文章编号:1003-0530(2017)02-0229-07

# 语音合成系统中语音库样本能量均衡方法研究

刘 伟 谢建志
(电子科技大学通信抗干扰技术国家重点实验室, 四川成都 611731)

**摘 要**: 语音合成(Text to Speech, TTS)技术是实现人机语音通信的一项关键技术, 语音库的质量是决定 TTS 效果的重要因素。本文针对 TTS 语音库制作周期长, 发音人录音状态(音色、能量)差异而导致的 TTS 语音数据库录制后能量不一致问题, 提出了一种语音能量均衡方法, 包括时域包络波动检测和帧能量平均两个步骤。首先分析获得标准语音的相关能量参数和波动参数作为模板, 利用时域包络波动检测算法对预调节语音样本的合格性进行检验; 最后, 根据帧能量平均准则, 对所有合格语音样本进行时域幅值调整, 以最大限度地保证语音库整体能量的一致性。实验结果表明, 本文提出的语音能量均衡方法可以有效提升 TTS 语音库质量, 具有实际工程意义。

**关键词**: 语音合成; 能量均衡; 时域包络波动检测
**中图分类号**: TN912.3 **文献标识码**: A **DOI**: 10.16798/j.issn.1003-0530.2017.02.014

# Voice Energy Balance Method for Text to Speech Database

LIU Wei XIE Jian-zhi
(National Key Laboratory of Science and Technology on Communications, University of Science and Technology of China, Chengdu, Sichuan 611731, China)

**Abstract**: The quality of speech library is an important factor, which determines the effect of Speech to Text (TTS). The production cycle of the TTS speech database needs about six months. During the period, the voice state recording needs to be consistent, that is, the tone and energy can not have a big difference, which is more difficult for pronunciation. Thus, this paper gives voice energy balance method, including the time-domain envelope detection algorithm and the frame energy average algorithm, aiming to solve the TTS speech database recording after the phenomenon of inconsistency. Firstly, obtaining the standard speech related energy parameters and wave parameters as a template; secondly, using the time-domain envelope fluctuation detection algorithm to check the pre-regulation speech samples test. Finally according to the frame energy average criterion of all qualified speech samples, adjusting the samples amplitude in time domain value, to maximize the overall energy of the speech database consistency. The experimental results show that the proposed method can effectively improve the quality of the TTS speech database, and has practical engineering significance.

**Key words**: speech to text; energy balance; time-domain envelope detection

## 1 引言

语音合成(Text to Speech, TTS)技术是实现人与计算机进行语言交互所需的一项关键技术,在人机通信等领域有着重要应用。与早期基于规则构建的 TTS 系统相比,目前大多数的 TTS 系统都是基于大量的语音数据和统计模型所建立的,该系统合成时是从语音库中挑选出最佳单元并结合韵律进行拼接。由于合成的语音基元都是来自自然的原始发音,合成语句的清晰度和自然度都将明显高于早期基于规则的 TTS 系统[1-2]。语音库是基于拼接的 TTS 系统中重要组成部分,它的好坏将直接影响合成语音的质量[3-4]。

TTS 语音库中标准语音也称为模板语音,其特

收稿日期: 2016-07-08; 修回日期: 2016-10-28

性为语音整体能量波动小,时域波形平滑变化,不会出现“尖峰”现象,TTS 语音库中样本能量均衡都是以标准语音为模板进行调节的。语音库通常都是在录音条件好、抗噪能力强的语音室内由专业人员录制的,通常,TTS 语音数据库制作需要耗费半年左右时间。期间,发音人的录音状态需要保持一致,即音色、能量、语速等皆不能有大的差异,尽可能与标准语音质量相同。这对于发音人员较为困难,如开始录音时,人的精神状态上佳,录音质量和标准语音质量相近;随着长时间发音录制,人的生理状态变化将导致录音效果变差,不能达到标准语音的水平。若不对录制好的语音进行检验和调节,将导致语音合成效果变差[5]。目前并没有针对 TTS 语音库中语音样本能量均衡处理方法的研究,本文的研究工作对提升 TTS 效果具有实际意义。

本文提出的能量均衡方法目的在于将 TTS 语音库中合格语音样本进行能量调节,使与标准语音的能量参数最大程度相等。该方法中包括时域包络波动检测算法和帧能量平均算法,二者分别用来获得语音的波动参数和能量参数。由于语音样本在录制时会出现起伏较大、“尖峰”等情况,其调节时会发生严重截幅现象,属于不合格语音样本,本文通过比较语音样本和标准语音的波动参数来判断语音样本的合格性。最终的时域能量调节因子由标准语音和语音样本的能量参数之比得到,由于该因子调节后可能会出现截幅,为保证信号的频率特性不被破坏,需要对该因子自适应微调,以保证语音样本调节后最大限度地接近标准语音的能量参数。

简述能量均衡方法的思想,其时域包络波动检测算法是通过对降采样后的语音信号进行希尔伯特变换,并结合小波变换方法得到信号时域包络,设置门限阈值并粗略过滤掉包络中“非语音”部分;由于标准差是对样本离散程度的度量,计算语音样本时域包络的标准差可表征该信号的波动性,即本文的波动参数[6]。帧能量平均算法中,首先用语音端点检测算法得到精确的非静音段,计算其平均帧能量;考虑到 TTS 语音库样本良好信噪比,选用运算复杂度和时间复杂度较小的短时能零积方法进行端点检测[7]。实验通过对录制好还未处理的语音样本进行能量均衡调节,结果表明了本文提出方法的有效性和实用性。

本文的安排如下:第 2 节介绍算法原理和系统框图,第 3 节给出时域包络波动检测算法的具体实现,第 4 节给出帧能量平均算法的具体实现,第 5 节为实验与结果分析,最后为本文的总结。

## 2 算法原理和系统框图

从整体来看,表征语音信号的参数均是实时变化的,在短时间范围(20 ~ 30 ms)内相对稳定,可以看作为准稳态过程,此时间范围被称作帧。本文提出的能量均衡方法目的在于调节 TTS 语音库中语音样本的平均帧能量,使得其最大限度的接近标准语音的平均帧能量,图 1 给出了标准语音样本的时域波形,可以看出其波形整体平稳变化,波动性较小,没有明显的“尖峰”现象。

不同的语音样本具有不同长度的静音段,而静音段中是不含有有效信息的,因此计算平均帧能量只针对样本中的语音部分。如前文所述,专业发音人长时间的录音会引起精神状态变化,导致所录语音样本的整体能量下降,波动性较大,如图 2 所示,这类样本是不合格的,需要淘汰掉;而波动性接近或者低于标准语音的样本,经过本文方法调节,使其平均帧能量最大限度的达到标准语音的程度。

本文提出的能量均衡方法系统框图如图 3 所示。

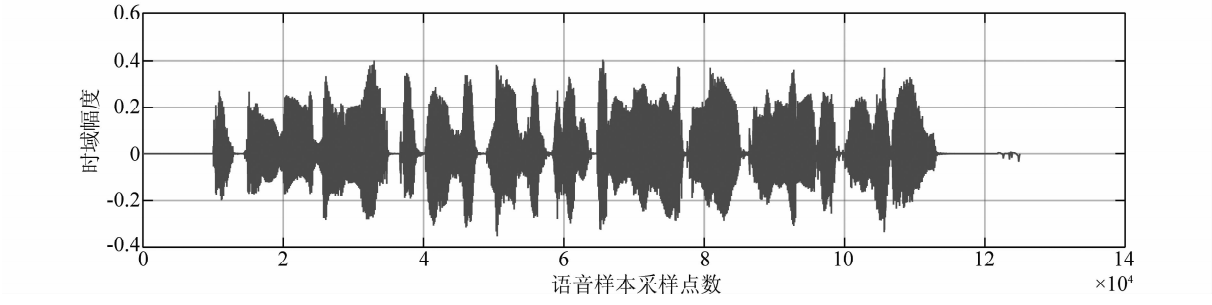


图 1 标准语音时域波形

Fig. 1 Time-domain waveform of standard speech

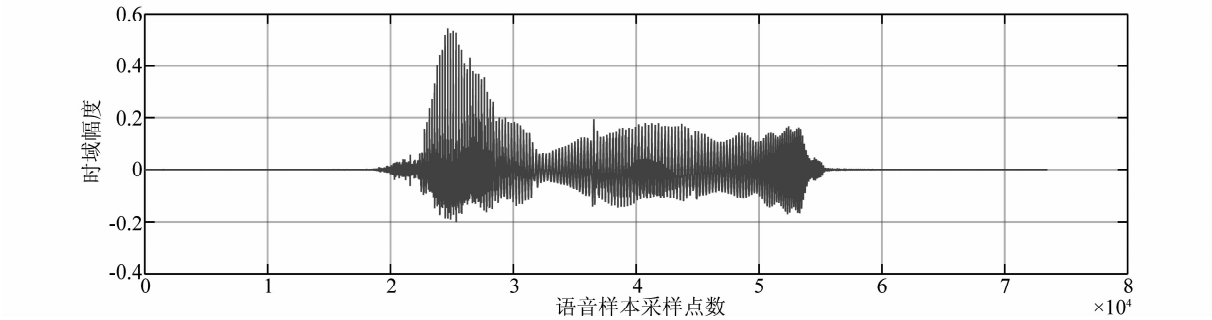


图 2　不合格语音样本时域波形

Fig. 2　Time-domain waveform of an unqualified speech sample

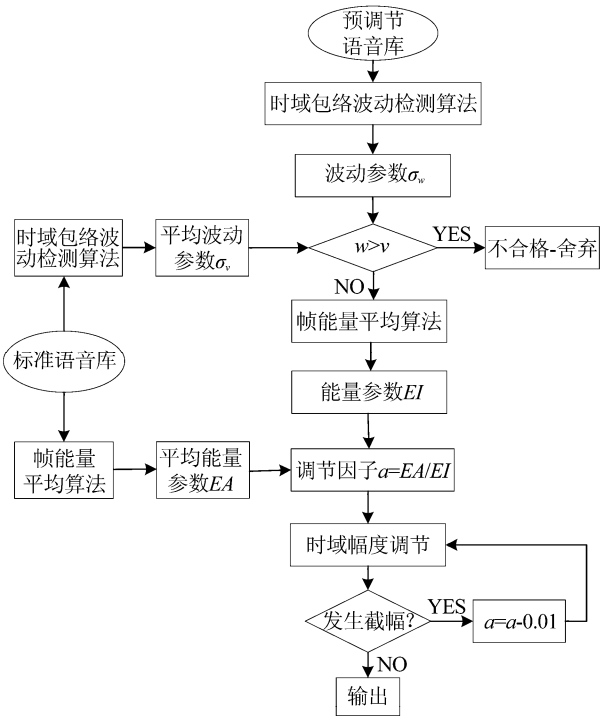


图 3　能量均衡方法系统框图

Fig. 3　Flow chart of the energy balance method

从图 3 的能量均衡方法系统框图可知,该方法中包含两个算法,其中时域包络波动检测算法用来计算语音信号的波动参数,帧能量平均算法用来计算语音信号的能量参数。具体流程为:

(1)计算标准语音库中所有标准语音的能量参数和波动参数,得到平均能量参数和平均波动参数。

(2)选择 TTS 预调节语音库中的一个语音样本开始调节,首先得到其波动参数,并与标准语音库的平均波动参数相比较;若小之,则执行后续调节处理,否则将其作为不合格语音样本淘汰掉,再开始处理 TTS 语音库中下一个语音样本。

(3)对合格的语音样本计算得到其能量参数,并通过与标准语音库的平均能量参数相比,得到调节因子 $a$,开始对样本的时域幅度调节。

(4)调节时,如果检测到发生截幅现象,为保证语音样本固有的频率特性不被破坏,对调节因子进行小幅度的减小,继续调节,直至不发生截幅,最后输出保存调节好的语音样本数据。

从方法的流程图可看出信号波动参数和能量参数的计算,将直接决定着能量均衡的效果。性能良好的时域包络波动检测算法和帧能量平均算法,可以使得本方法淘汰掉 TTS 语音库中因录音人员的生理状态变化所导致的不合格语音,以及调节整体能量较低的语音样本,达到标准语音的能量水平。

## 3　时域包络波动检测算法

本文利用时域包络波动检测算法计算语音信号的波动参数以表征其波动性。该算法主要思想为:首先通过小波变换和希尔伯特变换相结合的方法得到降采样后语音信号的时域包络;设置该包络随均值变化的参数为门限阈值,粗略去除静音得到有语音部分的包络;最后求此包络时域的均方差,得到波动参数。下面详述各个步骤的实现方法。

(1)提取信号时域包络

语音信号 $x(t)$ 的小波变换(Wavelet Transform, WT)定义如公式(1)所示:

$$WT_x^{\psi}(\tau,s) = |s|^{-1/2}\int x(t)\,\psi^*\left(\frac{t-\tau}{s}\right)\mathrm{d}t \qquad (1)$$

式中 $\psi(t)$ 称为基本小波,小波变换的基函数可以表示为:

$$\psi_{(\tau,s)}(t)=|s|^{-1/2}\psi(\frac{t-\tau}{s}) \tag{2}$$

它是基本小波$\psi(t)$在时域平移$\tau$,波形尺度伸缩$s$而得到的。上面式中$\tau\in\Re$,$s>0$,$\frac{1}{\sqrt{|s|}}$为能量归一化因子,$*$代表复共轭。通过控制$\tau$和$s$可以从不同的维度分析信号,使其在时域和频域都有很好的局部性。从小波变换定义可以看出,它相当于是一组不同中心频率带通滤波器,可对信号进行多通带的滤波以得到不同频带内的信息。而该带通滤波器的中心频率和带宽与尺度因子成正比,随着中心频率的变化自动调节,可实现对信号的自适应分析[8-10]。

传统的语音信号包络提取是先对语音信号进行带通滤波,然后对滤波后的信号进行希尔伯特变换生成复解析信号,从而得到包络,此方法存在固有缺陷是需要根据信号不同而设计滤波器。根据小波变换的性质,只要选择了合适的尺度因子和伸缩因子,就可分析想要的频段。而复解析小波函数是既可以实现带通滤波又能实现提取包络的小波函数,其表达形式如式(3)所示:

$$\psi_{C(\tau,s)}(t)=\psi_{(\tau,s)}(t)+j\cdot\psi_{H(\tau,s)}(t) \tag{3}$$

式中$\psi_{C(\tau,s)}(t)$是小波基函数$\psi_{(\tau,s)}(t)$经希尔伯特变换得到的解析信号,也即:

$$\psi_{H(\tau,s)}(t)=\psi_{(\tau,s)}(t)*\frac{1}{\pi t} \tag{4}$$

其中$1/\pi t$为希尔伯特变换时域表达。

利用得到的复解析小波基函数$\psi_{C(\tau,s)}(t)$进行语音信号包络包络分析如下:

$$W_x(\tau,s)=W_r(\tau,s)+j\cdot W_i(\tau,s)=\int_{-\infty}^{\infty}x(t)\,\psi_{C(\tau,s)}(t)\,\mathrm{d}t \tag{5}$$

式中$W_r(\tau,s)$和$W_i(\tau,s)$频率成分相同,但由于解析信号的希尔伯特变换特性,使得$W_r(\tau,s)$的相位比$W_i(\tau,s)$延迟了$\pi/2$,相互正交,此时对$W_x(\tau,s)$求幅值即得到语音信号$x(t)$的时域包络。

(2)语音包络过滤

得到语音包络中包含“有语音”和“无语音”部分,后者主要是噪声和静音部分,可近似认为无波动性的。由于TTS语音库中的样本特性不同,其“无语音”部分长度也各不相同,若计算样本波动性时未将此部分过滤掉,会增加波动误差;其次,衡量语音信号的波动性就是相对“语音”部分而言的。因此,通过在时域中设置一个随语音信号波形变化的阈值,只使“有语音”部分通过,其余滤掉。阈值如式(6)所示:

$$T(t)=\overline{D(t)}\cdot f(\log\overline{D(t)}) \tag{6}$$

其中$\overline{D(t)}$为语音信号时域绝对值均值,Sigmoid函数$f(t)=\alpha+\frac{\beta}{1+e^{-t}}$,通常选择$\alpha$为0.8,$\beta$为1,可以根据实际效果进行调节。

根据所设置的阈值,对语音信号时域包络进行调节,对小于阈值的时域值都舍弃,大于阈值的都保留,最后重新拼接过滤后的信号,得到只有语音部分的信号时域包络。

(3)计算波动参数

至此,得到了语音样本的“有语音”部分的时域包络,可利用标准差对此包络的波动性进行表征。在统计学中,标准差是反映个体间变异大小的指标,反映了整个样本对样本平均数的离散程度,是数据精密度的衡量指标。而语音信号包络是由多个时域离散点构成的,可以把这些离散点看作样本,则对语音包络求其标准差,即可看出该包络的波动性,表达如式(7)所示:

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2} \tag{7}$$

其中,$N$为时域信号$x(t)$包络的总采样点数,$x_i$为包络的第$i$个采样点值,$\mu$为$x(t)$包络的均值,$\sigma$即为语音样本的波动参数。

## 4 帧能量平均算法

语音信号可以看做为在单位帧内是短时平稳的过程,因此可求得语音样本的单位平均帧能量作为能量参数。由于语音样本采样率的差异,相同时长的语音对应的时域采样点数也不相同,因此本文的帧能量是指一帧时间内的语音能量。在求语音样本的帧能量时,不能将其“无语音”部分考虑进去,因为不同的语音样本中,“无语音”部分所占比例各不相同,得到的平均帧能量不具有统一性。因此,帧能量平均算法中首先需要用语音端点检测

(Voice Activity Detection, VAD)算法,检测出“语音部分”,再计算所有语音段能量最终得到其平均帧能量。

在现有的 VAD 算法中,主要是通过对信号噪声进行估计、语音增强,设置阈值并检测得到语音端点,这种方法适用于不同信噪比情况下的语音,其弊端在于时间复杂度和空间复杂度都非常大,不适合处理大规模的 TTS 语音库。TTS 语音库中的语音样本是在录音棚中录制得,其信噪比较好,因此本系统选用时间复杂度和空间复杂度都较小的短时能零积算法实现对语音样本的端点检测[11-13]。

短时能零积是指语音短时平均能量与短时平均过零率的乘积,设短时平均能量为 $E_n$、短时平均过零率为 $Z_n$,则短时能零积为:

$$EZ_n = E_n \times Z_n \tag{8}$$

其中,

$$Z_n = \sum_{n=1}^{N} \left| \operatorname{sgn}[s_w(n)] - \operatorname{sgn}[s_w(n-1)] \right| \tag{9}$$

式(8)和(9)中 $n$ 代表为该语音的第 $n$ 帧,$N$ 为帧长,$s_w(n)$为经过加窗处理后的语音。该方法的主要检测流程为:

(1)对待检测语音的前 10 帧计算其短时平均能量和短时过零率,并得到其短时能零积值 $EZ_n$;再取后 10 帧的短时能零积值求出其平均值,得到噪声门限阈值如式(10)所示,其中 $k$ 为常数,按照经验一般取 5。

$$\mathrm{TH} = k \times EZ \tag{10}$$

(2)将噪声门限阈值 TH 同每帧的短时能零积值 $EZ_n$ 比较,若 $EZ_n$>TH,认为此时进入语音段,记录该帧的帧序号作为有语音段的起始位置帧 $M_1$,之后若出现 $EZ_n$<TH,则该帧的帧序号就是有语音段的终点。反之,若出现 $EZ_n$<TH 时,$M_1$ 还未得到,则认为该帧处于非语音部分段。

得到语音样本的“有语音”首尾端点后,即可计算其平均帧能量如式(11)所示:

$$EV = \frac{1}{Mn} \cdot \sum_{m=1}^{M} \left\{ \sum_{i=b_m}^{t_m} |x(i)|^2 \right\} \tag{11}$$

其中,$Mn$ 为语音样本中“有语音”部分的总帧数,$M$ 为“有语音”部分的段数,$b_m$ 和 $t_m$ 分别为第 $m$ 段语音的起始采样点数和末尾采样点数,$x(i)$为语音信号时域值[14-16]。

## 5　仿真实验与结果分析

为了验证本文所提方法的性能,选择一组常用的标准语音作为标准集,对刚录制未处理的语音库进行能量均衡实验。实验所选的标准语音集中共有 7 个标准语音,将其在时域中进行拼接,如图 4 所示。

根据时域包络波动检测算法求得该标准语音集的平均波动参数为 $\sigma_{std}$=0.0056,用帧能量平均算法求得该标准语音集的平均能量参数为 $EV_{std}$ = 0.0158;根据此标准语音集参数,对刚录制好语音库进行实验。其中语音库中的语音样本共有 150 个,其时域波形合成如图 5 所示。

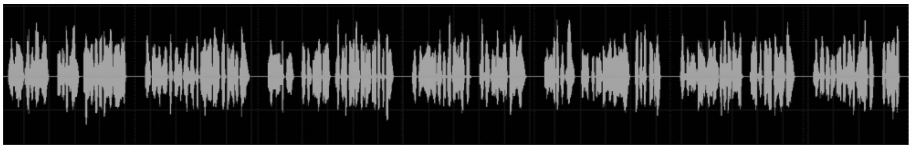

图 4　标准语音集时域合成波形(横轴为采样点,纵轴为幅度)

Fig. 4　Time-domain synthesized waveform from standard speech database

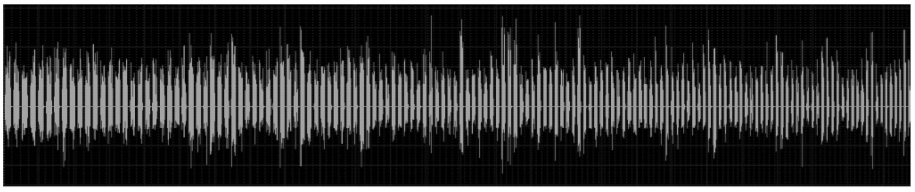

图 5　未处理的语音库样本时域合成波形(横轴为采样点,纵轴为幅度)

Fig. 5　Time-domain synthesized waveform from unhandled speech database

利用时域包络波动检测算法计算此语音库中的所有样本的波动参数,结果如图 6 所示。

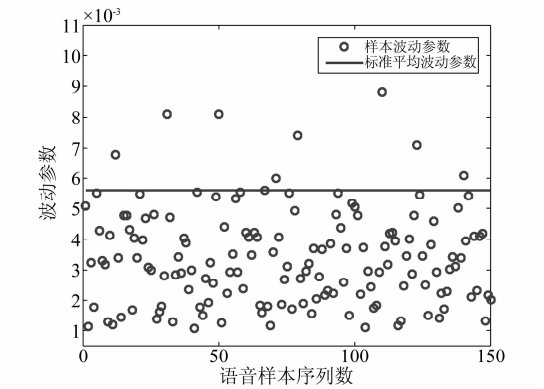


图 6 标准语音集和语音库中样本的波动参数

Fig. 6 Fluctuation parameters of a sample in the standard speech database

从图 6 中可以发现用于实验的 TTS 语音库的 150 个样本中,有 8 个语音样本的波动参数大于标准语音集的平均波动参数,说明这些语音属于不合格的样本,将其挑选出得到时域合成波形,如图 7 所示。

按照能量均衡方法系统框图所示,对剩余 142 个合格的语音样本进行进一步均衡处理,得到结果如图 8 所示。和未处理前时域信号图 5 比较,从图 8 粗略可看出经能量均衡调节后语音样本的能量有较大的改善,为了准确说明本文提出的能量均衡方法的有效性,计算处理后合格语音样本的平均帧能量,如图 9 所示。

从图 9 可以精确看出经过本文提出的语音能量均衡调节方法处理的 TTS 语音库中合格语音样本,其平均帧能量都接近于和标准语音集平均帧能量 0.0158 相等,这也进一步验证了本文所提方法的有效性。

图 7 不合格的 8 个语音样本(横轴为采样点,纵轴为幅度)

Fig. 7 Eight unqualified speech samples

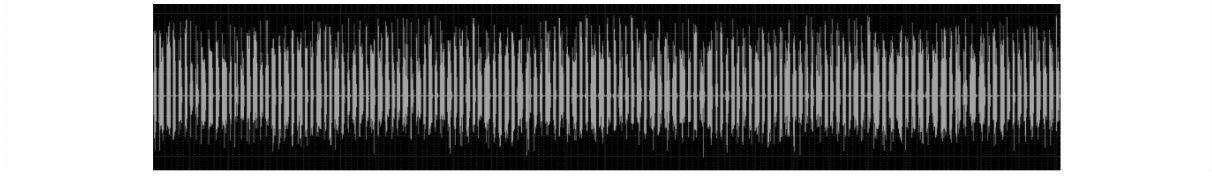

图 8 处理后合格语音样本时域合成波形(横轴为采样点,纵轴为幅度)

Fig. 8 Time-domain synthesized waveform of speech samples after processing

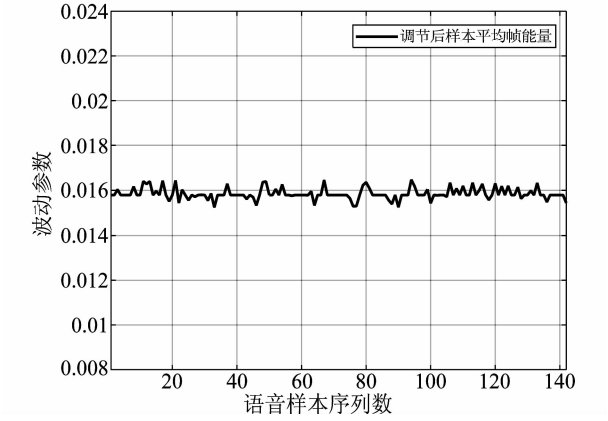


图 9 处理后合格语音样本时域平均帧能量

Fig. 9 Time-domain averaged frame energy of qualified speech sample after processing

## 6 结论

本文提出一种用于 TTS 语音库样本的能量均衡调节方法。该方法中包括时域包络波动检测算法和帧能量平均算法,分别用来计算语音信号的波动参数和能量参数。处理时,首先得到语音样本的波动参数,以理想的标准语音为模板,对其合格性进行检验;计算合格语音样本的能量参数,得到能量调节因子并进行自适应的时域幅度调节,使得调节后以最大限度的接近标准语音的平均帧能量。实验结果表明,经过本文方法处理后的 TTS 语音库,淘汰了有明显“尖峰”的不合格语音样本,调节后的合格语音样本平均帧能量与标准语音集的平均帧能量接近相等,有效提

升了 TTS 语音库的质量。

## 参考文献


[1] 杨辰雨. 语音合成音库自动标注方法研究[D]. 安徽合肥:中国科学技术大学,2014.
Yang Chenyu. Research on automatic labeling of speech synthesis corpora[D]. Anhui, Hefei: University of Science and Technology of China,2014. (in Chinese)

[2] 蔡明琦. 融合发音机理的统计参数语音合成方法研究[D]. 安徽合肥:中国科学技术大学,2015.
Cai Mingqi. Study of statistical parametric speech synthesis method based on Mechanism of pronunciation [D]. Anhui, Hefei: University of Science and Technology of China, 2015. (in Chinese)

[3] Heiga Zen, Andrew Senior, Mike Schuster. Statistical parametric speech synthesis using deep neural networks[C]// IEEE International Conference on Aconstic, Speech and Signal Processing,2013: 7962-7966.

[4] Youcef tabet, Mohamed boughazi. Speech synthesis techniques. a survey[C]//7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA),2011: 67-70.

[5] 庞敏辉. 语音库自动构建技术的研究[D]. 山东青岛:中国海洋大学,2010.
Pang Minhui. Study on antomatic construction of speech database[D]. Shandong, Qingdao: Ocean University of China,2010. (in Chinese)

[6] Samar V J, Bopardikar A, Rao R, et al. Wavelet analysis of neuroelectric waveforms: A conceptual, tutorial [J]. Brain & Language, 1999, 66(1):7-60.

[7] Ramirez J, Segura J C, Benitez C, et al. A new Kullback-Leibler VAD for speech recognition in noise[J]. IEEE Signal Processing Letters,2004, 11(2):266-269.

[8] Sharma D, Naylor P A. Evaluation of pitch estimation in noisy speech for application in nonintrusive speech quality assessment[C]//Proc European Signal Processing Conf, Aug. 2009: 2514-2518.

[9] Charles K. Chui. An Introduction to Wavelets[M]. New-York:Academic Press, 1992.

[10] Sunil Tyagi. Wavelet Analysis and Envelope Detection for Rolling Element Bearing Fault Diagnosis-A Comparative Study[J]. Center of Marine Engineering Technology INS Shivaji, Lonavla,2001:402-410.

[11] Ling Z H, Richmond K, Yamagishi J. Articulatory Control of HMM-Based Parametric Speech Synthesis Using Feature-Space-Switched Multiple Regression[J]. IEEE Transactions on Audio Speech & Language Processing, 2013, 21(1):207-219.

[12] 张勇,刘轶,刘宏. 结合人耳听觉感知的两级语音增强算法[J]. 信号处理,2014,30(4): 363-373.
Zhang Yong, Liu Yi, Liu Hong. A two-stage speech enhancement algorithm combined with human auditory perception[J]. Journal of Signal Processing, 2014, 30(4): 363-373. (in Chinese)

[13] 刘凤山,吕钊,张超,等. 改进小波阈值函数的语音增强算法研究[J]. 信号处理,2016,32(2):203-212.
Liu Fengshan, Lv Zhao, Zhang Chao, et al. Research on Speech Enhancement Algorithm Based on Modified Wavelet Threshold Function[J]. Journal of Signal Processing, 2016,32(2):203-212. (in Chinese)

[14] Sira Gonzalez, Mike Brookes. A Pitch Estimation Filter Robust to High Levels of Noise[C]//Proc European Signal Processing Conference, Barcelona, Spain, 2011: 451-455.

[15] 钟林鹏. 说话人识别系统中的语音信号处理技术研究[D]. 四川成都:电子科技大学,2010.
Zhong Linpeng. Studies on the Speech Signals Processing of the Speaker Recognition System[D]. Sichuan, Chengdu: University of Electronic Science and Technology of China,2010. (in Chinese)

[16] Di W U, Zhao H, Huang C, et al. Speech endpoint detection in low-SNRs environment based on perception spectrogram structure boundary parameter[J]. Chinese Journal of Acoustics, 2014, 39(4):392-399.


## 作者简介


**刘　伟**　男,1971 年生,山东文登人。硕士,电子科技大学通信抗干扰技术国家重点实验室讲师,主要研究方向为信号检测、信号处理、扩调频通信、无线网络。
E-mail:liuwei_71@outlook.com



**谢建志**　男,1970 年生,山东栖霞人。硕士,电子科技大学讲师,研究方向为信号处理、成像技术以及阵列信号处理。
E-mail:jzxie@uestc.edu.cn