应用声学
第 38 卷 第 1 期
2019 年 1 月
Journal of Applied Acoustics
Vol. 38, No. 1
January, 2019

◇ 研究报告 ◇

采用骨导语音自适应的语句分割方法*

苗晓孔† 张雄伟

(陆军工程大学指挥控制工程学院 南京 210007)

摘要 为了解决含噪语句分割问题，也为了解决某些低信噪比环境下传统气导语句分割算法分割效果差、分割准确度低且算法自适应性弱等问题，提出一种基于骨导语音自适应的分段双门限语音分割方法。将骨导语音和气导语音同步采集，获取抗噪性能更好的骨导语音，然后在融合过零率与短时能量中引入随机动态阈值的自适应方法进行端点检测，最后利用分段双门限和语音聚类等手段实现语音分割，提高语音分割算法的鲁棒性。通过实验验证了所提算法的有效性和可行性，同时与其他语音分割算法进行了对比，证明该文所提分割算法精度更高，效果更好。

关键词 骨导语音，语音分割，分段双门限，语音聚类

中图法分类号: TP391 **文献标识码:** A **文章编号:** 1000-310X(2019)01-0068-08

DOI: 10.11684/j.issn.1000-310X.2019.01.010

The adaptive speech segmentation method based on bone conduction voice

MIAO Xiaokong ZHANG Xiongwei

(*Command & Control Engineering College, Army Engineering University, Nanjing* 210007, *China*)

Abstract In order to solve the problem of segmentation of noisy sentences, and to solve the problems of poor segmentation efficiency, low segmentation accuracy and poor adaptive ability of traditional air-guided speech segmentation algorithm in some low SNR environments, a segmentation two-threshold speech segmentation method based on bone conduction speech adaptation is proposed. Firstly, bone-guided speech and air-guided speech are acquired synchronously to obtain better anti-noise performance. Then an adaptive method of random dynamic threshold is introduced to detect endpoints in the fusion of zero-crossing rate and short-term energy. Finally, segmentation double threshold and speech clustering are used to realize sentence segmentation and improve the robustness of speech segmentation algorithm. The effectiveness and feasibility of the proposed algorithm are verified by experiments. At the same time, compared with other speech segmentation algorithms, the proposed segmentation algorithm is proved to be more accurate and effective.

Key words Bone guided speech, Speech segmentation, Segmented double threshold, Speech clustering

2018-03-18 收稿; 2018-07-13 定稿

*国家自然科学基金项目 (61471394)

作者简介: 苗晓孔 (1991-), 男, 河北石家庄人, 博士, 研究方向: 智能信息处理。

†通讯作者 E-mail: miao_xk@163.com

0 引言

近些年随着神经网络、机器学习等技术在语音智能等方面的运用，语音数据库制作也受到关注。语音数据库可用来帮助训练和改善语音处理算法，为了丰富语音数据库内容，同步录制包含周围环境噪声的语音数据也逐步得到重视。含噪语音可以用来检验相关语音算法在不同真实场景中的处理效果。而语音分割技术就是将不同情况下的连续语句进行分割、提取，以制取完备的语音数据库。针对含噪语音或者某些低信噪比环境下的语音数据分割，高效、鲁棒的分割算法对提升语音转换、语音识别、语音截取 [1] 等技术的性能将起到一定的作用。

语音分割关键是准确得到语句起始和结束端点，按其端点检测方式目前语音分割方法大致可分三类：(1) 基于特征参数提取的分割方法：主要是提取语音信息中的时频特征参数进行端点检测，然后分割。时域特征如过零率、短时能量以及自相关函数等 [2−3]；频域特征主要有梅尔倒谱距离、频率方差、LPC 以及谱熵等 [4−7]。这类算法操作简单，便于实现，但算法鲁棒性差，在低信噪比环境适应效果不理想。(2) 基于模型的分割方法：通常是将端点检测问题转化为分帧问题，分别对噪声和语音进行二分类建模，然后检测语音端点并分割。常用模型有隐马尔科夫模型 (Hidden Markov models, HMM)、支持向量机 (Support vector machine, SVM)、深度神经网络 (Deep neural network, DNN)[8−9] 等。这些算法比较复杂，其分割效果取决于模型与环境噪声的匹配程度，匹配度越高效果越好，所以其依赖性较强，适应性较差。(3) 基于一些新理论的方法：运用混沌理论、分形理论的端点检测分割算法。这些算法的运算量大，只适用于一些特殊噪声，具有一定的局限性 [10]。

针对上述分割算法存在的问题，本文提出了基于骨导语音的自适应分段双门限语句分割方法。首先利用骨导语音的抗噪性提升时域参数特征融合算法鲁棒性 (因骨导语音通过捕获振动源的机械振动获取语音，去除了周围环境噪声影响，且骨导设备廉价易得，可操作性强)，然后引入随机动态阈值进行自适应的端点检测，最后通过分段双门限和层聚类的方式实现语音分割。实验证明：本文所提分割算法提高了语音分割的精度和准确度，算法适应性强，鲁棒性好且便于实现。同时与其他几种算法对比，本文分割算法的分割效果获得明显改善。

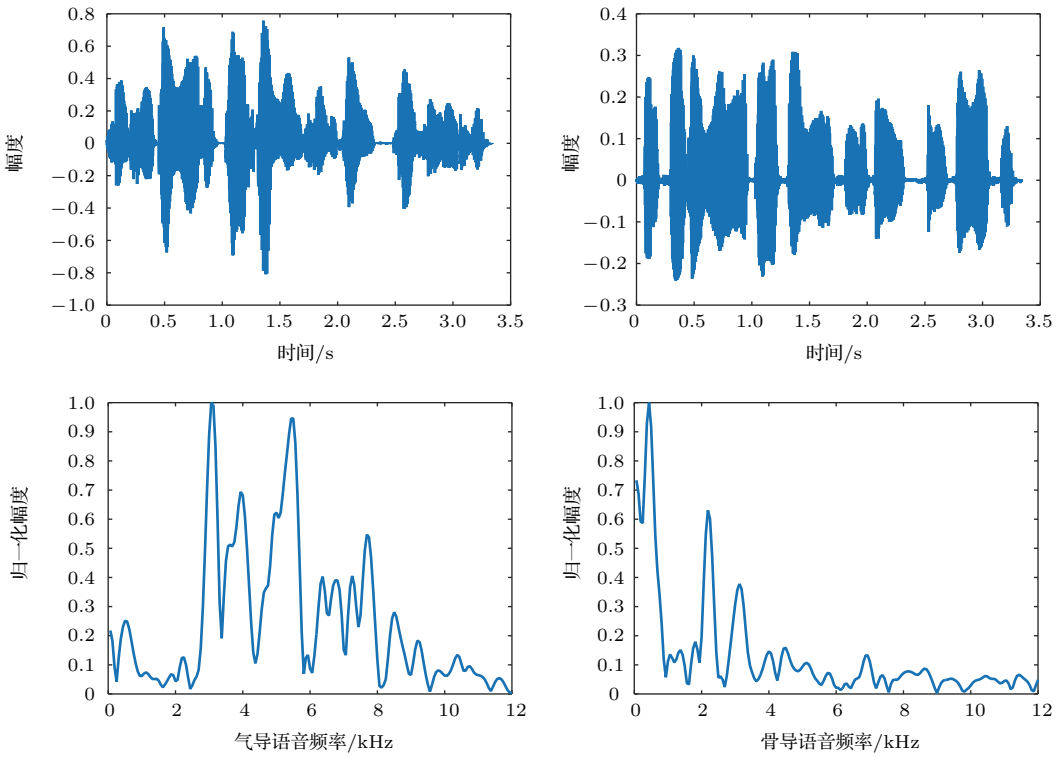
1 改进预处理方法

传统的时域参数融合分割方法，在语音预处理阶段主要是对气导语音进行信号预加重加窗分帧，通过预处理提升语音信号的信噪比。但是大多数情况下采集到的气导语音信号含有噪声，对受到不同噪声影响的气导语音进行分割，需要考虑不同的去噪方法，这使得算法的适应性降低。本文提出在预处理阶段引入骨导语音，利用骨导语音的低频抗噪性来提升算法的适应性，通过对骨导语音简单的噪声滤波，减少去噪复杂度进而实现鲁棒的端点检测。

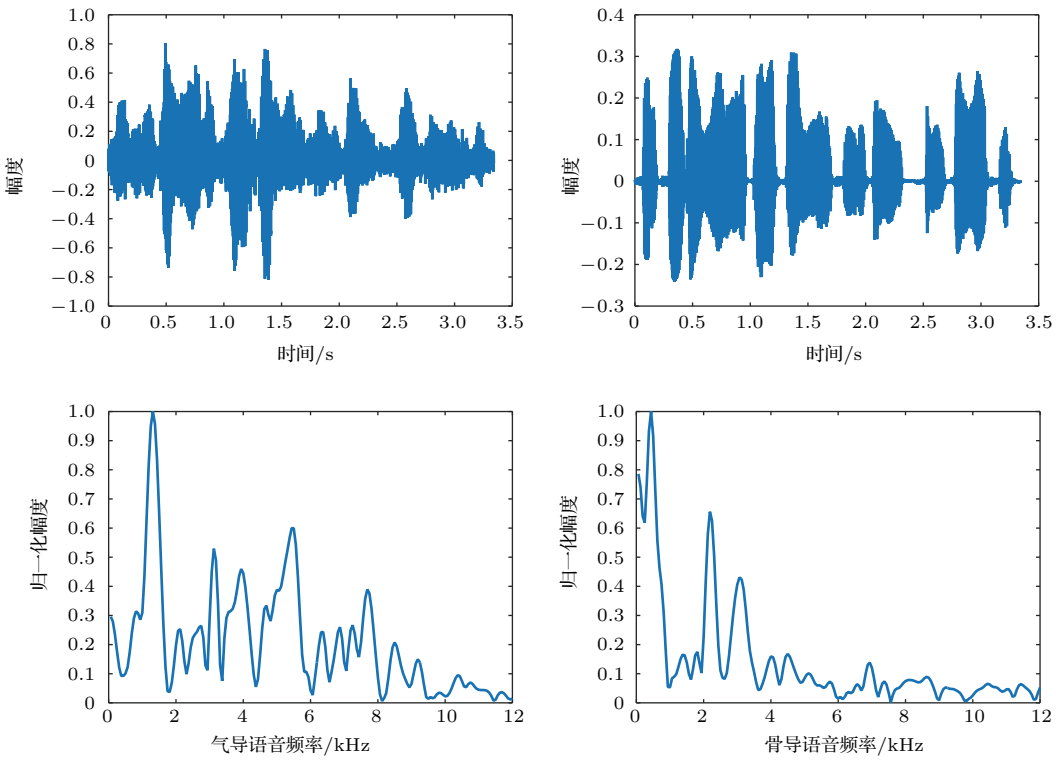
骨导语音是骨导麦克风通过捕捉头骨或喉头振动采集的语音信号，由于其不受空气中的噪声干扰，得到的语音具有很强的抗噪性能。虽然骨导语音本身仍存在有待改善的问题，例如：语音中高频成分较弱，可懂度低等，但是充分利用其较强的抗噪性能，对于改善语音切割效果会起到很大作用。为了验证骨导语音的抗噪性，通过实验得到如图 1 所示结果。

图 1 是同步采集语音信号的气导语音与骨导语音时域信号图形及其对应帧的频率成分分析图。两者在时间和内容上都具有一致性。

图 1 中左侧图形均为气导语音的相关图，右侧均为骨导语音的相关图。图 1(a) 展示了相同语句内容，气导和骨导的时域图和频率成分图。可以看出，该语句内容的气导语音在中高频部分幅度较大，其保存信息相对较多，而骨导语音在低频部分幅度较大，说明骨导低频部分保存信息相对较多。图 1(b) 展示了在受嘈杂人声背景噪声影响下，气导和骨导分别对应的时域波形和频率成分图。分析图 1(b) 可知，气导语音受噪声干扰后其中高频信息已受到严重干扰，由频率成分图可知，此时气导语音的低频信息较强而中高频信息则相对较弱，与图 1(a) 中的气导频率成分图产生较大变化。而骨导语音几乎不受外界任何干扰，其频率成分分析图与图 1(a) 中基本保持不变，由此可见骨导语音的抗噪性相对气导更加明显。所以在制备语音数据库时，同步录制骨导语音，在预处理阶段提取骨导语音信号，对其进行去噪处理，可以很大程度上减少外界噪声对算法适应性的干扰。后续在进行端点检测或语音分割时，可提升其检测或分割的适应性。



(a) 未受噪声影响的语音时域信号图和频率成分图



(b) 受嘈杂人声背景影响下的语音时域信号图和频率成分图

图 1　气导语音与骨导语音的频率成分

Fig. 1　The frequency component of air conduction speech and bone conduction speech

2　改进的语句分割算法

改进分割算法主要体现在两个方面：一是采用骨导语音应对多种复杂噪声，提升算法抗噪性和适应性；二是结合分段双门限检测和语音聚类等方法实现语音分割，有效降低了传统固定阈值分割带来的“一句分割成多句”或“多句分为一句”等问题。在处理过程中还提出了一些其他改进步骤，如：引入随机动态阈值、改善相似度度量方式以及自适应等改进方案，最终从整体上提升了语音分割算法的准确度和鲁棒性。

改进分割算法的基本流程如图 2 所示，针对改进流程下面具体介绍各步骤实现方法。

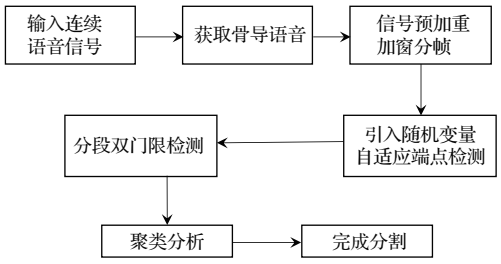


图 2　改进语音分割算法基本流程

Fig. 2　The basic process of improving the speech segmentation algorithm

2.1　随机变量的自适应端点检测

语音端点检测算法主要包括特征提取和端点判定两个环节 [10]。传统的方法在特征提取时，提取单一的时域参数或频域参数，作为区分语音段与噪声段的特征。本文在利用骨导语音良好的抗噪性能的前提下，使用了短时能量和过零率时域融合的参数特征，克服了单一参数特征抗噪性差与区分性差的缺点，一定程度上提升了端点检测算法的准确性 [2,11−12]。但在连续语音分割时，因语句内容的长短不一，并且在一句话内部中也会产生停顿和间隙，所以在进行端点检测时，还是会容易造成误检 (将语句内部的停顿作为新语句的起始点分割)。为了有效避免此类误检，引入随机变化的动态阈值进行端点判定，将固定区间的截取变成动态区间的截取，克服了固定阈值不能自适应环境的缺点，使端点检测算法适应性更强。其实现的具体方法如下：

$$\tau_i = \frac{\text{length}(S)}{\omega n} + \kappa \times \text{rand}, \tag{1}$$

式 (1) 中，τ_i 表示初始设置第 i 段中包含静音帧的数量值，$\text{length}(S)$ 表示初始选取一段语音的时间长度 (单位：ms)，ω 表示帧和秒的转换关系 (单位：ms/帧)，n 表示选取的随机语音段内能量出现能量峰的个数，κ 表示信号分帧时选取的帧移位置的大小，rand 表示随机生成的 (0,1] 区间上的数，通过式 (1) 能够确定检测出语音起始端点后需要向后位移的帧数。因为引入了 rand 随机量的生成，所以间隙起点与语音结束点恰好重合的概率大大降低。

2.2　分段双门限检测

双门限检测，是指通过设置检测门限的最低值和最高值来判定语音是否开始和结束。分段双门限则是为了有效应对语句内部间隙停顿和语句间间隔类似情况下造成语句分割点误判的问题。简单来说，就是为了防止因为语句内部间隙原因而将一句话误分成两句话或多句话的问题。分段第二段采用的检测方法与第一段相同，但是其在设置语音信号间隔中的静音帧数和随机参量做了调整，其计算公式如下：

$$\tau_i = a \times \left[\frac{\text{length}(S)}{\omega n}\right] + \frac{1}{2} \times \kappa \times \text{rand}. \tag{2}$$

由公式 (2) 可以看出，其主要是在对静音段帧数进行了一定程度上的减少，a 是一个比例系数，其取值范围 (0,1]，相关系数主要通过实验测试所得，本实验中取 2/3。通过上述分段的两段检测，通常会得到一段较长的语音和一段相对较短的语音段，在连续语音中两者交替出现。

2.3　语音聚类

当对说话人语音进行分割之后，输入音频被切分成了若干片段，通常希望分割后每个片段中只包含一个人的一句话，而聚类就是将这些语音片段依次重新组合，把一句话的片段聚为一类。常见的聚类策略有基于支持向量机、层次凝聚聚类等。层次凝聚聚类是一种贪心聚类方式，在聚类的过程中把相似度高的两个类别合并，简单高效，示意图如图 3 所示。

本文正是采用层次凝聚聚类方法，其具体实现步骤如下：

(1) 将端点检测之后得到的语音片段作为初始类别，对每个类别进行建模。

(2) 算出两两类别之间的相似程度，得到距离度量矩阵。

(3) 依次对相邻两个语音片段进行相似度比较，如果相似度高(相似度大于某一阈值) 即合并为一句，然后将合并后的句子与接下来一句继续进行比较，直到其相似度小于阈值。如果第一次比较就小于相似度阈值，则不合并前两句，分别将第一句生成单独的语音片段，第二语音片段作为下次比较的第一个片段，继续比较。

(4) 重复步骤(2)和步骤(3)，当遍历所有语音片段后停止聚类。

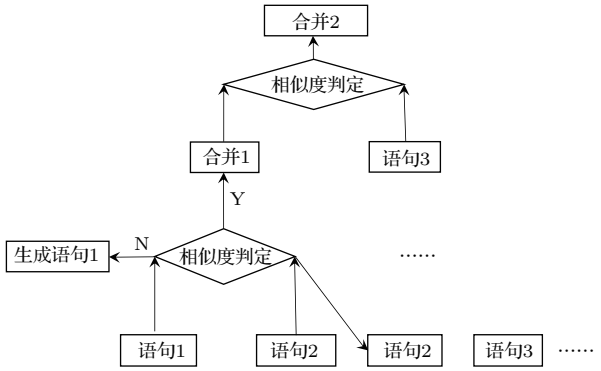


图3 语句聚类图

Fig. 3 Statement clustering

聚类过程中一个重要的问题就是相似度判别方式，大多情况下首先以其两者之间的距离作为度量。在本文中，采用了欧氏距离2范数的方法来进行相似度度量，求各个元素的平方求和然后求平方根。其计算公式如式(3)所示：

$$d_i = \left(\sum_{i=1}^{n} |x_{i+1} - y_i|^2\right)^{1/2},$$
$$l_i = \frac{2}{3}\frac{|y_i - x_i|}{|y_{i+1} - x_{i+1}|}, \tag{3}$$

式(3)中，d_i表示第i个相邻语音片段之间的欧氏距离，x_i表示第i段语音片段的开头位置，y_i表示第i段语音截止位置，l_i表示另一种对其相似度的判断条件。计算出d_i和l_i后分别与阈值比较，阈值设定则根据实际观测取定值。

3 实验结果及对比分析

为了验证本文所提改进算法的有效性和可行性，在windows操作系统下的Matlab 13.0环境中进行了实验。实验选取了20名男生，20名女生，每人200句连续语句作为样本进行分割，共计8000个样本。

3.1 本文算法实验效果

实验样本语音是选取了32.00 kHz的采样和16 bit量化情况下同步录制真实包含周围嘈杂人声的语音数据，帧长取240采样点，帧移取80采样点。本文算法的分割效果如图4所示。

图4(a)是实验中男8(编号为8的男生)录取包含背景噪声情况下混合双声道的部分语句时域波形图。图4(b)为分离后气导语音时域波形图，图4(c)分离后的骨导语音时域波形图。对比图4(b)和图4(c)可以看出，气导语音受到严重干扰，而骨导语音受外界环境的影响很小，较好地保持了说话人语句起始和终止的位置信息。图4(b)中的黑色竖线和图4(c) 红色竖线分别表示本文算法在气导语音和骨导语音上分割出第一句语音的起始和终止位置。图4(d)中蓝色部分表示气导情况下截取的第一段语音，红色表示骨导情况下截取的第一段语音。放大提取后的语音片段，可以明显看出，基于骨导语音的分割更加准确，这也说明骨导语音具有良好的抗噪性，可以更好地利用这一特点，对含噪语音进行分割和提取。

为了更加充分证明实验分割的准确度，对8000句语音进行切割，其统计结果如表1所示。

表1 本文方法分割后语音的数量和正确率

Table 1 The number and accuracy of the speech after this method is segmented

	男1~10	女1~10	男11~20	女11~20
句子总量	2000	2000	2000	2000
分割后数量	2043	2018	2022	2035
正确率	97.5%	98.8%	98.4%	97.8%

表1统计了本文分割算法对20名男生和20名女生的分割结果，分割语句的数量并不代表准确率，因为分割中出现的误聚类、分割丢失或一句多分等情况，正确率计算公式为分割正确数量/分割所得总数量。表1中所给出的正确率为经人工检验分割语句内容后计算所得。因为在语音库制取过程中，还存在人为因素，比如语句不流畅、发音不明显等问题，语句与语句间隔有时甚至不如一句话内部停顿时间长，对于上述情况本文算法仍然无法有效分割，但是对于受噪声影响的语句则可以有效准确地分割。由表1可得其分割正确率比较高。核对数量则相对较少，所以其可以大大减少人工分割的时间，提高工作效率。

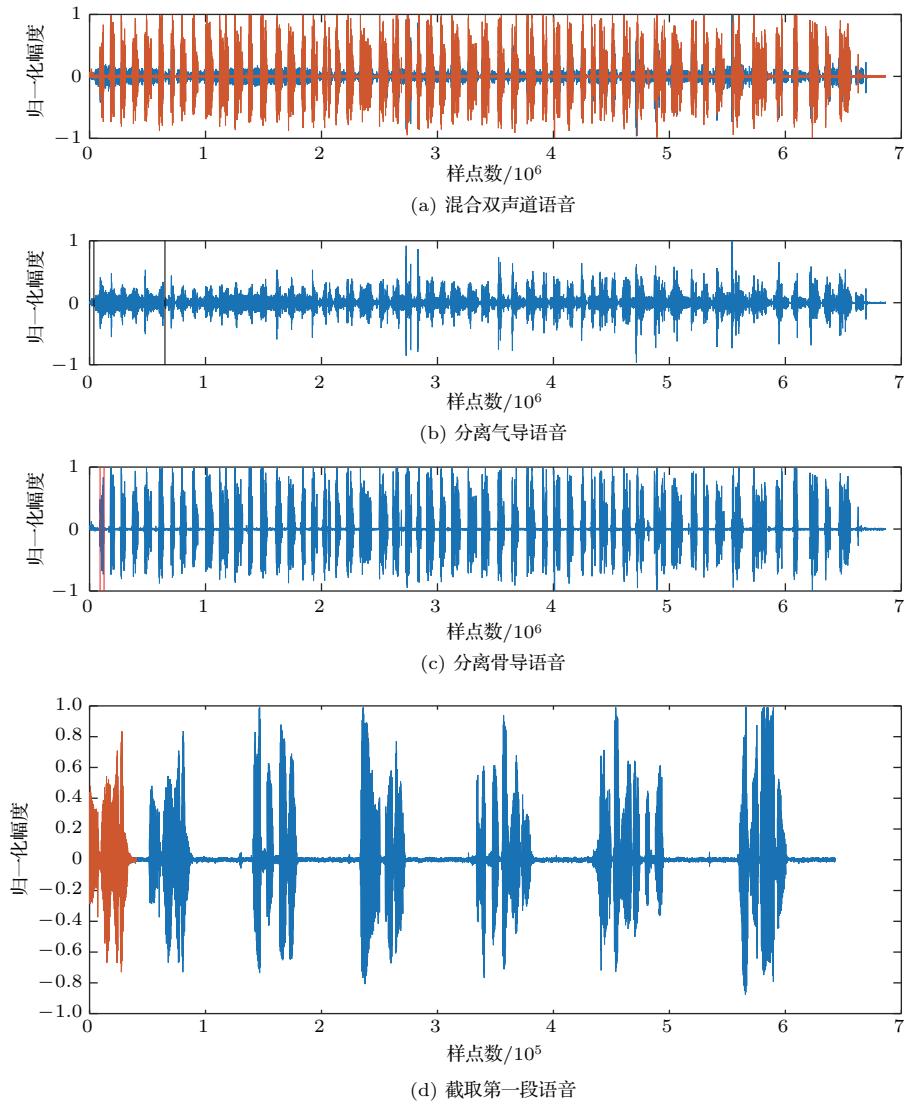


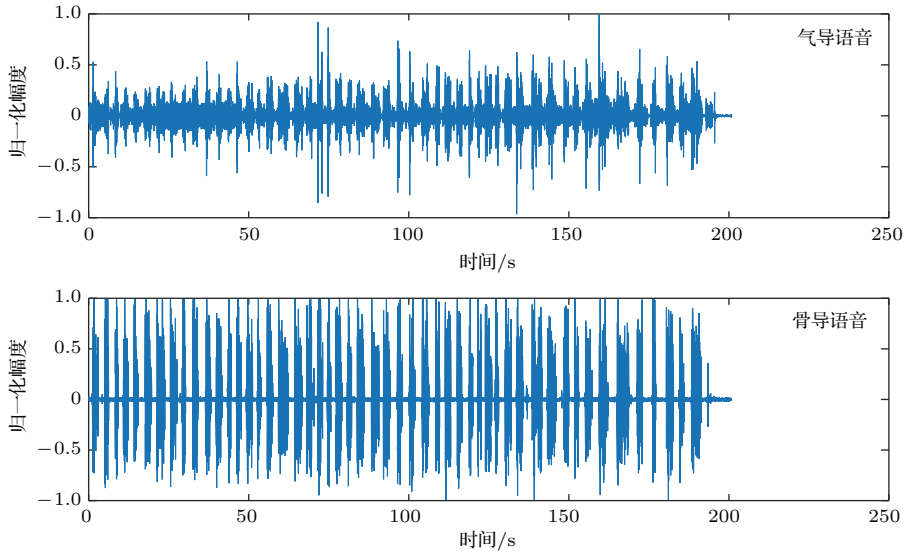
图 4　本文所提算法实验效果图

Fig. 4　Experimental effect of the algorithm in this paper

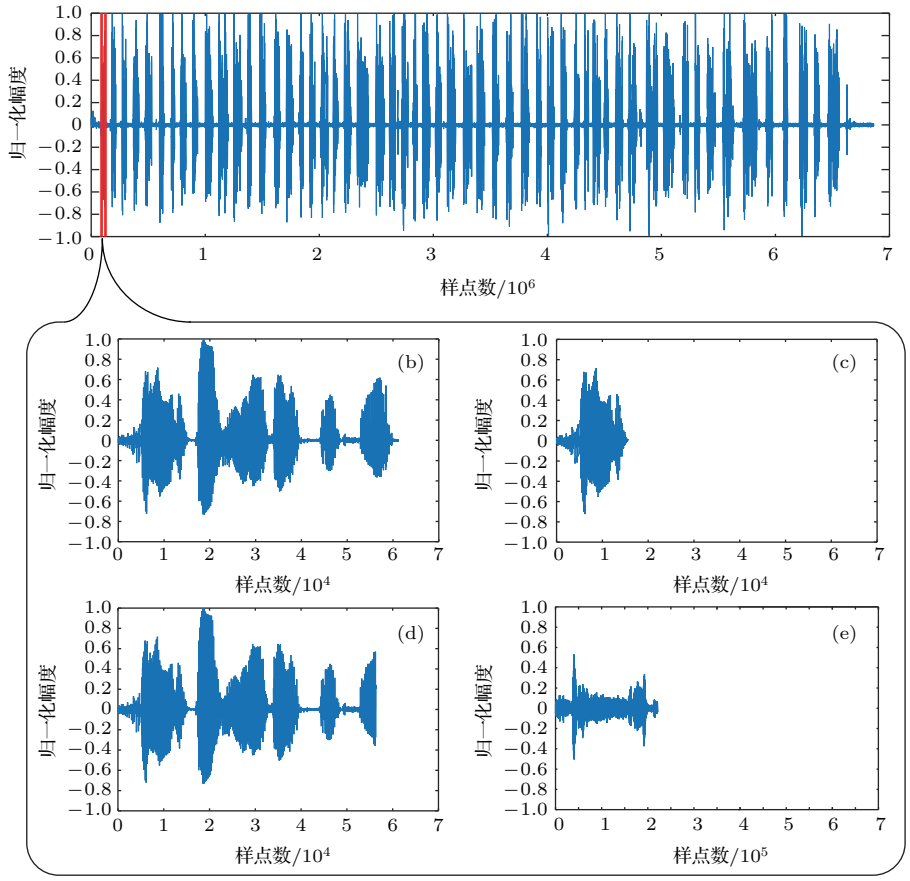
3.2　对比实验效果

为了进一步说明本文所提算法的优良性，本文还与骨导情况下的传统单特征和混合特征的单门限分割算法以及气导情况下的融合特征单门限分割算法进行了对比。针对含有嘈杂人声、音乐背景、汽车鸣笛背景等多种噪声环境进行了测试。本次实验选取了同步录制时含有背景音乐的语音数据进行测试，效果如图 5 所示。

图 5 展示了四种分割算法对同一段语音分割出第一个语句的分割效果。图 5(a) 表示含有音乐背景噪声的气导语音和骨导语音时域波形图。图 5(b) 表示本文的分割方法，经检验其分割正确，同时可以看出语句中间存在的较长间隙，但本文所提方法依旧能够实现正确分割。图 5(c) 表示基于骨导语音选取单特征时分割算法的分割效果，其误语句中间的间隙作为语句截止点，将一句话分割成多句，分割错误。图 5(d) 表示基于骨导语音的融合特征固定阈值分割方法，可看出由于分割阈值固定，当语句内部停顿或静音段较大时会造成分割错误，引起后半句语句内容丢失。在实际实验中需要不断调整固定阈值，因为后续语句间隔的不确定性仍旧会造成“一句多分”的现象出现。图 5(e) 表示基于气导语音的融合特征分割法，由于受到背景音乐的干扰，直接对其分割，其分割准确率很低。如果采用去噪方法，不同的背景噪声所需去噪方法不同，分割算法的适应性降低。由此可见本文的分割算法优于其他三种分割算法。同时对四种算法进行了分割效果统计，如表 2 所示。



(a) 同步录制含音乐噪声的语音时域信号图



(b) 本文分割方法 (c) 骨导单特征分割方法 (d) 骨导融合特征固定能够阈值法 (e) 气导融合特征分割法

图 5 四种方法分割效果对比

Fig. 5 Comparison of four methods of segmentation

表 2 是四种分割算法对男生和女生分别同步录制三种不同背景噪声 (音乐背景噪声、嘈杂人声背景噪声、车辆鸣笛噪声) 情况下的语句分割正确率的统计结果。结合图 5 的分割效果，可以看出融合特征算法优于单特征分割算法，基于骨导的优于基于气导的分割效果，进而验证了本文分割算法效果更好，分割准确率更高，也说明算法的适应性较强。

表 2　四种分割算法在不同噪声环境下分割的正确率

Table 2　Accuracy rate of four segmentation algorithms for different noise environments

	本文算法	(骨导) 单特征分割法	(骨导) 融合特征固定阈值法	(气导含噪) 融合特征法
男 (音乐背景噪声)	96%	79.4%	89.5%	74%
男 (嘈杂人声背景噪声)	94.5%	71.2%	86.2%	67.2%
男 (车辆鸣笛噪声)	92.2%	67.5%	80.2%	64.6%
女 (音乐背景噪声)	98.6%	83.6%	93%	76.5%
女 (嘈杂人声背景噪声)	96.7%	78%	89.6%	74%
女 (车辆鸣笛噪声)	94.2%	70.6%	85%	65.8%

4　结论

本文利用骨导语音具有的优良抗噪性，在对骨导语音预处理的基础上提出了一种自适应的分段双门限语音分割算法，通过时域特征融合、引入随机动态阈值以及分段双门限检测等多个方面改善语音分割效果，并通过实验证明了其有效性和鲁棒性。针对需要同步录制背景噪声或某些信噪比较低情况下的语句分割问题，找到一种最接近手工分割结果的端点位置，从而达到对噪声环境下的连续语音进行分割的目的，且分割精度和准确度获得一定程度上的提高。当然，算法还存在一些可以继续完善的地方，例如：在相似度判定上还可以做进一步的自适应调整，根据语音长度和整个语句信息的相关参数确定一个变化的相似度阈值，可以使算法的性能进一步拓展，后续相关工作也可以在这方面进行相应的实验和改进。

参　考　文　献

[1] 洪奕鑫, 张浩川, 余荣, 等. 语音端点检测在实时语音截取中的应用 [J]. 无线互联科技, 2017(22): 50–53.
Hong Yixin, Zhang Haochuan, Yu Rong, et al. Application of speech endpoint detection in real time speech interception[J]. Wireless Internet Technology, 2017(22): 50–53.

[2] 吕卫强, 黄荔. 基于短时能量加过零率的实时语音端点检测方法 [J]. 兵工自动化, 2009, 28(9): 69–70, 73.
Lyu Weiqiang, Huang Li. Realtime voice activity detection based on short time energy plus rate of passing zero[J]. Ordnance Industry Automation, 2009, 28(9): 69–70, 73.

[3] 纪振发, 杨晖, 李然, 等. 基于短时自相关及过零率的语音端点检测算法 [J]. 电子科技, 2016, 29(9): 52–55.
Ji Zhenfa, Yang Hui, Li Ran, et al. Speech endpoint detection algorithm based on short time autocorrelation and short-time zero crossing rate[J]. Electronic Science and Technology, 2016, 29(9): 52–55.

[4] 赵新燕, 王炼红, 彭林哲. 基于自适应倒谱距离的强噪声语音端点检测 [J]. 计算机科学, 2015, 42(9): 83–85, 117.
Zhao Xinyan, Wang Lianhong, Peng Linzhe. Adaptive cepstral distance-based voice endpoint detection of strong noise[J]. Computer Science, 2015, 42(9): 83–85, 117.

[5] 王群, 曾庆宁, 郑展恒. 低信噪比下语音端点检测算法的改进研究 [J]. 科学技术与工程, 2017, 17(21): 50–56.
Wang Qun, Zeng Qingning, Zheng Zhanheng. Research of speech endpoint detection in low SNR environment[J]. Science Technology and Engineering, 2017, 17(21): 50–56.

[6] You D, Han J, Zheng G, et al. Sparse power spectrum based robust voice activity detector[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012: 289–292.

[7] 曹亮, 张天骐, 周圣, 等. 一种基于奇异谱的语音激活检测方法 [J]. 应用声学, 2013, 32(2): 137–143.
Cao Liang, Zhang Tianqi, Zhou Sheng, et al. A method of voice activity detection based on spectrum of singular value[J]. Applied Acoustics, 2013, 32(2): 137–143.

[8] Wu J, Zhang X L. Efficient multiple kernel support vector machine based voice activity detection[J]. IEEE Signal Processing Letters, 2011, 18(8): 466–469.

[9] Zhang X L, Wu J. Deep belief networks based voice activity detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(4): 697–710.

[10] 王晓华, 屈雷. 基于时频参数融合的自适应语音端点检测算法 [J]. 计算机工程与应用, 2015, 51(20): 203–207, 212.
Wang Xiaohua, Qu Lei. Self-adaptive voice activity detection algorithm based on fusion of time-frequency parameter[J]. Computer Engineering and Applications, 2015, 51(20): 203–207.

[11] 戴元红, 陈鸿昶, 乔德江, 等. 基于短时能量比的语音端点检测算法的研究 [J]. 通信技术, 2009, 42(2): 181–183.
Dai Yuanhong, Chen Hongchang, Qiao Dejiang, et al. Speech endpoint detection algorithm analysis based on short-term energy ratio[J]. Communications Technology, 2009, 42(2): 181–183.

[12] Patel R, Shrawankar U. Security issues in speech watermarking for information transmission[J]. Computer Science, 2013: 830–839.