] HWANG S, JIN Y G, SHIN J W, et al. Dual microphone voice activity detection based on reliable spatial cues[J]. Sensors, 2019, 19(14): 3056-3064.

[12] TEKI S, BARASCUD N, PICARD S, et al. Neural correlates of auditory figure-ground segregation based on temporal coherence[J]. Cerebral Cortex, 2016, 26(9): 3669-3680.

[13] KLEINSCHMIDT M. Spectro-temporal Gabor features as a front end for automatic speech recognition[C]//In 3rd European Congress on Acoustics. [S.n.:s.l.], 2002: 1-6.

[14] 罗智勇, 杨旭, 孙广路, 等. 基于马尔可夫的有限自动机入侵容忍系统模型[J]. 通信学报, 2019, 40(10): 79-89.
LUO Z Y, YANG X, SUN G L, et al. Finite automaton intrusion tolerance system model based on Markov[J]. Journal on Communications, 2019, 40(10): 79-89.

[15] NIELSEN J K, JENSEN T L, JENSEN J R, et al. Fast fundamental frequency estimation: making a statistically efficient estimator computationally efficient[J]. Signal Process, 2017(135): 188-197.

[16] SHE L M, NIELSEN J K, JENSEN J R, et al. Bayesian pitch tracking based on the harmonic model[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1737-1751.

[17] SCHLUTER R, BEZRUKOV L, WANGNER H, et al. Gammatone features and feature combination for large vocabulary speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2007: 649-652.

[18] SHAO Y, JIN Z, WANG D L, et al. An auditory-based feature for robust speech recognition[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2009: 4625-4628.

[19] MAARTEN V S, ANDREAS T, NARAYANAN S. A robust front end for VAD: exploiting contextual, discriminative and spectral cues of human voice[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2013: 704-708.

[20] MAARTEN V S, ANDREAS T, NARAYANAN. A robust front end for VAD: exploiting contextual, discriminative and spectral cues of human voice[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2013: 704-708.

[21] LU L, JIANG H, ZHANG H, et al. A robust audio classification and segmentation method[C]//ACM Multimedia. New York: ACM Press, 2001: 203-211.

[22] ZHAO J H, GAO H B, LIU Y C, et al. Speech recognition algorithm based on neural network and hidden Markov model[J].The Journal of China Universities of Posts and Telecommunications, 2018, 25(4): 28-37.

[作者简介]

龙华（1963− ），女，回族，云南大理人，博士，昆明理工大学教授，主要研究方向为无线网络及音频信号处理。

杨明亮（1994− ），男，四川宜宾人，昆明理工大学硕士生，主要研究方向为音频信号处理、语音识别。

邵玉斌（1970− ），男，云南曲靖人，昆明理工大学教授，主要研究方向为移动通信和个人通信系统以及信号处理。