JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 40, 283-302 (2024)
DOI: 10.6688/JISE.202403_40(2).0006

Development of a Taiwanese Speech Synthesis System Using Hidden Markov Models and a Robust Tonal Phoneme Corpus[*]

YUNG-JI SHER[1,+], MING-CHUN HSU[2], YU-HSIEN CHIU[3], YEOU-JIUNN CHEN[5],
CHUNG-HSIEN WU[4] AND JIUNN-LIANG WU[6]

[1]*Department of Special Education and Graduate Institute of Rehabilitation Counseling*
National Taiwan Normal University
Taipei, 106 Taiwan
[2]*Wistron Neweb Corporation*
Hsinchu, 300 Taiwan
[3]*NSTC AI Biomedical Research Center*
[4]*Department of Computer Science and Information Engineering*
National Cheng Kung University
Tainan, 701 Taiwan
[5]*Department of Electrical Engineering*
Southern Taiwan University of Science and Technology
Tainan, 710 Taiwan
[6]*Department of Otorhinolaryngology, National Cheng Kung University Hospital*
College of Medicine, National Cheng Kung University
Tainan, 704 Taiwan
E-mail: siaa@ntnu.edu.tw[1,+]; smjeremy@gmail.com[2]; chiuyuhsien@gmail.com[3];
chenyj@stust.edu.tw[5]; chwu@csie.ncku.edu.tw[4]; jiunn@mail.ncku.edu.tw[6]

The number of young native speakers of Taiwanese, the variant of Southern Min spoken in Taiwan, has decreased. Technological advancements such as text-to-speech (TTS) systems could help arrest this decline. The aim of this study was to design a robust tonal phoneme corpus and a speech synthesis system for Modern Literal Taiwanese (MLT). MLT subsyllables were analyzed using phonetics and phonology to establish tonal phoneme models. These robust tonal phoneme models and hidden Markov models were used to construct an MLT TTS synthesis system. Algorithm-based training resulted in 869 balanced sentences containing 12,544 syllables, with each sentence containing an average of 14.4 syllables. In total, 218 sentences, which included rare phonemes, were manually drafted to supplement the corpus. The synthesized phonemes were deemed to have high intelligibility and could be included in the developed TTS system. According to the HTK speech recognition tool, the overall phoneme recognition rate was 96.47%. Testers, who were native Taiwanese speakers, assigned the synthesized sentences a mean opinion score of 4, indicating that they sounded natural. This developed system and the results described herein can inspire future developments in speech technology and computational linguistics.

Keywords: Taiwanese, text-to-speech, speech corpus, hidden Markov model, modern literal Taiwanese

1. INTRODUCTION

Language is a means of communication, strengthening emotional identity, and passing on cultures [1, 2]. However, linguists predict that 90% of all languages worldwide

Received April 12, 2022; revised November 29, 2022; accepted April 14, 2023.
Communicated by Berlin Chen.

[+] Corresponding author.

[*] This project was supported by the funding of the National Science Council (NSC 96-2221-E-003-023-MY3) and Ministry of Science and Technology (MOST 109-2511-H-003-037-) in Taiwan.

(> 3000) will vanish by 2100 [3-5]. Researchers [6] estimate that Southern Min, or Minnan (a language group that includes Amoy), is spoken by approximately 50.1 million people, ranking 27th worldwide. The variant of Southern Min spoken in Taiwan is collectively referred to as Taiwanese in the present study and is the native language of approximately 76.9% of Taiwan's population [7-9]. The continual migration toward northern Taiwan, where Mandarin is the dominant language, has led to a decrease in young fluent speakers of Taiwanese [10, 11] and a heritage crisis [12, 13]. The government therefore enacted the Development of National Languages Act in 2019 [14] to guarantee equality among all national languages in Taiwan and to encourage research into language development.

The Taiwanese language features many tones, nasals, and sandhi [15-19]. Liim Keahiong [20, 21] developed Modern Literal Taiwanese (MLT), which uses the Latin alphabet to represent tones and nasals and spells out sandhi to form multisyllable words with complete semantic meanings. The eight tones under MLT are a raised tone (Tone 1), pushed out tone (Tone 2), depressed tone (Tone 3), low stop tone (Tone 4), bend tone (Tone 5), bend-up tone (Tone 6), fundamental tone (Tone 7), and high stop tone (Tone 8).

2. TEXT-TO-SPEECH SYSTEM

Klatt released a text-to-speech (TTS) system in 1980 that introduced speech synthesis technology and speech quality assessment [22]. Pitch Synchronous Overlap and Add, first introduced by France Telecom, improves the naturalness and clarity of synthesized speech [23]. Hidden Markov model (HMM) may be applied to capture acoustic parameters to synthesize steady and fluent speech, and they have the advantages of being portable and adaptable [24, 25].

3. TAIWANESE TTS SYSTEM

Sin-Horng Chen analyzed the phonetic characteristics of Taiwanese to develop a TTS system [26-29]. Hsin-Hsi Chen developed an online Taiwanese translation system using a parts-of-speech parser [30, 31]. Ren-Yuan Lyu [32, 33] and Kao-Chi Chung [19] have developed Taiwanese TTS systems using syllables as the synthesis units.

4. RESEARCH PURPOSE AND AIMS

This research purpose was to design and establish a robust tonal phoneme corpus and develop an MLT speech synthesis system. The research aims were to (1) design an algorithm to train and establish a speech corpus of balanced sentences in MLT; (2) analyze and establish the robustness of MLT-based tonal phoneme models using statistical methods; and (3) develop a TTS system through applying HMMs and an MLT tonal phoneme corpus.

5. MATERIALS AND METHODS

This study analyzed MLT sub-syllables using phonetics and phonology to establish tonal phoneme models. A training algorithm was designed for an MLT-based balanced speech corpus. The HMM Toolkit (HTK) was applied to recognize tonal phonemes and validate the robustness of an MLT-based tonal phoneme set through a Bayes screening test. The robust tonal phoneme models and HMMs were used to develop an MLT-TTS system.

The modeling detail and processing pipeline in this research are: First, to analyze MLT sub-syllables through Taiwanese phonetics and phonology to establish tonal phoneme models; Second, to design a training algorithm for Taiwanese balanced speech database; Then, to apply HMM Toolkit (HTK) to recognize tonal phoneme and validate the robust Taiwanese tonal phoneme set through Bayes screening test; Finally, to apply the robust tonal phoneme models and HMM-based TTS system to develop Taiwanese TTS system. A statistics-based syllable pitch contour model for Mandarin speech is proposed by Chen [34]. This approach takes the mean and the shape of a syllable log-pitch contour as two basic modeling units and it also gives a quantitative and more complete description of the co-articulation effect of neighboring tones rather than conventional qualitative descriptions of the tone sandhi rules [34]. Our HMM MLT TTS system established in this research should be considered in the future works.

5.1 Establishment of MLT Tonal Phoneme Models

Taiwanese is a tonal language, and therefore, when synthesis systems are being trained, a speech corpus with tones is required for the synthesized speech to exhibit tonal changes [35]. MLT contains 2,805 tonal syllables, which are further distinguished into 9 types of sub-syllables and 63 phonemes. Table 1 presents the established tonal phoneme model.

$$Syllable = Consonant + Glide + Vowel + Nasal(CGVN) \tag{1}$$

Table 1. MLT subsyllables and phonemes.

Sub-syllables	Phonemes	No.
consonant	b, c, ch, g, h, j, k, kh, l, m, n, p, ph, s, t, th, z, zh	18
nasal consonant	ch^, c^, g^, h^, kh^, k^, ph^, p^, s^, th^, t^, zh^, z^	13
glide	i-, o-	2
nasal glide	i~, o~	2
Beginning vowel	a:, e:, i:, o:, 0:, u:, ng:	7
vowel	a, e, i, m, ng, o, 0, u	8
nasal vowel	a^, e^, i^, o^, u^	5
Beginning nasal vowel	a@, e@, i@, o@, u@	5
nasal	m_, n_, ng_	3

5.2 Development of an MLT Balanced Speech Corpus

This involved the collection of a text corpus and the design of an algorithm for training balanced sentences to identify sentences containing rare phoneme units.

5.3 Text Database Collection

Text databases were collected from websites and audiobooks, such as the MLT website [21] and the audio version of the *Medicinal Textbook in Modern Literal Taiwanese* [36].

5.4 Design of a Training Algorithm for Balanced Sentences

The training algorithm consists of spell-checking, screening, unit scoring, right context dependency score determination, and sentence ranking.

(A) Spell-checking and screening

Spell-checking involves using spelling rules to detect and correct misspellings and typographical errors. Screening is used to delete overly long or overly short sentences; this is conducive to recording fluent speech, which ensures the precision and stability of database training.

(B) Unit scores

Unit scores are the weights of units determined based on the frequency with which individual units appear. The less frequently a unit appears, the more weight it carries. The equation is as follows,

$$UnitScore_i = \sum_{j=1}^{U_i} \frac{Unit_{all\ times}}{Unit_{i,j}}. \tag{2}$$

Unit Score$_i$ is the score of the unit weight of sentence I, and $Unit_{all\ times}$ is the total number of appearances of that unit. $Unit_{i,j}$ is the number of times that unit j appears in sentence i. U_i is the number of units in sentence i.

(C) Right context dependency

Right context dependency (RCD) is the sequence formed by a unit and those that follow. For example, Taioaan (Taiwan in Taiwanese) can be broken into the phonemes {t, aM, iM, oHead, aG, aML, and nNLM}; the RCD sequence is {{t-aM}, {aM-iM}, {iM-oHead}, {oHead-aG}, {aG-aML}, and {aML-nNLM}}. The formula for calculating the RCD score is as follows,

$$RCDScore_i = \sum_{j=1}^{R_i} \frac{RCD_{all\ times}}{RCD_{i,j}}. \tag{3}$$

RCD $Score_i$ is the weight of the RCD of sentence i, and $RCD_{all\ times}$ is the total number of times that RCD is present. RCD$_{i,j}$ is the number of times that RCD j is present in sentence i. R_i is the number of times RCD is present in sentence i.

(D) Sentence ranking

The total score of each sentence, calculated by summing its unit score and RCD score, determines the amount of phonetic information in a sentence, forming the basis for ranking sentences by importance. The formula is as follows,

$$Sentence\ Score_i = Unit\ Score_i + RCD\ Score_i. \tag{4}$$

Sentence $Score_i$ is the total score for sentence i, Unit $Score_i$ is the unit score of sentence i, and RCD $Score_i$ is the RCD score of sentence i.

5.5 Design of Sentences Containing Rare Phoneme Units

After the collected text databases and the balanced speech corpus training results were analyzed, the frequency of some phonemes was found to be low. These rare phonemes were mostly nasals, as indicated in Table 2. Designing sentences that contain these phonemes can supplement data on rare phonemes without having to spend time compiling a rare phoneme corpus [19].

Table 2. Number of appearances of 21 rare phonemes.

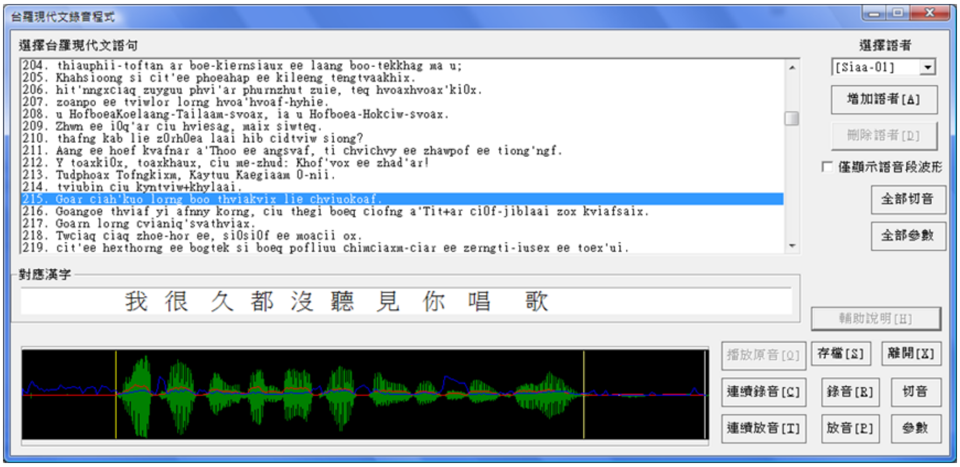
Phonemes	Appearances	Phonemes	Appearances	Phonemes	Appearances
uVHead	1	eVLS	2	zhV	7
mmL	1	iVMS	2	aVHead	7
mmLM	1	iVLS	2	aVMS	7
aVLS	1	mmML	3	oVML	9
eVML	2	eVMS	4	oVLM	9
eVLM	2	mmH	6	eVL	10
eVHS	2	aVHS	6	eVHead	14

To record a rare phoneme, that phoneme is placed in the middle of sentence to control for its pitch, duration, and volume and retain high-quality phonological data. The sentence containing a rare phoneme has the following structure:

Goar thak「rare phoneme」piauzurn.

5.6 Recording an MLT Balanced Speech Corpus

Recordings were made using a unidirectional condenser microphone with noise cancelling technology and a bandpass filter of 100–17 kHz. The sampling rate was 16 kHz, and the resolution was 16 bits. A total of 1,087 sentences were recorded, comprising 869 balanced sentences and 218 sentences containing rare phonemes (Fig. 1). The gender of the speaker in the recorded speech utterances is male.



```
204. thiauphii-toftan ar boe-kiernsiaux ee laang boo-tekkhag ma u;
205. Khahsioong si cit'ee phoeahap ee kileeng tengtvaakhix.
206. hit'nngxciaq zuyguu phvi'ar phurnzhut zuie, teq hvoaxhvoax'ki0x.
207. zoanpo ee tviwlor lorng hvoa'hvoaf-hyhie.
208. u HofboeaKoelaang-Tailaam-svoax, ia u Hofboea-Hokciw-svoax.
209. Zhwn ee i0q'ar ciu hviesag, maix siwteq.
210. thafng kab lie z0rh0ea laai hib cidtviw siong?
211. Aang ee hoef kvafnar a'Thoo ee angsvaf, ti chvichvy ee zhawpof ee tiong'ngf.
212. Y toaxki0x, toaxkhaux, ciu me-zhud: Khof'vox ee zhad'ar!
213. Tudphoax Tofngkixm, Kaytuu Kaegiaam 0-nii.
214. tviubin ciu kyntviw+khylaai.
215. Goar ciah'kuo lorng boo thviakvix lie chviuokoaf.
216. Goangoe thviaf yi afnny korng, ciu thegi boeq ciofng a'Tit+ar ci0f-jiblaai zox kviafsaix.
217. Goarn lorng cvianiq'svathviax.
218. Twciaq ciaq zhoe-hor ee, si0si0f ee moacii ox.
219. cit'ee hexthorng ee bogtek si boeq pofliuu chimciaxm-ciar ee zerngti-iusex ee toex'ui.
```

[Siaa-01]

我 很 久 都 沒 聽 見 你 唱 歌

Fig. 1. MLT sound recorder interface.

5.7 Development of Robust Tonal Phoneme Set

The HTK developed by Cambridge University was used to recognize tonal phonemes and to analyze and verify their robustness [29, 36].

5.8 Tonal Phoneme Recognition

This involved capturing feature parameters, training acoustic models, and recognizing speech [29].

(A) Feature parameter capture

The audio files were set to have sound frame lengths of 20 ms and overlap lengths of 10 ms. These files were also set to use 1-dimensional log energy, 12-dimensional mel-frequency cepstral coefficient (MFCC), 13-dimensional delta MFCC, and 13-dimensional delta-delta MFCC. Following the preemphasis and frame blocking of a speech signal, the signal's log was taken to calculate the sum of the square values of each sample point within the sound frame.

(B) Acoustic model training

The state observation probability was employed to determine whether to remain in the original state or proceed to the next state. A Gaussian mixture model was adopted as the state observation probability function, in which forward and backward algorithms were used to determine the probability of a sequence being produced in the following. Subsequently, Baum-Welch reestimation algorithm was used to repeatedly adjust the parameters until the results converged, resulting in a fully trained acoustic model.

5.9 Speech Recognition

The Viterbi reestimation algorithm was used to determine the optimal path by recognizing and comparing acoustic models. After recognition was completed, recognition files for the input sentences were generated; the files contained tonal phoneme sequences that were identified in each sentence. These sequences were assessed according to the figure of merit metric of the US National Institute of Standards and Technology, and optimized comparisons were performed using dynamic programming to determine the rates of sentence and word recognition. These results were then used to assess the robustness of the acoustic models and recognizers.

(A) Analysis and validation of robust tonal phonemes in MLT

On the basis of Bayes' rule, unit confusion matrices were built to calculate sensitivity and specificity and generate receiver operating characteristic (ROC) curves. These results were used to validate the robustness of the intelligibility of the tonal phonemes.

(B) Development of an MLT-TTS system using an HMM

Modifications by the HMM/DNN-based Speech Synthesis System (HTS) working group [37] using the core technology of the HTK voice recognition tool means the HTS already has its own API [38].

(C) Training system

The proposed training system included a tonal phoneme unit set, a Taiwanese balanced speech database (for training purposes), a text analysis processing module, an acoustic parameter capture module, a context clustering-dependent problem set, and an HMM training module.

1. Tonal phoneme unit set

Through the use of 2,805 MLT tonal syllables, which included 157 units, the volumes of synthesized data and calculations were substantially compressed to obtain a state of balance. As a result, all possible MLT syllables could be combined with a small number of unit models.

2. MLT balanced speech corpus for training

The speech corpus for training contains speech files and their corresponding text information. The speech files were tagged using Cool Edit Pro version 2.0, which was developed by Syntrillium Software Corporation for audio editing. The corresponding text information was processed using a text analysis and tagging program developed by this study, resulting in context clustering-dependent tonal phonemes and the automatic generation of tagged text files.

3. Text analysis and processing module

This step involved segmenting MLT sentences, capturing tonal phonemes, and generating context clustering-dependent sequences. MLT spelling rules were also imposed.

4. Acoustic parameter capture module

This step involved capturing the acoustic parameters of speech signals, including the excitation parameter (log F_0 and its dynamic acoustic parameters) and the frequency spectrum parameter (the MFCC and its dynamic acoustic parameters), for HMM training.

5. Context clustering-dependent problem set

The text analysis and processing module was designed a problem set that correspond to different parameter changes to improve the state merging and splitting results. The five classes used in this study were phoneme, syllable, word, phrase, and sentence.

6. HMM training module

This step involved generating context clustering-dependent HMMs for training the decision tree system. The step, which can be distinguished into model initialization, tonal phoneme and duration modeling, and finally the production of fundamental frequency (F_0), spectrum, and duration models.

5.10 Synthesis System

The synthesis system features text analysis and processing modules, context-dependent HMMs, a speech parameter generation algorithm [39], and synthesis filtering modules.

(A) Text analysis and processing

The synthesis process transforms input text directly into a synthesized speech output. The target of processing is text that corresponds to the desired speech output; the text is immediately transformed after input and used to generate speech parameters for speech synthesis.

(B) Context-dependent HMMs

The HMM-trained decision tree system was used to perform clustering and generate pitch, spectrum, and duration.

(C) Speech parameter generation algorithm and the synthesis filter

Speech parameters were generated using an algorithm [39]. Subsequently, an excitation signal generator and a mel log spectrum approximation filter were used to directly restore and synthesize the excitation and spectrum parameters into a speech signal in Taiwanese [40].

5.11 Speech Quality Assessment

The synthesized speech should have sufficient quality such that the produced speech is comprehensible, natural, and smooth to users and resembles human speech. Speech quality assessment was used to determine the intelligibility of the synthesized speech signals. Subjective assessments, such as mean opinion score (MOS), involved five grade levels as follows: 5, excellent; 4, good; 3, fair; 2, poor; 1, unsatisfactory. The assessed sentences can be further distinguished into parts that were either inside or outside the speech corpus. For the first part, 15 sentences were randomly selected from the corpus and synthesized into speech as the Inside testing corpus. For the second part, 15 newly written sentences were synthesized into speech as the Outside testing corpus.

The testees were 30 adults fluent in Taiwanese. Prior to testing, the testees received training on the scoring standards. They were asked to grade each sentence from 1 to 5; the median of the grading results for 30 sentences was calculated and set as the system MOS, which was the standard for assessing the naturalness of the synthesized speech.

6. RESULTS AND DISCUSSION

The collected text corpus consisted of 8,905 MLT sentences and 100,000 syllables from website and books. An MLT balanced speech corpus featuring 869 MLT sentences was established using a training and analysis system developed through Windows programming, and another 218 sentences containing rare phoneme units were generated for inclusion in the database. A phonetic set of 156 Taiwanese tonal phonemes was generated from the HTK recognition results, and the robustness of the phonetic set was validated through sensitivity and specificity analysis and a ROC curve. The HMM-based Taiwanese TTS system was developed using Linux and Windows operating systems, and the synthetic speech samples were determined to have an MOS of 4, indicting good naturalness. The study results provide fundamental information and new techniques for the development of indigenous clinical speech technology and Taiwanese computational linguistics.

6.1 Development of MLT Tonal Phoneme Sets

Initially, 176 tonal phonemes were established; 19 of those units were found to be redundant and were removed, leaving 157 units. This substantially reduced the volume of data for calculation, resulting in increased efficiency. The removed phonemes were in 10 major categories (18 consonants, 13 nasal consonants, 2 glides, 2 nasal glides, 7 beginning monophthong, 58 monophthongs, 33 nasal monophthongs, 5 beginning nasal monophthongs, 18 nasal syllable codas, and 1 silent segment). The 157 tonal phonemes are presented in Table 3.

Table 3. Taiwanese tonal phoneme set.

Categories	157 Phonemes in 10 Categories	Qty
Consonants	b, c, ch, g, h, j, k, kh, l, m, n, p, ph, s, t, th, z, zh	18
Nasal consonants	chV, cV, gV, hV, khV, kV, phV, pV, sV, thV, tV, zhV, zV	13
Glides	iG, oG	2
Nasal glides	iVG, oVG	2
Beginning monophthongs	aHead, eHead, iHead, oHead, 0Head, uHead, ngHead	7
Monophthongs	aH, aM, aL, aML, aLM, aHS, aMS, aLS eH, eM, eL, eML, eLM, eHS, eMS, eLS iH, iM, iL, iML, iLM, iHS, iMS, iLS mmH, mmM, mmL, mmML, mmLM ngH, ngM, ngL, ngML, ngLM, oH, oM, oL, oML, oLM, oHS, oMS, oLS 0H, 0M, 0L, 0ML, 0LM, 0HS, 0MS, 0LS uH, uM, uL, uML, uLM, uHS, uMS, uLS	58
Nasal monophthongs	aVH, aVM, aVL, aVML, aVLM, aVHS, aVMS, aVLS eVH, eVM, eVL, eVML, eVLM, eVHS, eVMS, eVLS iVH, iVM, iVL, iVML, iVLM, iVMS, iVLS oVH, oVM, oVL, oVML, oVLM, uVH, uVM, uVL, uVML, uVLM	33
Beginning nasal monophthongs	aVHead, eVHead, iVHead, oVHead, uVHead	5
Nasal syllable codas	mNH, mNM, mNL, mNLM, mNMS, mNLS nNH, nNM, nNL, nNLM, nNMS, nNLS, ngNH, ngNM, ngNL, ngNLM, ngNMS, ngNLS	18
Silence	silence	1

6.2 Development of an MLT Speech Corpus using Balanced Sentences

The training corpus comprised 8,905 MLT sentences. After training an algorithm with balanced sentences developed in this study, window programming languages were used to develop a training and analysis system that could effectively determine sentence scores and ranking. The initial screening yielded 869 balanced sentences. After the inclusion of 218 sentences containing rare phonemes, 1,087 audio files were recorded.

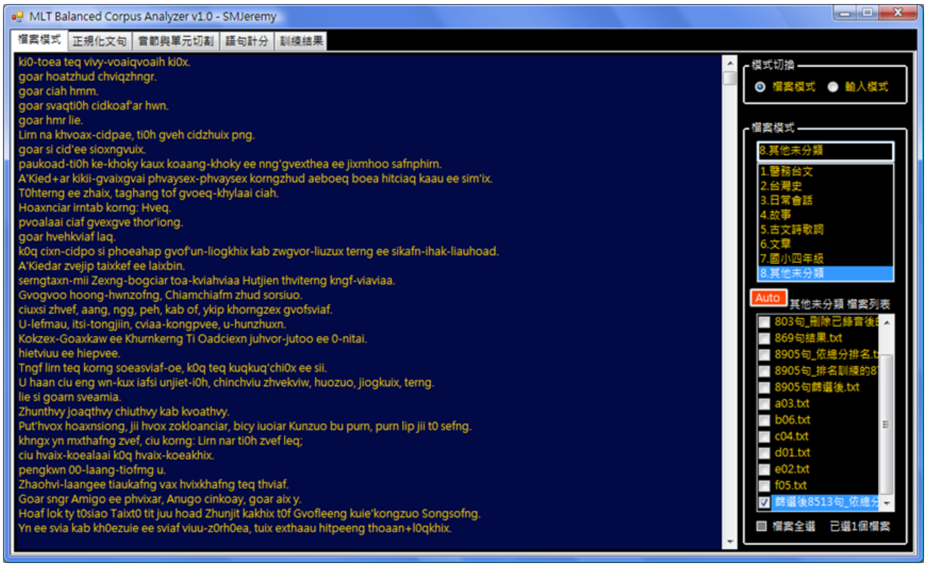
(A) Corpus of continuous sentences

The corpus of continuous sentences included 245 articles from the MLT promotion website [21] and the text from an audiobook version of the *Medicinal Textbook in Modern Literal Taiwanese*. The collected corpus consisted of 8,905 MLT sentences, with an average of 11.8 syllables per sentence.

(B) Balanced sentence analysis system and training results

The balanced sentence training and analysis system was developed using Microsoft Visual Studio 2005 IDE in Windows Vista, and the programming language Visual C#.Net was used to design the program interface. The corpus was subjected to spell-checking, screening, unit scoring, RCD scoring, sentence ranking, and balanced sentence selection. The execution interface was divided into corpus file reading (Fig. 2), sentence normalization (Fig. 3), syllabic and unit segmentation (Fig. 4), sentence scoring (Fig. 5), and balanced sentence training result (Fig. 6) sections.

The corpus file reading section has two modes: file and input modes. Users can select multiple files or enter sentences for analysis. Fig. 2 presents a screenshot of the related interface's operation.



```
ki0-toea teq vivy-voaiqvoaih ki0x.
goar hoatzhud chviqzhngr.
goar ciah hmm.
goar svaqti0h cidkoaf'ar hwn.
goar hmr lie.
Lirn na khvoax-cidpae, ti0h gveh cidzhuix png.
goar si cid'ee sioxngvuix.
paukoad-ti0h ke-khoky kaux koaang-khoky ee nng'gvexthea ee jixmhoo safnphirn.
A'Kied+ar kikii-gvaixgvai phvaysex-phvaysex korngzhud aeboeq boea hitciaq kaau ee sim'ix.
T0hterng ee zhaix, taghang tof gvoeq-khylaai ciah.
Hoaxnciar irntab korng: Hveq.
pvoalaai ciaf gvexgve thor'iong.
goar hvehkviaf laq.
k0q cixn-cidpo si phoeahap gvof'un-liogkhix kab zwgvor-liuzux terng ee sikafn-ihak-liauhoad.
A'Kiedar zvejip taixkef ee laixbin.
serngtaxn-mii Zexng-bogciar toa-kviahviaa Hutjien thviterng kngf-viaviaa.
Gvogvoo hoong-hwnzofng, Chiamchiafm zhud sorsiuo.
ciuxsi zhvef, aang, ngg, peh, kab of, ykip khorngzex gvofsviaf.
U-lefmau, itsi-tongjiin, cviaa-kongpvee, u-hunzhuxn.
Kokzex-Goaxkaw ee Khurnkerng Ti Oadciexn juhvor-jutoo ee 0-nitai.
hietviuu ee hiepvee.
Tngf lirn teq korng soeasviaf-oe, k0q teq kuqkuq'chi0x ee sii.
U haan ciu eng wn-kux iafsi unjiet-i0h, chinchviu zhvekviw, huozuo, jiogkuix, terng.
lie si goarn sveamia.
Zhunthvy joaqthvy chiuthvy kab kvoathvy.
Put'hvox hoaxnsiong, jii hvox zokloanciar, bicy iuoiar Kunzuo bu purn, purn lip jii t0 sefng.
khngx yn mxthafng zvef, ciu korng: Lirn nar ti0h zvef leq;
ciu hvaix-koealaai k0q hvaix-koeakhix.
pengkwn 00-laang-tiofmg u.
Zhaohvi-laangee tiaukafng vax hvixkhafng teq thviaf.
Goar sngr Amigo ee phvixar, Anugo cinkoay, goar aix y.
Hoaf lok ty t0siao Taixt0 tit juu hoad Zhunjit kakhix t0f Gvofleeng kuie'kongzuo Songsofng.
Yn ee svia kab kh0ezuie ee sviaf viuu-z0rh0ea, tuix exthaau hitpeeng thoaan+l0qkhix.
```

模式切換
◉ 檔案模式 ○ 輸入模式

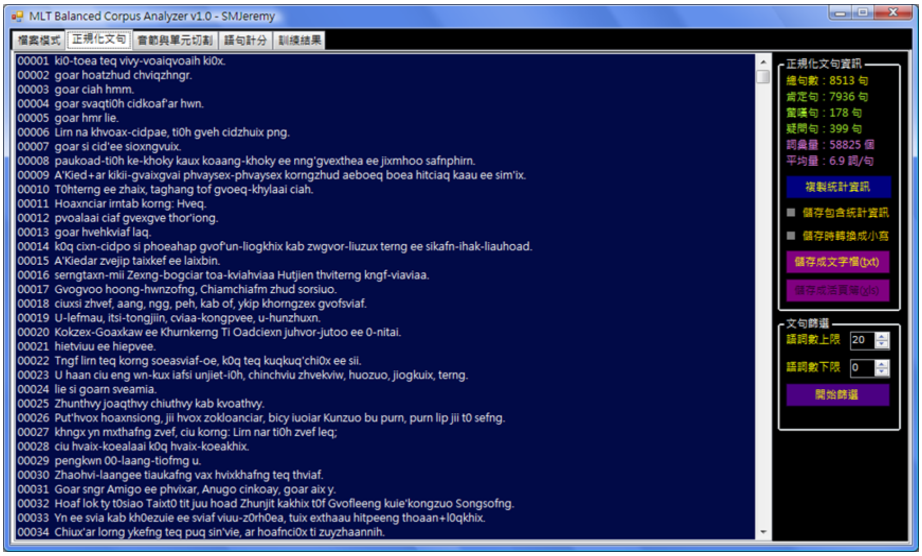
檔案模式
8.其他未分類
1.醫務台文
2.台灣史
3.日常會話
4.故事
5.古文詩歌詞
6.文章
7.國小四年級
8.其他未分類

Auto 其他未分類 檔案列表
□ 803句_刪除已錄音後
□ 869句結果.txt
□ 8905句_依總分排名.t
□ 8905句_排名訓練的8
□ 8905句篩選後.txt
□ a03.txt
□ b06.txt
□ c04.txt
□ d01.txt
□ e02.txt
□ f05.txt
☑ 篩選後8513句_依總分

□ 檔案全選 已選1個檔案

Fig. 2. Screenshot of file reading operation in the balanced sentence training and analysis system.

Sentence normalization involves the summation of misspellings and vocabulary; it can be used to filter sentences based on word length (Fig. 3).



```
00001 ki0-toea teq vivy-voaiqvoaih ki0x.
00002 goar hoatzhud chviqzhngr.
00003 goar ciah hmm.
00004 goar svaqti0h cidkoaf'ar hwn.
00005 goar hmr lie.
00006 Lirn na khvoax-cidpae, ti0h gveh cidzhuix png.
00007 goar si cid'ee sioxngvuix.
00008 paukoad-ti0h ke-khoky kaux koaang-khoky ee nng'gvexthea ee jixmhoo safnphirn.
00009 A'Kied+ar kikii-gvaixgvai phvaysex-phvaysex korngzhud aeboeq boea hitciaq kaau ee sim'ix.
00010 T0hterng ee zhaix, taghang tof gvoeq-khylaai ciah.
00011 Hoaxnciar irntab korng: Hveq.
00012 pvoalaai ciaf gvexgve thor'iong.
00013 goar hvehkviaf laq.
00014 k0q cixn-cidpo si phoeahap gvof'un-liogkhix kab zwgvor-liuzux terng ee sikafn-ihak-liauhoad.
00015 A'Kiedar zvejip taixkef ee laixbin.
00016 serngtaxn-mii Zexng-bogciar toa-kviahviaa Hutjien thviterng kngf-viaviaa.
00017 Gvogvoo hoong-hwnzofng, Chiamchiafm zhud sorsiuo.
00018 ciuxsi zhvef, aang, ngg, peh, kab of, ykip khorngzex gvofsviaf.
00019 U-lefmau, itsi-tongjiin, cviaa-kongpvee, u-hunzhuxn.
00020 Kokzex-Goaxkaw ee Khurnkerng Ti Oadciexn juhvor-jutoo ee 0-nitai.
00021 hietviuu ee hiepvee.
00022 Tngf lirn teq korng soeasviaf-oe, k0q teq kuqkuq'chi0x ee sii.
00023 U haan ciu eng wn-kux iafsi unjiet-i0h, chinchviu zhvekviw, huozuo, jiogkuix, terng.
00024 lie si goarn sveamia.
00025 Zhunthvy joaqthvy chiuthvy kab kvoathvy.
00026 Put'hvox hoaxnsiong, jii hvox zokloanciar, bicy iuoiar Kunzuo bu purn, purn lip jii t0 sefng.
00027 khngx yn mxthafng zvef, ciu korng: Lirn nar ti0h zvef leq;
00028 ciu hvaix-koealaai k0q hvaix-koeakhix.
00029 pengkwn 00-laang-tiofmg u.
00030 Zhaohvi-laangee tiaukafng vax hvixkhafng teq thviaf.
00031 Goar sngr Amigo ee phvixar, Anugo cinkoay, goar aix y.
00032 Hoaf lok ty t0siao Taixt0 tit juu hoad Zhunjit kakhix t0f Gvofleeng kuie'kongzuo Songsofng.
00033 Yn ee svia kab kh0ezuie ee sviaf viuu-z0rh0ea, tuix exthaau hitpeeng thoaan+l0qkhix.
00034 Chiux'ar lorng ykefng teq puq sin'vie, ar hoafnci0x ti zuyzhaannih.
```

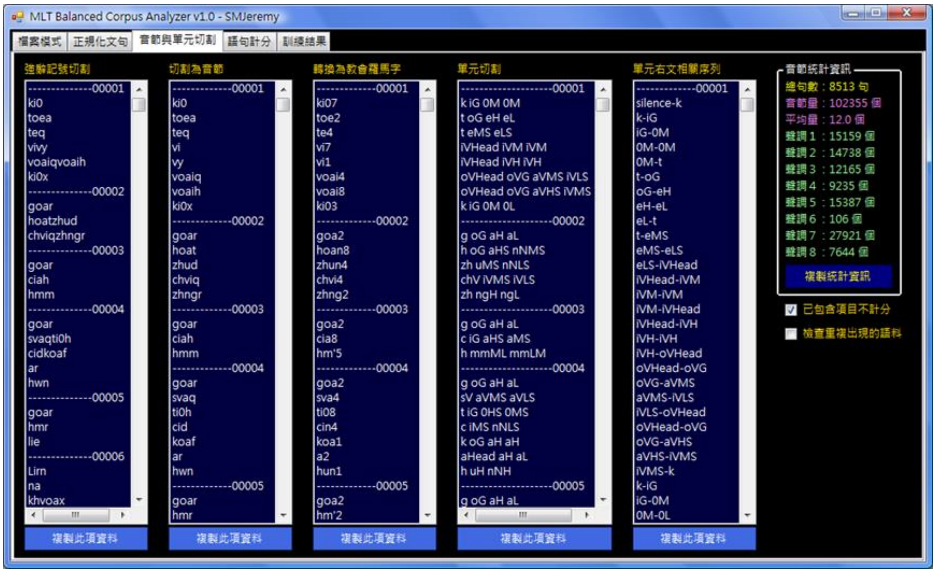
正規化文句資訊
總句數：8513 句
肯定句：7936 句
驚嘆句：178 句
疑問句：399 句
詞彙量：58825 個
平均量：6.9 詞/句

□ 儲存包含統計資訊
□ 儲存時轉換成小寫

文句篩選
語詞數上限 20
語詞數下限 0

Fig. 3. Screenshot of sentence normalization in the balanced sentence training and analysis system.

The syllabic and unit segmentation function is responsible for generating the tonal phonemes of sentences and their RCD sequences. This function also calculates the number of syllables and tones (Fig. 4).



連聯記號切割

```
-------------00001
ki0
toea
teq
vivy
voaiqvoaih
ki0x
-------------00002
goar
hoatzhud
chviqzhngr
-------------00003
goar
ciah
hmm
-------------00004
goar
svaqti0h
cidkoaf
ar
hwn
-------------00005
goar
hmr
lie
-------------00006
Lirn
na
khvoax
```

切割為音節

```
------------00001
ki0
toea
teq
vi
vy
voaiq
voaih
ki0x
------------00002
goar
hoat
zhud
chviq
zhngr
------------00003
goar
ciah
hmm
------------00004
goar
svaq
ti0h
cid
koaf
ar
hwn
------------00005
goar
hmr
```

轉換為教會羅馬字

```
------------00001
ki07
toe2
te4
vi7
vi1
voai4
voai8
ki03
------------00002
goa2
hoan8
zhun4
chvi4
zhng2
------------00003
goa2
cia8
hm'5
------------00004
goa2
sva4
ti08
cin4
koa1
a2
hun1
------------00005
goa2
hm'2
```

單元切割

```
------------------00001
k iG 0M 0M
t oG eH eL
t eMS eLS
iVHead iVM iVM
iVHead iVH iVH
oVHead oVG aVMS iVLS
oVHead oVG aVHS iVMS
k iG 0M 0L
------------------00002
g oG aH aL
h oG aHS nNMS
zh uMS nNLS
chV iVMS iVLS
zh ngH ngL
------------------00003
g oG aH aL
c iG aHS aMS
h mmML mmLM
------------------00004
g oG aH aL
sV aVMS aVLS
t iG 0HS 0MS
c iMS nNLS
k oG aH aH
aHead aH aL
h uH nNH
------------------00005
g oG aH aL
```

單元右文相關序列

```
------------00001
silence-k
k-iG
iG-0M
0M-0M
0M-t
t-oG
oG-eH
eH-eL
eL-t
t-eMS
eMS-eLS
eLS-iVHead
iVHead-iVM
iVM-iVM
iVM-iVHead
iVHead-iVH
iVH-iVH
iVH-oVHead
oVHead-oVG
oVG-aVMS
aVMS-iVLS
iVLS-oVHead
oVHead-oVG
oVG-aVHS
aVHS-iVMS
iVMS-k
k-iG
iG-0M
0M-0L
```

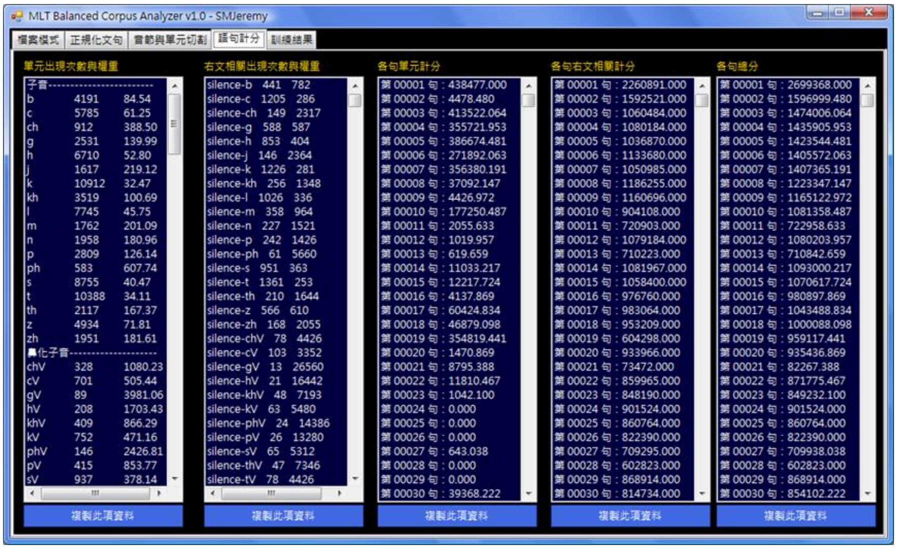
音節統計資訊

總句數：8513 句
音節量：102355 個
平均量：12.0 個
聲調 1：15159 個
聲調 2：14738 個
聲調 3：12165 個
聲調 4：9235 個
聲調 5：15387 個
聲調 6：106 個
聲調 7：27921 個
聲調 8：7644 個

☑ 已包含項目不計分
□ 檢查重複出現的語料

Fig. 4. Screenshot of syllabic and unit segmentation in the balanced sentence training and analysis system.

Sentence scoring involves presenting the distribution of all phonemes and RCD sequences and then calculating the score of each sentence based on the weighted scores (Fig. 5).



單元出現次數與權重

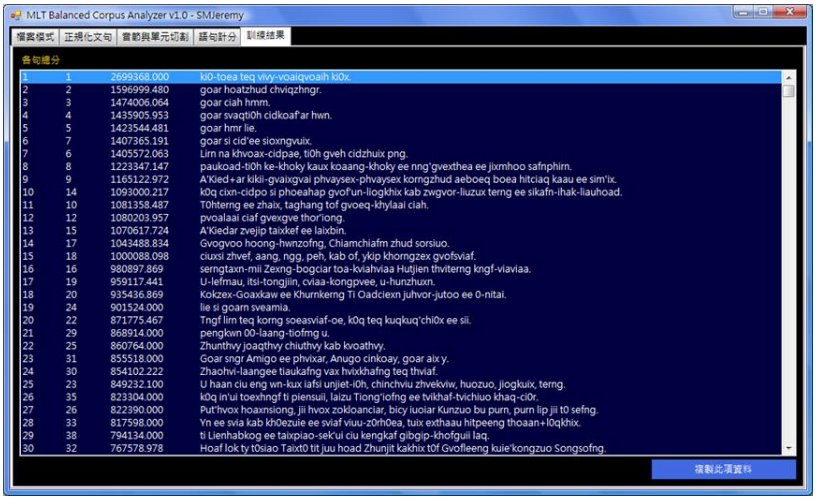
子音-----------------------		
b	4191	84.54
c	5785	61.25
ch	912	388.50
g	2531	139.99
h	6710	52.80
j	1617	219.12
k	10912	32.47
kh	3519	100.69
l	7745	45.75
m	1762	201.09
n	1958	180.96
p	2809	126.14
ph	583	607.74
s	8755	40.47
t	10388	34.11
th	2117	167.37
z	4934	71.81
zh	1951	181.61
鼻化子音-------------------		
chV	328	1080.23
cV	701	505.44
gV	89	3981.06
hV	208	1703.43
khV	409	866.29
kV	752	471.16
phV	146	2426.81
pV	415	853.77
sV	937	378.14

右文相關出現次數與權重

silence-b	441	782
silence-c	1205	286
silence-ch	149	2317
silence-g	588	587
silence-h	853	404
silence-j	146	2364
silence-k	1226	281
silence-kh	256	1348
silence-l	1026	336
silence-m	358	964
silence-n	227	1521
silence-p	242	1426
silence-ph	61	5660
silence-s	951	363
silence-t	1361	253
silence-th	210	1644
silence-z	566	610
silence-zh	168	2055
silence-chV	78	4426
silence-cV	103	3352
silence-gV	13	26560
silence-hV	21	16442
silence-khV	48	7193
silence-kV	63	5480
silence-phV	24	14386
silence-pV	26	13280
silence-sV	65	5312
silence-thV	47	7346
silence-tV	78	4426

各句單元計分	各句右文相關計分	各句總分
第 00001 句：438477.000	第 00001 句：2260891.000	第 00001 句：2699368.000
第 00002 句：4478.480	第 00002 句：1592521.000	第 00002 句：1596999.480
第 00003 句：413522.064	第 00003 句：1060484.000	第 00003 句：1474006.064
第 00004 句：355721.953	第 00004 句：1080184.000	第 00004 句：1435905.953
第 00005 句：386674.481	第 00005 句：1036870.000	第 00005 句：1423544.481
第 00006 句：271892.063	第 00006 句：1133680.000	第 00006 句：1405572.063
第 00007 句：356380.191	第 00007 句：1050985.000	第 00007 句：1407365.191
第 00008 句：37092.147	第 00008 句：1186255.000	第 00008 句：1223347.147
第 00009 句：4426.972	第 00009 句：1160696.000	第 00009 句：1165122.972
第 00010 句：177250.487	第 00010 句：904108.000	第 00010 句：1081358.487
第 00011 句：2055.633	第 00011 句：720903.000	第 00011 句：722958.633
第 00012 句：1019.957	第 00012 句：1079184.000	第 00012 句：1080203.957
第 00013 句：619.659	第 00013 句：710223.000	第 00013 句：710842.659
第 00014 句：11033.217	第 00014 句：1081967.000	第 00014 句：1093000.217
第 00015 句：12217.724	第 00015 句：1058400.000	第 00015 句：1070617.724
第 00016 句：4137.869	第 00016 句：976760.000	第 00016 句：980897.869
第 00017 句：60424.834	第 00017 句：983064.000	第 00017 句：1043488.834
第 00018 句：46879.098	第 00018 句：953209.000	第 00018 句：1000088.098
第 00019 句：354819.441	第 00019 句：604298.000	第 00019 句：959117.441
第 00020 句：1470.869	第 00020 句：933966.000	第 00020 句：935436.869
第 00021 句：8795.388	第 00021 句：73472.000	第 00021 句：82267.388
第 00022 句：11810.467	第 00022 句：859965.000	第 00022 句：871775.467
第 00023 句：1042.100	第 00023 句：848190.000	第 00023 句：849232.100
第 00024 句：0.000	第 00024 句：901524.000	第 00024 句：901524.000
第 00025 句：0.000	第 00025 句：860764.000	第 00025 句：860764.000
第 00026 句：0.000	第 00026 句：822390.000	第 00026 句：822390.000
第 00027 句：643.038	第 00027 句：709295.000	第 00027 句：709938.038
第 00028 句：0.000	第 00028 句：602823.000	第 00028 句：602823.000
第 00029 句：0.000	第 00029 句：868914.000	第 00029 句：868914.000
第 00030 句：39368.222	第 00030 句：814734.000	第 00030 句：854102.222

Fig. 5. Screenshot of sentence scoring in the balanced sentence training and analysis system.



```
1    1    2699368.000   ki0-toea teq vivy-voaiqvoaih ki0x.
2    2    1596999.480   goar hoatzhud chviqzhngr.
3    3    1474006.064   goar ciah hmm.
4    4    1435905.953   goar svaqti0h cidkoaf'ar hwn.
5    5    1423544.481   goar hmr lie.
6    7    1407365.191   goar si cid'ee sioxngvuix.
7    6    1405572.063   Lirn na khvoax-cidpae, ti0h gveh cidzhuix png.
8    8    1223347.147   paukoad-ti0h ke-khoky kaux koaang-khoky ee nng'gvexthea ee jixmhoo safnphirn.
9    9    1165122.972   A'Kied+ar kikii-gvaixgvai phvaysex-phvaysex korngzhud aeboeq boea hitciaq kaau ee sim'ix.
10   14   1093000.217   k0q cixn-cidpo si phoeahap gvof'un-liogkhix kab zwgvor-liuzux terng ee sikafn-ihak-liauhoad.
11   10   1081358.487   T0hterng ee zhaix, taghang tof gvoeq-khylaai ciah.
12   12   1080203.957   pvoalaai ciaf gvexgve thor'iong.
13   15   1070617.724   A'Kiedar zvejip taixkef ee laixbin.
14   17   1043488.834   Gvogvoo hoong-hwnzofng, Chiamchiafm zhud sorsiuo.
15   18   1000088.098   ciuxsi zhvef, aang, ngg, peh, kab of, ykip khorngzex gvofsviaf.
16   16   980897.869    serngtaxn-mii Zexng-bogciar toa-kviahviaa Hutjien thviterng kngf-viaviaa.
17   19   959117.441    U-lefmau, itsi-tongjiin, cviaa-kongpvee, u-hunzhuxn.
18   20   935436.869    Kokzex-Goaxkaw ee Khurnkerng Ti Oadciexn juhvor-jutoo ee 0-nitai.
19   24   901524.000    lie si goarn sveamia.
20   22   871775.467    Tngf lirn teq korng soeasviaf-oe, k0q teq kuqkuq'chi0x ee sii.
21   29   868914.000    pengkwn 00-laang-tiofmg u.
22   25   860764.000    Zhunthvy joaqthvy chiuthvy kab kvoathvy.
23   31   855518.000    Goar sngr Amigo ee phvixar, Anugo cinkoay, goar aix y.
24   30   854102.222    Zhaohvi-laangee tiaukafng vax hvixkhafng teq thviaf.
25   23   849232.100    U haan ciu eng wn-kux iafsi unjiet-i0h, chinchviu zhvekviw, huozuo, jiogkuix, terng.
26   35   823304.000    k0q in'ui toexhngf ti piensuii, laizu Tiong'iofng ee tvikhaf-tvichiuo khaq-ci0r.
27   26   822390.000    Put'hvox hoaxnsiong, jii hvox zokloanciar, bicy iuoiar Kunzuo bu purn, purn lip jii t0 sefng.
28   33   817598.000    Yn ee svia kab kh0ezuie ee sviaf viuu-z0rh0ea, tuix exthaau hitpeeng thoaan+l0qkhix.
29   38   794134.000    ti Lienhabkog ee taixpiao-sek'ui ciu kengkaf gibgip-khofguii laq.
30   32   767578.978    Hoaf lok ty t0siao Taixt0 tit juu hoad Zhunjit kakhix t0f Gvofleeng kuie'kongzuo Songsofng.
```

Fig. 6. Screenshot of the training results in the balanced sentence training and analysis system.

Fig. 6 presents the training results produced by the system, namely the balanced sentence selection results determined according to the scoring criteria.

After the spell-checking and screening of the 8,905 sentences in the corpus, 392 sentences were removed due to misspellings or redundancy. This resulted in 8,513 sentences for training, comprising 102,355 syllables. The mean number of syllables was 12.0 per sentence. After the unit and RCD scores of each sentence were calculated, the total sentence score was used to rank the speech by volume of information. Sentences were input to the training algorithm in order from greatest to least volume of information contained in each sentence for training, which yielded 869 balanced sentences, with 12,544 syllables in total and an average of 14.4 syllables per sentence (Table 4).

Table 4. Trained 869 balanced sentences.

Total number of balanced sentences: 869		
Number of words: 7,140 Number of syllables: 12,544		Mean number of words: 8.2 (per sentence) Mean number of syllables: 14.4 (per sentenc
Type		
Affirmative sentences: 810	Exclamations: 14	Questions: 45
Tone		Numbers of syllables
Tone 1		1,843
Tone 2		1,735
Tone 3		1,510
Tone 4		1,246
Tone 5		1,858
Tone 7		3,360
Tone 8		992

After training, tonal phonemes appeared an average of 261 times among the 869 balanced sentences. To supplement the corpus, some sentences were drafted such that they contained 21 tonal phonemes that appeared fewer than 20 times. Furthermore, after training, the collected corpus only covered approximately 65% of the 4,875 possible phoneme sequences. In future research, the remaining possible sequences should be entered into the balanced sentence corpus through manual sentence drafting to increase the coverage of phoneme sequences in corpus, thereby increasing the volume of speech information and thus providing the training system with a greater abundance of context cluster-dependent information.

(C) Drafting of sentences containing rare phonemes

Sentences were manually drafted such that they contained phoneme units that appeared fewer than 21 times for the purpose of supplementing the target units. Syllables were selected to cover these 21 units and incorporated into sentences based on whether units had an insufficient number of appearances. A total of 218 supplemental sentences were drafted.

(D) Distribution of tonal phonemes and RCD sequences

Each unit had 3,215 possible RCD sequences for tonal phonemes, approximately 66% of which were covered in the corpus. The sequences not covered must be supplemented to ensure that the principles of all RCD sequences of tonal phonemes are covered.

6.3 Validation Results for the Robust Tonal Phonemes

After the speech recognition tool had identified the phoneme sequences of each sentence, statistical methods were used to validate the reliability of the tonal phonemes.

(A) Tonal phoneme sequence recognition results

The initial recognition results for the acoustic model trained using the HTK speech recognition tool are presented in Table 5. The overall phoneme recognition rate was 96.47%.

Table 5. HTK tool's recognition of tonal phonemes.

Consonants									
Units	b	c	ch	g	h	i	k	kh	l
Appearances	486	670	160	530	826	202	1,174	408	812
Errors	25	10	2	52	80	25	75	17	34
Recognition rate	94.9%	98.5%	98.8%	90.2%	90.3%	87.6%	93.6%	95.8%	95.8%
Units	m	n	p	ph	s	t	th	z	zh
Appearances	260	276	575	131	898	1,136	479	773	281
Errors	1	5	10	1	33	76	7	15	7
Recognition rate	99.6%	98.2%	98.3%	99.2%	96.3%	93.3%	98.5%	98.1%	97.5%

(B) Sensitivity and specificity of the recognition results

The recognition results for phonemes demonstrated considerably high levels of sensitivity and specificity. Phonemes units with sensitivity less than 0.95 were consonants *b*, *g*, *h*, *j*, *k*, and *t*; the beginning monophthong *oHead*; monophthongs *aH*, *aM*, *Em*, *iH*, *iM*, *iL*, and *uH*; and the nasal monophthong *aVM*. The units with sensitivity lower than 0.9 were *j*, *aM*, and *aVM*.

The specificity for all phoneme units exceeded 0.99. The sensitivity results demonstrated the considerably high intelligibility of the tonal phoneme unit set, and the specificity results indicated that the speech recognition tool had a considerably low probability of incorrect recognition. As such, the recognition rate is robust.

(C) Analysis of tonal phoneme robustness using ROC curves

The sensitivity value and value (1 – specificity) of the tonal phonemes were depicted as ROC curves. The area under the ROC curve (AUC) is the indicator of tonal phoneme robustness. The mean AUC among all groups was approximately 0.990; nasal consonants, beginning nasal monophthongs, nasal glides, nasal monophthongs, nasal syllable codas had ROC curves of more favorable shapes. Therefore, these groups had AUCs of 0.99 or higher.

The overall tonal phonemes had a mean AUC of 0.988, indicating competence in describing the speech characteristics of Taiwanese and considerably robust intelligibility. Therefore, they can be used in the HMM-based Taiwanese TTS system developed in this study.

6.4 Development of the Taiwanese TTS System

HTS was the platform used for building the HTS training modules and designing the speech synthesis user interface. These elements were then combined to develop an HMM-based Taiwanese TTS system.

(A) Building the training module

Ubuntu Linux version 8.10 was chosen as the environment for constructing the HTS training module due to its high stability, high expandability, simple interface, and free open-source code.

The training corpus comprised MLT audio files and corresponding text, which consisted 1,087 sentences. The audio files were approximately 100 min in total duration and had a size of 171 MB. The raw audio files were in .wav format; they were converted into .raw files for training. After the corresponding text files were analyzed by the text analysis and processing module in the training phase, two file types were generated, one with single phoneme tags and the other with context clustering-dependent sequence tags, for HMM training. Furthermore, a context clustering-dependent problem set were constructed using contextual information on phonemes, syllables, words, phrases, and sentences. The file size of the completed problem set was 124 KB.

Training the HMMs of each tonal phoneme required several hours. The F_0 modeling, spectrum modeling, and duration modeling files were generated for each phoneme after training and saved in the HTS training folder, which had a total size of approximately 2.27MB. F_0, spectrum, and duration decision trees were also generated, with a total file size of approximately 988 KB.

(B) Establishment of a synthesis system

The tagged files of the sentences to be synthesized into speech were placed in the HTS system's tag file folder, and the speech synthesis command was executed. The HTS system's synthesis function involves generating .wav and .raw speech files and saving them in a folder. Under considerations of universal compatibility and convenience of use, the synthesis subsystem (which originally ran on a Linux command line interface) was redesigned to run on the Microsoft Windows operating system, which has a larger user base. A window application program was written using API developed based on the HTS, which addressed the inability of the HTS system to play the synthesized speech instantly.

A graphical user interface was designed for the proposed program through appropriate modifications and program compilations. The training files required by the system were taken from the model parameter files and decision tree system files originally trained in the Ubuntu Linux system, which facilitated convenient user operations in human–computer interactions.

The system's synthesis function was developed in the Windows operating system. The HTS API was written using Visual C^{++} in Microsoft Visual Studio. An HTS speech synthesis engine was produced after appropriate modifications and compilations. The synthesis engine was used as the core program of the system's synthesis function, and the core engine and trained files were used to design a window GUI with the Visual C# .NET language. Users enter the sentence to be synthesized in the MLT text input area, and the system automatically generates a tagged text file and synthesized speech file before playing the synthesized speech. The system also generates a .wav synthesized audio file, which is saved in the program folder; users can extract the synthesized audio file for future use.

Conventional corpus-based speech synthesis systems require the support of a massive speech database for synthesis, which typically results in the system taking up several megabytes or even gigabytes of space and long computing times. Furthermore, the connections between the synthesized speech syllables are often broken. By contrast, the proposed Taiwanese TTS system – including the user interface and the parameter training files for synthesizing purposes – requires no more than 4 MB of storage space to synthesize Taiwanese speech that is stable, continuous, and smooth. The system can be applied as an embedded system to address weaknesses of low portability and discontinuous synthesized speech.

6.5 Speech Quality Assessment Results

The quality of synthesized speech was assessed according to the MOS. Thirty testees were asked to subjectively assess the synthesized speech on a scale of 1 to 5, and the median of their scores was taken to measure quality. The assessment results related to the naturalness of the synthesized speech for the Inside corpus are presented in Table 6.

Table 6. Speech assessment for the Inside corpus.

No.	Sentences in the Inside corpus
01	Y ti simlai ciu teq liam ji-safm-gvor-pad-sux-kiuo-liok. He was reciting "236-8496" in his head.
02	Hoe'mm na khuy, hiaf ciu k0q boeq u hoef. If flowers bloom, there will be flowers again.
03	Thafng kab lie z0rhoea laai hib cidtviw siong. I can take a photo with you.
04	Tvy ee ymliau kab zhamthngg ee miqkvia, lorng maix. I don't want sweet soft drinks or anything with sugar.
05	K0q si iwkoafn svesie-hengboong ee buxntoee. It is another live-or-die matter.

06	Paukoad Laam-Pag-Kosog-Konglo, Thohngg Kokzex Kitviuu. It includes the North–South Freeway and Taoyuan International Airport.
07	Goar thak vuix piauzurn. I read the standard “vuix.”
08	Cidkoaf chvy ti thviterng teq siafmsiarm hoatkngf. Some stars are twinkling in the sky.
09	Zhuolai ma boo symmih'tadcvii ee miqkvia thafng theqkhix tngx. There is nothing valuable in the house that can be pawned.
10	A'Efng huisioong thiarm, ciu phag ti zhngpvy khuxn+khix.. Ah-Ying was very tired and fell asleep by the bed.
11	Cvii si papaf thaxnlaai ee. The money was earned by Dad.
12	Goar si cid'ee sioxngvuix. I am a military captain.
13	Kiexnpoftvoaf ee hoxbea kab lirn-taw ee tiexn'oe-hoxbea. The health insurance policy number and your home phone number.
14	Cit'ee sizun, u thviakvix cid'ee soeasoea'ar ee sviaf Phuqphuq'phuh. At this time, a very soft “Pu, pu, pu” sound could be heard.
15	Zoeakin cidtviuu kafmmo-au ciu sengkhw pviecviaa boo-zuxiuu. Since a recent bout of common cold, my body has not been moving freely.

The 30 testees, all of whom were adults fluent in Taiwanese, were trained on the scoring standards and then asked to assess the naturalness of each sentence from the Inside and Outside corpuses on a scale of 1 to 5 for a MOS. The medians of their assessments were then taken as the measure of naturalness. The sentences from the Inside and Outside corpuses both scored 4 points; the overall system scored 4 points.

7. CONCLUSION AND OUTLOOKS

A Taiwanese speech synthesis system was developed. The development process involved an in-depth review and analysis of MLT subsyllables according to the phonetic changes and characteristics produced by the seven tones specific to Taiwanese and the construction of a Taiwanese tonal phoneme model; the model was then used to construct a Taiwanese balanced sentence training algorithm. A balanced sentence training and analysis system was then developed using window programming, which enabled the rapid and reliable execution of complex processes related to the balanced sentence training algorithm. After the required audio files were recorded, a Taiwanese speech synthesis system with uses in clinical speech technology and computational linguistics was constructed. The synthesized speech had an MOS of 4. This study provides a foundation for research and development in clinical speech technology and computational linguistics; the proposed system has potential applications in medical services, educational training, and multimedia speech-assisted rehabilitation.

REFERENCES

1. J. C. Chang, “Influence of a grade 4 Hakka language immersion program on school children’s Hakka listening and speaking competence,” *Global Hakka Study*, Vol. 10,

2018, pp. 59-90.

2. K. P. J. Tse, *An Introduction to Linguistics*, 3rd ed., San Min Book Co., Ltd., Taipei, 2011.
3. D. Crystal, *Language and the Internet*, 2nd ed., Cambridge University Press, Cambridge, 2006, pp. 1-7.
4. M. Krauss, "The world's languages in crisis," *Language*, Vol. 68, 1992, pp. 4-10.
5. UNESCO, "UNESCO atlas of the world's languages in danger," http://www.unesco.org/languages-atlas/en/atlasmap.html, 2021.
6. Wikipedia, "List of languages by number of native speakers," https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers, 2021.
7. J. DeBernardi, "Linguistic nationalism: The case of southern min," Sino-Platonic Papers, Department of East Asian Languages and Civilizations, University of Pennsylvania, Philadelphia, USA, 1991, pp. 1-25.
8. S. F. Huang, "Language, society and ethnicity: a study of the sociology of language in Taiwan," *Crane*, Taipei, Taiwan, 1993.
9. C. T. Hsu, "The direction of Taiwanese textualization," *Independence Evening Post Press*, Taipei, 1992, pp. 3-55.
10. K. H. Yap, "Family language choice in Taiwan," *Taiwanese Journal of Sociology*, Vol. 62, 2017, pp. 59-111.
11. H. K. Tiun, "Mother tongues first: planning Taiwanese native languages education policy for language revitalization," *Journal of Research in Education Sciences*, Vol. 65, 2020, pp. 175-200.
12. J. Fishman, "Reversing language shift: theory and practice of assistance to threatened languages," *Language in Society*, Vol. 23, 1994, pp. 115-119.
13. J. M. Huang, "A study of organization reinventing and policy design on revitalization of native language: A case of Taiwanese," *The Journal of Chinese Public Administration*, Vol. 11, 2012, pp. 233-248.
14. Development of National Language Act, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=H0170143, 2019.
15. U. Ang, "The motivation and direction of sound change: on the competition of Minnan dialects Chang-chou and Chüan-chou, and the emergence of general Taiwanese," Doctoral dissertation, Institute of Linguistics, National Tsing Hua University, 2003, http://hdl.handle.net/11296/2u847v.
16. U. Ang, "The distribution and regionalization of varieties in Taiwan," *Language Lingustics*, Vol. 14, 2013, pp. 315-369.
17. S. J. Lin, "Myna: A development tool for Mandarin-Taiwanese machine translation/text-to-speech system," Master Thesis, Department of Computer Science and Information Engineering, National Cheng Kung University, 2000, https://hdl.handle.net/11296/9j6a6y.
18. J. W. Lin, "A study on the linguistic style of the Tale of Lychee Mirror," Master Thesis, Department of Chinese Literature, National Chengchi University, https://hdl.handle.net/11296/fxdnpx, 2017.
19. Y. J. Sher, K. C. Chung, and C. H. Wu, "Design and develop Taiwanese syllable-based synthesis units' database," *Journal of Medical and Biological Engineering*, Vol. 19, 1999, pp. 47-58.

20. K. Liim, *Textbook of Modern Literal Taiwanese*, 1st ed., Dasia Publishing Ltd., Taipei, Taiwan, 1990.
21. K. Liim, "Website for 21st century Taiwanese language and art web," EduTech Foundation, http://www.edutech.org.tw, 2012.
22. D. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, Vol. 67, 1980, pp. 971-995.
23. F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proceedings of Eurospeech*, Vol. 89, 1989, pp. 13-19.
24. Y. J. Chen, C. H. Wu, Y. H. Chiu, and H. C. Liao, "Generation of robust phonetic set and decision tree for Mandarin using chi-square testing," *Speech Communication*, Vol. 38, 2002, pp. 349-364.
25. K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of IEEE Workshop on Speech Synthesis*, 2002, pp. 227-230.
26. S. H. Chen, "A statistical-based pitch contour model for Mandarin speech," *Journal of the Acoustical Society of America*, Vol. 117, 2005, pp. 908-925.
27. S. H. Chen, S. Chang, and S. M. Lee, "A statistical model based fundamental frequency synthesizer for Mandarin speech," *Journal of the Acoustical Society of America*, Vol. 92, 1992, pp. 114-120.
28. S. H. Hwang and S. H. Chen, "A neural network synthesizer of pause duration for Mandarin test-to-speech," *Electronics Letters*, Vol. 28, 1992, pp. 720-721.
29. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. Woodland, and C. Zhang, *The HTK Book*, Cambridge University, Cambridge, 2015.
30. C. H. Huang, L. C. Huang, and C. C. Kuo, "Text processing of Taiwanese POJ for text-to-speech," *Computer and Communications Consumer Electronics Technology Journal*, Vol. 133, 2010, pp. 95-102.
31. C. J. Lin and H. H. Chen, "A Mandarin to Taiwanese Min Nan machine translation system with speech synthesis of Taiwanese Min Nan," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 4, 1999, pp. 59-84.
32. M. S. Liang , R. C. Yang, Y. C. Chiang, D. C. Lyu, and R. Y. Lyu, "A Taiwanese text-to-speech system with applications to language learning," in *Proceedings of IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 91-95.
33. S. H. Chen *et al*., "Modeling of speaking rate influences on Mandarin speech prosody and its application to speaking rate-controlled TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, 2014, pp. 1158-1171.
34. C. Huang, Y. Shi, J. Zhou, M. Chu, T. Wang, and E. Chang, "Segmental tonal modeling for phone set design in mandarin LVCSR," *IEEE Transactions on Acoustics Speech and Signal*, 2004, pp. 901-904.
35. K. Liim, L. Liim, K. Ciofng, E. Liie, P. Liim, and E. Siaa, *Medicinal Textbook in Modern Literal Taiwanese*, 1st ed., New Wun Ching Developmental Publishing Co., Ltd., Taipei, 2005, pp. 1-62.
36. HTS Working Group, "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp, 2009.

37. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis," in *Proceedings of EUROSPEECH*, Vol. 5, 1999, pp. 2347-2350.
38. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, 2000, pp. 1315-1318.
39. Y. C. Yang, "An implementation of Taiwanese text-to-speech system," Master Thesis, Department of Communication Engineering, National Yang Ming Chiao Tung University, 1998.

Yung-Ji Sher received the BS degree in physical therapy from National Yang-Ming Medical College, Taipei, Taiwan, in 1992, and the MS and Ph.D. degrees in Biomedical Engineering from National Cheng Kung University, Tainan, Taiwan, in 1997 and 2006, respectively. He is currently an Associate Professor at the Department of Special Education and the Graduate Institute of Rehabilitation Counseling at National Taiwan Normal University. His research interests include Modern Literal Taiwanese, assistive technology, adaptive physical education, natural language processing, computational linguistics, and special education.

Ming-Chun Hsu received the BS degree from National Chiayi University, Taiwan, in 2007, and the MS degree from National Cheng Kung University, Taiwan, in 2009. He is currently the Supervisor of Software Development Department of Wistron NeWeb Corporation (WNC) and also serves as the security software analyst. His research interests include software architecture design/analysis, telematics, and cybersecurity engineering.

Yu-Hsien Chiu received the BS degree from I-Shou University, Taiwan, in 1997, and the MS and Ph.D. degrees from National Cheng Kung University, Taiwan, in 1999 and 2003, respectively. He is currently the Consultant of Digital Biomedical Research Center, National Cheng Kung University, and the Convener of Digital Health Research and Industry Strategic Alliance, Taiwan. His research interests include artificial intelligence, digital health, medical instrumentation and devices, rehabilitation engineering and assistive technology.

Yeou-Jiunn Chen received his BS degree in Mathematics from Tatung Institute of Technology, Taipei, Taiwan, and his Ph.D. degree from the Institute of Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 1995 and 2000, respectively. He was with the Advanced Technology Center, Computer and Communications Laboratories, Industrial Technology Research Institute, from 2001 to 2005 as a Researcher. He is currently a Professor at the Department of Electrical Engineering, Southern Taiwan University of Science and Technology, Tainan, Taiwan. His research interests include biomedical signal processing, spoken language processing, and artificial intelligence. Dr. Chen is a member of the Biomedical Engineering Society, Taiwan Rehabilitation Engineering and Assistive Technology Society, and the Association for Computational Linguistics and Chinese Language Processing.

Chung-Hsien Wu received the BS degree in Electronics Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the MS and Ph.D. degrees in Electrical Engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively. Since 1991, he has been a member of the Department of Computer Science and Information Engineering at NCKU, where he was appointed as Chair Professor in 2017. In the summer of 2003, he worked as a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory of the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He served as an Associate Editor of several journals, including IEEE Transactions on Audio, Speech and Language Processing (2010–2014), IEEE Transactions on Affective Computing (2010-2014), ACM Transactions on Asian and Low-Resource Language Information Processing, and APSIPA Transactions on Signal and Information Processing (2014~2020). He was also a member of the APSIPA BoG from 2019 to 2021. He received the 2018 APSIPA Sadaoki Furui Prize Paper Award and the Outstanding Research Award from the Ministry of Science and Technology in Taiwan in 2010 and 2016. His research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing.

Jiunn-Liang Wu received his M.D. degree from China Medical University in Taichung, Taiwan, in 1989. He is currently an Associate Professor of Otolaryngology at the Department of Medicine, National Cheng Kung University, Tainan, Taiwan, and an Otolaryngologist at the National Cheng Kung University Hospital, Tainan, Taiwan. His research interests include otology, hearing balance rehabilitation, hearing aids, cochlear implant, and children’s language development.