Computer Science and Application 计算机科学与应用, 2018, 8(9), 1437-1446
Published Online September 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.89157

Hans汉斯

Indonesian Text Analysis and Processing for Speech Synthesis

Xuan Kong, Jian Yang*

School of Information Science and Engineering, Yunnan University, Kunming Yunnan
Email: *jianyang@ynu.edu.cn

Received: Sep. 6th, 2018; accepted: Sep. 21st, 2018; published: Sep. 28th, 2018

Abstract

This paper focused on the development of Indonesian speech synthesis system, and it studied Indonesian text analysis and processing methods. It mainly studied text normalization and syllable division methods. By using a combination of regular expressions and keywords, the numbers and special symbol in the text are normalized. Furthermore, a combination of syllable lists and special rules are used to achieve syllable segmentation. From the pronunciation corpora, 500 sentences containing special characters were selected for normalization testing. The correct rate of the number of words according to special characters was 96.0%. The in-set testing selected 1000 words in the dictionary, and the syllable results were compared with the results of artificial division with the correct rate of 98.2%. From the text corpora, 480 sentences were randomly selected for a total of 5850 words for out-of-set testing, and the correct rate was 97.1%. The above experiments laid a good foundation for the development of the Indonesian speech synthesis system.

Keywords

Indonesian, Speech Synthesis, Text Normalization, Syllable Division

面向语音合成的印尼语文本分析与处理

孔　璇，杨　鉴*

云南大学信息学院，云南　昆明
Email: *jianyang@ynu.edu.cn

收稿日期：2018年9月6日；录用日期：2018年9月21日；发布日期：2018年9月28日

*通讯作者。

文章引用: 孔璇, 杨鉴. 面向语音合成的印尼语文本分析与处理[J]. 计算机科学与应用, 2018, 8(9): 1437-1446.
DOI: 10.12677/csa.2018.89157

摘　要

本文以开发印尼语语音合成系统为目的，研究印尼语文本分析与处理方法，主要研究了文本归一化和音节划分方法。采用正则表达及关键字相结合的方法，对文本中数字及特殊字符进行归一化处理；采用基于音节列表及特殊规则相结合的方案实现音节划分。从发音语料库中挑选出500个包含特殊字符的句子进行归一化测试，按特殊字符词数统计的正确率达96.0%。选取词典中的1000个单词进行集内测试，其音节化结果和人工划分结果相比，正确率为98.2%；从文本语料库中任意选取480个句子共计5850个单词进行集外测试，其正确率为97.1%。实验结果表明，上述方法为印尼语音合成系统的开发奠定了良好的基础。

关键词

印尼语，语音合成，文本归一化，音节划分

1. 引言

在语音合成技术中，前端文本分析结果的质量直接影响合成语音的可懂度和自然度。因此，前端文本分析是语音合成系统的重要模块。作为世界第四大人口国的官方语言，近年来关于印尼语语音合成的研究相对较少。

目前印尼语公开发布的语料库是从一部流行的印度尼西亚小说中提取并以男女对话的方式进行录音的[1]，该语料库的覆盖范围较窄。对于印尼语语音合成系统，Mengko 和 Ayuningtyas 研究了基于音节拼接的印尼语文语转换系统[2]，其主要针对音节声音数据库质量和播放过程中音节的整合问题进行改进，但该系统音节列表不全且未考虑音节组合的韵律特征。Sutarman 研究了使用双音素拼接的印尼语文语转换系统，此系统在构造双音素数据库和文本到语音的过程中发现，在音素表中查找单词时不够精确，并且在分割过程中，使用双音素进行切分得到的结果也不尽如人意[3]。

本文聚焦于印尼语语音合成系统中的前端文本分析模块，着重关注文本语料库中数字及特殊字符的归一化以及基于音节列表和特殊规则相结合的印尼语音节的自动划分[4]。

本文的结构如下：第 1 节为印尼语的简单概述；第 2 节对印尼语发音语料库的构建进行阐述；第 3 节介绍了印尼语中非标准词的归一化方法；第 4 节介绍印尼语的音节划分；第 5 节对整个实验过程进行了总结。

2. 印尼语简介

印度尼西亚语(Bahasa Indonesia)是印度尼西亚共和国的官方语言，在整个印度尼西亚群岛被广泛使用。在语言学分类中，印尼语、马来语、爪哇语等一同构成了马来－波利尼西亚语系西印度尼西亚语支[5] [6]。印尼语由 5 个单元音，3 个双元音和 25 个辅音组成(此时不区分单元音 e 和 é)，并且它是一种没有声调的语言[3] [7] [8]。它是一种黏着语，新词形成的方式有三种：附加词缀(词缀分为前缀、中缀、后缀)到词根、几个词或部分词重复构成新词以及外来借词[5]。在印尼语中不使用语法性别，只使用自然性

别。没有语法复数；动词不涉及人称、数量和时态；时间的表示运用时间副词或其他时态指标(如 sudah“已经”和 belum“尚未”)。

以下为一个印尼语例句：

Gagal SNMPTN 2017, Masih Ada Jalan Menuju Perguruan Tinggi Negeri!

3. 发音语料库

语料库是基于语料库的语音合成系统的重要部分。语料库的质量和覆盖范围将影响合成语音的综合质量。本实验的初始文本是从印度尼西亚语网站下载的涵盖多个领域的印尼语文本，其大小为 587 MB。但初始文本中包含许多不完整的句子、非法字符和网页标签，并且网站之间存在相互转载，因此需要通过删除网页标签和重复句子来处理该初始文本，将得到的结果用作实验文本语料库，其大小为 566 MB。从该语料库中，挑选出 5000 个具有代表性的句子作为发音语料库。

3.1. 发音语料库的构建

在选取发音语料的过程中，本着尽可能用最少的句子覆盖到最多语言现象的原则来保证句子的最大信息量。因此，从语句的韵律和音段角度出发进行发音语料的挑选。在考虑音段信息时，由于印尼语是一种没有重音的语言，因此只需把连读现象、声调、自然停顿及各个音节组合在自然语流中出现的频率作为考量因素。在考虑韵律信息时，要关注所选取句子的句子类型[1]，这是因为在印尼语中，不同的句子类型在发音时主要是通过音调的变化来进行区分的，如升调表示疑问句，降调表示感叹句[7]。

从而在进行发音语料的挑选时，以句子长度以及句子类型为基础，采用优先收录高频词的思想设计了本次实验的语料挑选算法。此次试验共选取发音语料 5000 句，其中包括 4400 句陈述句，200 句感叹句，200 句一般疑问句及 200 句特殊疑问句。发音预料的挑选过程详见文献[9]。

3.2. 发音语料库的优劣评判

发音语料库是文本语料库中具有代表性的句子组成的集合，因此发音语料库的一些特征应该与文本语料库一致。所以发音语料库中要包含印尼语中的所有音素，且要保证发音语料的选取算法应该是稳定的。

在衡量发音语料库与文本语料库相似性时，比较的特征有：句子长度分布是否一致、音素出现次数的分布是否一致。

从图 1 和图 2 可以看出文本语料库中的句子和发音语料库中的句子二者句长分布是相似的，都呈现类似正态分布[9]。因此，发音语料库在句子长度方面具有代表性。从图 3 和图 4 可以看出发音语料库中的音素出现比例和文本语料库中的出现比例是一致的。从图 4 可以看出，发音语料库中所选句子已包含了印尼语中的所有音素。所以，在音素层面，发音语料覆盖了印尼语的全部语言特征。从图 5 可以看出，三次随机选取的发音语料库中音素的出现次数的分布是一致的，且不同次的选取结果中音素出现次数也是相近的，即证明选取发音语料的算法和过程是稳定的[9]。

4. 文本归一化

归一化是将非印度尼西亚语拼写的单词转换为标准印尼语单词的过程。包含数字和其他符号的文本称为非标准文本。文本中的数字及特殊符号称为非标准单词。归一化过程包括识别非标准词，歧义判断，消歧处理以及将非标准词转换为标准词。

4.1. 非标准词的识别

进行印尼语文本归一化处理时，首先应该进行非标准词识别。此次实验进行文本归一化处理时着重

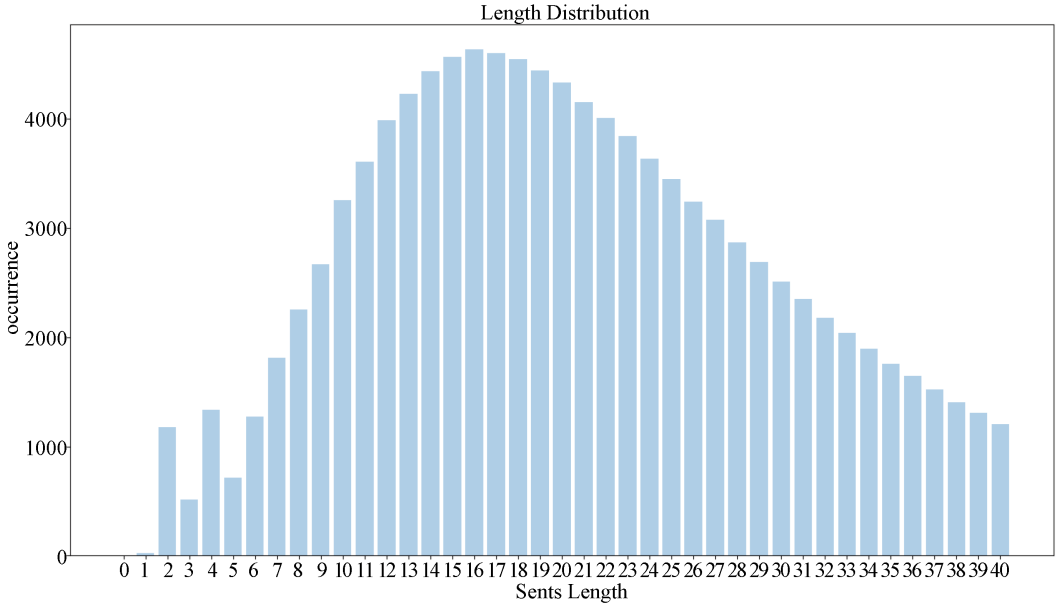

Figure 1. Text corpus sentence length distribution
图 1. 文本语料库句子长度分布

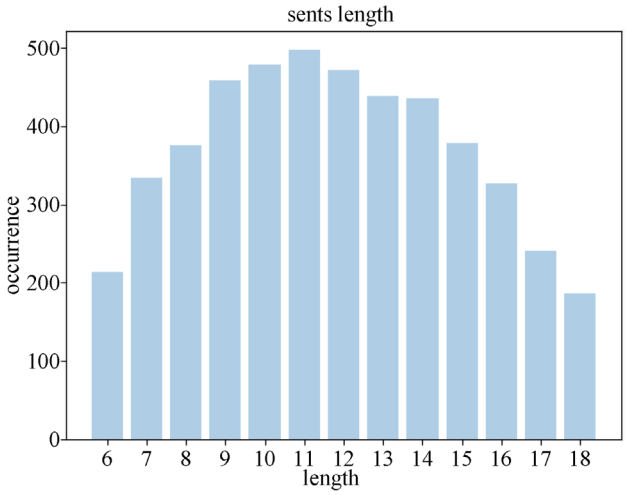

Figure 2. Pronunciation corpus sentence length distribution
图 2. 发音语料库句子长度分布

关注文本中的阿拉伯数字及特殊字符“-”。该过程将输入的文本解析为句子，再将句子分割为词，然后判断每个词是否含有阿拉伯数字“0~9”及特殊字符“-”，若有则进入歧义判断。

4.2. 歧义判断

一般来说，数字主要有两种读法，第一种读作数值(如 123 读为一百二十三)，第二种是读作数码(如 123 读为一二三)。此时需要根据上下文来判断其正确读法。而对于特殊字符“-”而言，其在不同的语境中出现时，发音情况也不同，故需分情况讨论。

4.2.1. 数字歧义判断

在文本中，按照数码的方法进行发音的情况较少，一般为身份证号、电话号码以及邮编。身份证号

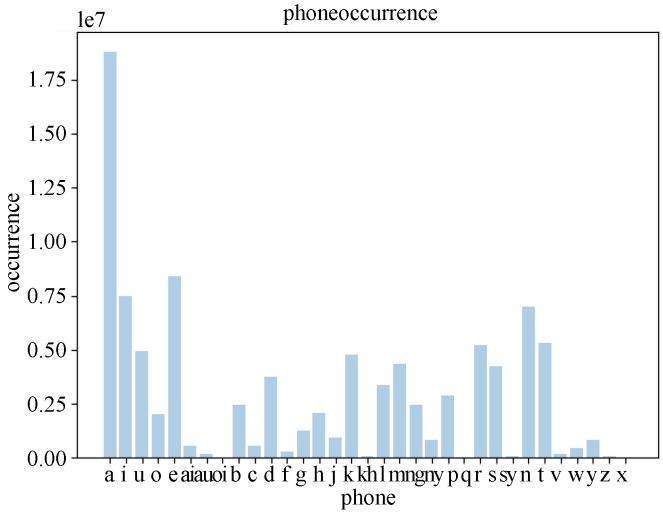

Figure 3. The number of phoneme in Text corpus
图 3. 文本语料库音素出现次数

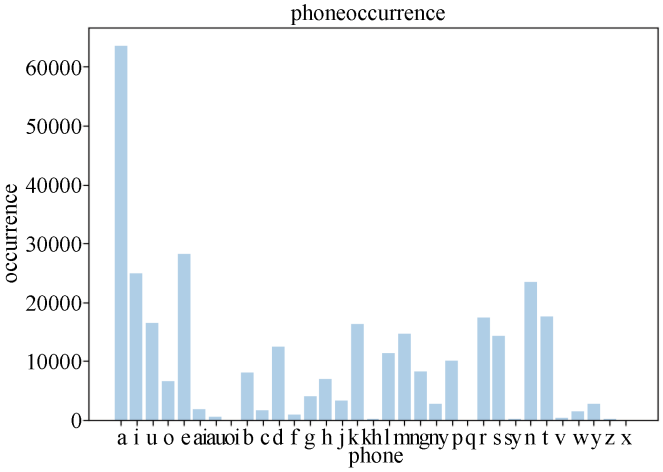

Figure 4. The number of phoneme in pronunciation corpus
图 4. 发音语料库音素出现次数

和电话号码的形式比较特殊，可以直接用正则匹配的方法进行歧义判断[10]。而邮编的形式需要用正则匹配加关键字的方法进行歧义判断。

- 印度尼西亚身份证号

印度尼西亚的身份证号码为 16 位纯数字。据此可编写印度尼西亚国民身份证号码正则表达式为：“(^|\s)\d{16}($|\s)”。

- 印度尼西亚电话号码

印度尼西亚的电话号码为 11 位纯数字且以 08 开头。据此可编写用来匹配印度尼西亚电话号码的正则表达式：“(^|\s)08\d{9}($|\s)”。

- 邮编

印尼邮编由 5 位纯数字组成，如 40115，但仅仅使用正则表达式进行匹配是不够的，因为当出现五位数字时，有可能表达的是数值。所以对邮编的进行歧义判断时，采用正则表达式和关键字相结合的方法，即当运用正则表达式“(^|\s)\d{5}($|\s)”判断出现五位数字时，还需要判断其上下文中是否出现关键

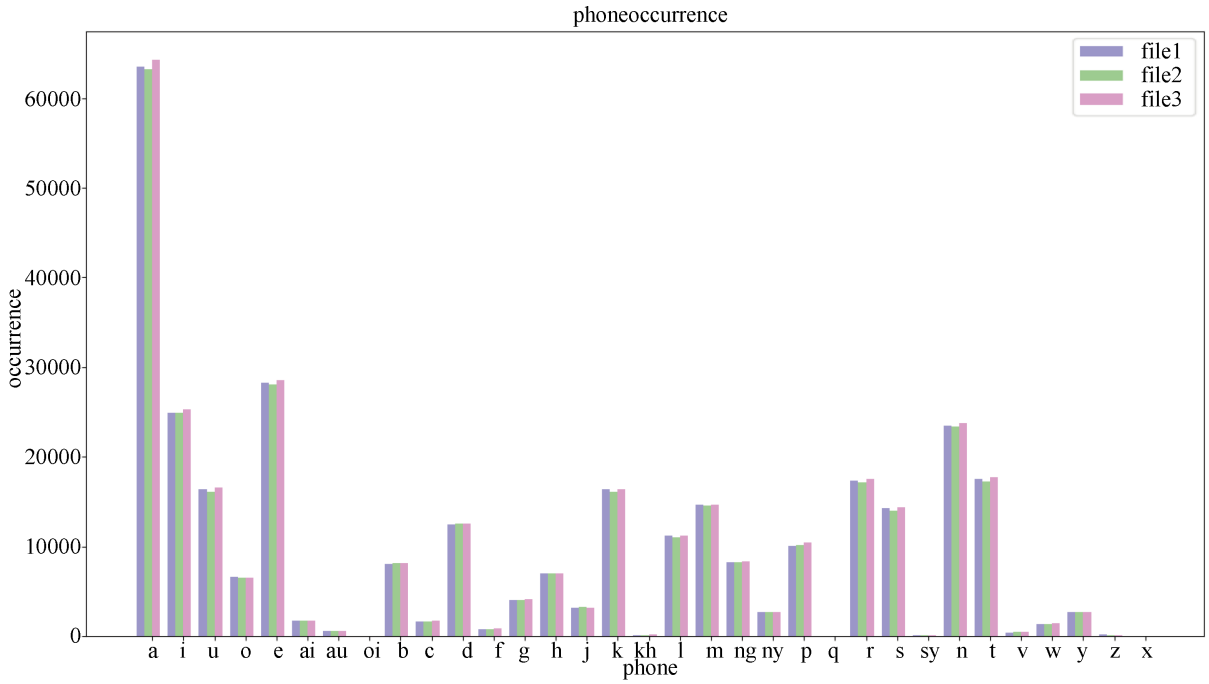

Figure 5. The number of phoneme in three sets of corpus
图 5. 三次产生发音语料库的音素出现次数

字“kode pos(邮编)”。若数字串既能匹配正则表达式且上下文中又出现关键字，则将数字串判断为邮编。

除上述三种情况读作数码，其余情况都读作数值[9]。

4.2.2. 特殊字符歧义判断

本文主要针对的是特殊字符“-”。在印尼语中，其出现的情况主要有：字母-字母(aba-aba)、数字-数字(1-2)、-数字(-2)。对于特殊字符“-”，采用分割后进行字符串匹配的方法进行歧义判断。

- 字母-字母

这种情况主要出现在两个词叠加形成一个新词时，如 aba-aba。此时采用的方法是判断其前后是否为字母：设“-”出现的位置为 i，则判断 i-1 及 i+1 是否为字母，若为字母，则“-”不存在歧义，此时不发音。

- 数字-数字

这种形式，可以表示电话号码、范围或减号。如“1-2”既可以读作 satusampaidua (一至二)，又可以读作 satukurangdua (一减二)。此时歧义判断的方法为：设字符“-”出现的位置为 i，判断 i-1 及 i+1 是否为数字，若为数字则需进行消歧处理。

- -数字

此种形式既可以表示负号，也可以表示序数词。如-2，可以读作 negatifdua (负二)，也可以读作 kedua (第二)。此时歧义判断的方法为：设字符“-”出现的位置为 i，判断 i+1 是否为数字，若为数字则需要进行消歧判断。

4.3. 消歧处理

本文主要针对阿拉伯数字“0~9”及特殊字符“-”消歧。对于纯数字的歧义，若判断为数码，则按照数码读法进行处理，其余情况按照数值读法进行消歧。对于特殊字符“-”，按照歧义判断结果进行消歧处理。

4.3.1. 数字的消歧处理

对于文本中的数字非标准词，若判断为数码的读法时，则按照表 1 所示印尼语数字将其转换为标准印尼语即可。例如出现数字“1”，转化为 satu。若不符合数码读法情况时，则按照数值的方法进行转换。具体转写规则详见文献[9]所述。

- 数值 0~9，按照表 1 印尼语数字写法进行转化。
- 数值 10 读为 sepuluh，数值 11 读为 sebelas。
- 数值 12~19，在个位数的基础上加上权位十(belas)，如 12 为 duabelas。
- 几十的时候，用权位“puluh”表示权位，如 20 为 duapuluh。
- 百位权重为 ratus，千位权重为 ribu，百万为 juta，十亿为 miliar。
- 除了如 10、100 此类的以“1”为开头的数值外，其余的数值都是加权位进行转换，而以“1”为开头的则是变为 se+权位进行转换，如数值 100 变为 seratus、数值 200 为 duaratus。
- 其余数值的读法和英语一样，如“100.000”读作“一百千”，即 seratusribu。

把数字转化为数值的过程中，首先要判断其是否为标准写法，即每三位用符号“.”进行隔开，若不是标准写法，需要对其进行改写。即从末尾一位开始向前进行计数，每三位插入一个“.”，直到整个数字串结束。在转化为标准词时，从最后一个“.”依次向前进行替换，分别替换为 ribu (千)、juta (百万)、miliar (十亿) [9]。

4.3.2. 特殊字符的消歧处理

特殊字符的消歧，通过字符串匹配判断出文本中的特殊字符“-”，搜寻上下文中的关键词，根据以下规则进行消歧处理：

- 数字-数字

这种情况出现时，首先判断“-”前面是否为四位数字 0062，若是的话再判断“-”后是否为 08 开头的 11 位纯数字。若符合上述情况则“-”不发音，数字按照数码的读法进行转写，如 0062-08123456789 转写为 kosong kosong enam dua kosong delapan satu dua tiga empat lima enam tujuh delapan sembilan。若不符合，则查找上下文是否存在关键字“samadengan (等于)”，若存在则将“-”转换为“kurang (减)”，如 3-2 转写为 tigakurangdua (三减二)。若上述情况都不符合，则默认为范围的读法，即“sampai (至)”，如 1-2 转化为 satusampaidua (一至二)。

Table 1. Indonesian digital writing
表 1. 印尼语数字写法

阿拉伯数字	印尼语数字
0	kosong
1	satu
2	dua
3	tiga
4	empat
5	lima
6	enam
7	tujuh
8	delapan
9	sembilian

- -数字

当此种情况出现时，先判断“-”前是否有 ke 出现，若出现，“-”后所跟数字转写为数值，“-”不发音，如 ke-2 转写为 kedua (第二)；若没有 ke 出现，则将“-”转化为“negatif (负)”，如 ke-2 转写为 negatifdua (负二)。

4.4. 归一化正确率

采用上述方法进行非标准文本的归一化处理，使文本中的非标准词转化为标准词。本次实验随机选取了发音语料库中的 500 个带有数字、特殊字符“-”的句子进行测试，按照特殊字符词数对归一化结果进行统计，正确率为 96.0%。其正确率未达 100.0%的原因在于特殊规则总结不够全面导致的非标准词转化产生歧义。

5. 音节划分

音节划分是在前端文本分析时，将给定的输入文本切分为音素序列或音节序列的过程。印尼语是一种以空格作为词边界、有确定的音节结构且音节中每个字母的发音都可以一对一映射到书面形式的拉丁字母语言[4]。在本实验中，音节的自动划分是基于音节列表和特殊规则完成的，它的意义在于为后端语音合成打下基础，以期改善合成语音的自然度及可懂度。

印尼语是一种黏着语。传统的印尼语音节结构有元音(Vowel)、辅－元音(Consonant-Vowel)、元-辅音(Vowel-Consonant)、辅－元－辅音(Consonant-Vowel-Consonant)四种。而辅－辅－元－辅音(CCVC)及辅－辅－辅－元音(CCCV)等结构则是出现在外来语中。例如“asrama(a/sra/ma)”是由 V/CCV/CV 结构组成的三音节词。印尼语单词的音节数最多有六个。

5.1. 音节划分方法

本文采用基于音节列表及特殊规则相结合的方法进行音节划分。基于音节列表的音节划分指的是选取词中最长的音节序列与音节列表进行匹配，若该音节序列存在于音节列表中，则将其分隔开；若不存在，则舍去一个字母继续进行匹配，直到该词中所有的字母都匹配到与音节列表相符的音节序列时划分结束。而运用特殊规则进行音节划分指的是在基于音节列表划分的基础上，加入一些特殊规则来提高音节划分的正确率。

此次实验将“/”作为音节边界。由于马来语和印尼语同属西印度尼西亚语支，并且两种语言之间具有相似性，故而在进行音节划分时，先运用已有的马来语音节列表对所构建的印尼语词典进行音节列表的音节划分[11]，并请印尼语专家对划分结果进行核对，利用核对后的结果更新音节列表，从而得到印尼语音节列表，之后运用印尼语音节列表对词典再次进行划分。将得到的结果进行错误原因统计，并进一步完善了音节划分的程序。本次实验采用的印尼语词典共计 22,220 个词汇。得到的印尼语音节列表共计 2167 个音节，对音节长度进行统计，如图 6 所示。

从图 6 可以看出，在整个音节列表中，最长的音节由五个音素组成。故而在音节自动划分的过程中，进行音节序列匹配的最大音素数目定为五个，并且采用逆向最大匹配的方法进行实验。即当程序读入一个词时，首先从词的末尾开始依此向前读入五个音素，将这五个音素组成的字母序列与音节列表进行匹配，若列表里存在该部分，则把其作为一个音节用“/”与从末尾读取的第六个音素分隔开，再从第六个音素向前读五个音素；若该五个音素在音节列表中没有与其匹配的音节序列时，则舍去其最左边的字母后，将剩余的四个字母与列表匹配，有的话用“/”分开，没有的话再舍去一个进行匹配。以此类推，直到该词的所有音素匹配完毕，再进行下一个词的音节划分。

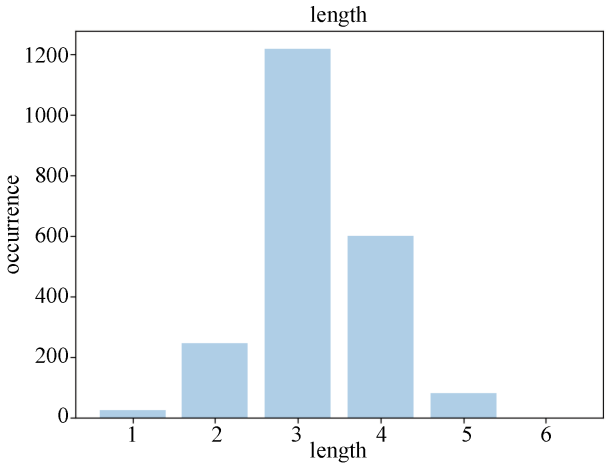

Figure 6. Syllable length distribution
图 6. 音节长度分布

然而，在音节划分的过程中存在一些错误，经过对错误的归纳总结后发现，当音节中出现“u”和“a”、“i”和“a”等连用的情况时，会出现划分错误。其错误的主要原因在于，当“u”、“a”连在一起进行发音时，有零声母现象存在，即“a”前面会省略声母“w”，如：“adikuasa”，用程序进行音节划分时得到的音节化结果为“a/di/kua/sa”，但经印尼语专家核对后其正确的划分方法为“a/di/ku/(w)a/sa”，而“i”、“a”连用在一起发音时，会省略“a”前面的声母“y”。为了提高音节划分的正确率，使后端语音合成得到良好的效果，我们采用的方法是在原有方案的基础上加上零声母规则，即当进入五个音素进行匹配时，若匹配到音节列表中有相对应的音节存在，在此基础上再进一步判断是否存在使零声母现象发生的音素连用情况，若存在的话，从这两个音素中间“/”用分开，进行二次匹配；若不存在的话，直接进行划分。

5.2. 音节划分正确率

采用基于音节列表和零声母规则相结合的方法，使音节划分的正确率有了极大提高。为验证实验正确率，把人工划分音节的词典作为集内测试数据，而将文本语料作为集外测试数据。

对词典音节划分后进行正确率统计，统计结果表明，词典中随机挑选的 1000 个词划分后正确率达到 98.2%，从文本语料中随机挑选的 480 个句子共计 5850 个单词划分，对其进行音节划分，正确率为 97.1%。

集内测试和集外测试正确率存在差异的原因主要在于，音节列表不够全面，文本语料中出现音节列表中未添加的音节；文本中存在外来借词，对未按印尼语规则书写的外来借词划分出现错误。

6. 结语

本文主要针对印尼语语音合成系统中的前端文本分析模块进行实验，实现了印尼语语音合成系统中的发音语料库构建，文本归一化以及音节划分。为印尼语后期语音合成工作提供了良好的数据准备，具有工程应用价值。本文的后续研究工作包括：对文本语料进行归一化处理时，在考虑数字及特殊字符的基础上，引入缩略词等其他情况。基于音节列表的音节划分，列表的完整性直接影响到音节划分的正确率，后期还需不断完善音节列表实现更好的划分效果。在进行音节划分时，加入的特殊规则只考虑了零声母的情况，而音节划分中对于外来借词的划分是出现错误的主要原因，在后期应该加入外来词音节划分的规则。

致　谢

作为印尼语专家，张会叶老师为本文研究提供了大力支持，在此致谢。

基金项目

本文获国家自然科学基金项目(61262068)资助。

参考文献

[1] Gunarso, M., Uliniansyah, T., Santosa, A., *et al*. (2016) Development of a Speech Corpus for an Indonesian Text-to-Speech System. 2016 *Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique* (*O-COCOSDA*), Bali, 26-28 October 2016, 258-261.

[2] Mengko, R. and Ayuningtyas, A. (2013) Indonesian Text-to-Speech System Using Syllable Concatenation: Speech Optimization. *International Conference on Instrumentation*, *Communications*, *Information Technology*, *and Biomedical Engineering* (*ICICI-BME*), Bandung, 7-8 November 2013, 412-415.

[3] Sutarman (2015) Indonesian Text-to-Speech System Using Diphone Concatenative Synthesis. *International Journal of Software Engineering & Computer Sciences* (*IJSECS*), 85-93.

[4] Dong, Y. and Li, D. 语音识别实践[M]. 俞凯, 钱彦旻, 等, 译. 北京: 电子工业出版社, 2016: 263-266.

[5] 张会叶. 印度尼西亚语词缀研究[D]: [硕士学位论文]. 昆明: 云南民族大学, 2009: 3-8.

[6] Koto, F. (2016) A Publicly Available Indonesian Corpora for Automatic Abstractive and Extractive Chat Summarization. 2016 *Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique* (*O-COCOSDA*), Bali, 26-28 October 2016, 293-297.

[7] Cahyaningtyas, E. and Arifianto, D. (2015) HMM-Based Indonesian Speech Synthesis System with Declarative and Question Sentences Intonation. 2015 *International Symposium on Intelligent Signal Processing and Communication Systems* (*ISPACS*), Bali, 9-12 November 2015, 153-158. https://doi.org/10.1109/ISPACS.2015.7432756

[8] Indonesian Language. http://encyclopedia.thefreedictionary.com/Indonesian+language

[9] Kong, X. and Yang, J. (2018) Indonesian Corpus Constructing and Text Processing for Speech Synthesis. *The International Conference on Asian Language Processing* (*IALP*), Bandung, 15-17 November 2018.

[10] Indra, Z., Jaafar, J. and Zamin, N. (2015) A Language Identifier for Indonesian and Malay Text Document. 2015 *International Symposium on Mathematical Sciences and Computing Research* (*iSMSC*), 127-131.

[11] 施梅芳. 面向语音合成的马来语文本分析[D]: [硕士学位论文]. 昆明: 云南大学, 2018: 7-17.