Research Article

Xiaofeng Liu, Pradeep Kumar Singh*, and Pljonkin Anton Pavlovich

Accent labeling algorithm based on morphological rules and machine learning in English conversion system

https://doi.org/10.1515/jisys-2020-0144
received December 29, 2020; accepted March 30, 2021

Abstract: The dependency of a speech recognition system on the accent of a user leads to the variation in its performance, as the people from different backgrounds have different accents. Accent labeling and conversion have been reported as a prospective solution for the challenges faced in language learning and various other voice-based advents. In the English TTS system, the accent labeling of unregistered words is another very important link besides the phonetic conversion. Since the importance of the primary stress is much greater than that of the secondary stress, and the primary stress is easier to call than the secondary stress, the labeling of the primary stress is separated from the secondary stress. In this work, the labeling of primary accents uses a labeling algorithm that combines morphological rules and machine learning; the labeling of secondary accents is done entirely through machine learning algorithms. After 10 rounds of cross-validation, the average tagging accuracy rate of primary stress was 94%, the average tagging accuracy rate of secondary stress was 94%, and the total tagging accuracy rate was 83.6%. This perceptual study separates the labeling of primary and secondary accents providing the promising outcomes.

Keywords: text-to-speech conversion, unregistered words, stress annotation, accent labeling, machine learning

1 Introduction

In the English TTS system, whether the accent marking of English words is correct is directly related to the quality of speech synthesis. In the analysis of stress, people usually use a three-layer processing method, that is, stress is divided into primary stress (Primary Stress), secondary stress (Secondary Stress), and unstressed stress (Unstressed Stress). Because stress is realized through syllables, if it is distinguished by whether it is stressed or not, the syllables of English words are divided into three levels accordingly: the main stressed syllable, the secondary stressed syllable, and the unstressed syllable. For example, the word administration [d_,mi_nis_‘trei_n] has five syllables, among which the syllable treii is the main stressed syllable, mii is the secondary stressed syllable, and the remaining three are unstressed syllables.

The stress situation in English is very complicated and takes the lexicon used in this system as an example. Of all 27,300 words, the word with the most syllables has eight syllables (such as interoperability [in_tr_,_pr_’bi_ li_ti]), but the main stress may appear in six different positions. The situation of secondary

* **Corresponding author: Pradeep Kumar Singh,** Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, UP, India, e-mail: pradeep_84cs@yahoo.com
Xiaofeng Liu: Department of Aircraft Maintenance, Sichuan Southwest Vocational College of Civil Aviation, Chengdu 610000, Sichuan Province, China, e-mail: xiaofengliu204@gmail.com
Pljonkin Anton Pavlovich: Institute of Computer Technologies and Information Security, Southern Federal University, Russia, e-mail: pljonkin@mail.ru

stress is more complicated. Some multisyllable words do not have secondary stress, but some may have more than one secondary stress, such as arteriosclerosis [a:_,ti_ri_u_,skli_′ru_sis], which has 2 Secondary stress [1].

The stress situation is so complicated, how to mark English words with accent? The easiest thing to think of is to mark the accent through the Morphological Structure of the word. For example, the main stress of nouns ending in ion is generally located at the penultimate syllable (Penult), such as acceleration [æk_,se_l_‘rei_n]. However, not every word's stress can be marked by morphological structure, such as adulterous [_′dl_t_rs] and numerical [′nju:_m_rs] have the ending ous, but the position of the main stress It cannot be judged by this ending. What's more, there are many words with a single morphological structure, such as adamantine [,æ_d_′mæn_tain]. Therefore, only relying on morphological structure cannot complete the task of accent annotation [2].

Another method of accent labeling is to perform accent labeling through machine learning. For example, J. Zhang and Cercone of the University of Regina in Canada use Iterated Version Space Learning to use syllables as learning objects and extract for each syllable. Seven attributes are used for accent annotation learning; John Coleman uses Probabilistic Grammar for accent annotation learning [3]. Machine learning has provided significant development in various fields of image processing, image recognition and speech classification, etc. [4,5]. However, because the accent situation is very complicated, the effect of accent annotation by these machine learning methods is not very satisfactory. According to the author, the accuracy of the former is only 84%, while the latter is even worse, only about 75%.

Based on the above situation, in order to better solve the problem of accent marking, in this system, the following measures are planned to be taken:

(1) ***Separate the primary and secondary accents into labels:*** The importance of primary stress is far greater than secondary stress. In the English TTS system, as long as the main stress of a word is marked correctly, even if the sub-stress markings are deviated, the effect of speech synthesis is acceptable, and *vice versa*. Moreover, the main stress is simpler than the secondary stress. Under normal circumstances, there is only one main stress in two-syllable words or polysyllable words (a few words have two main stresses, and out of 27,300 words, there are only 387 words that have two main accents, and these words can be treated as special cases) [6]. The situation of secondary stress is relatively complicated, and the position of secondary stress often has a great relationship with the position of primary stress. The previous machine learning methods often used the primary and secondary stresses together for annotation learning, which increased the complexity of learning, and the learning effect was not ideal. In this system, the primary stress and secondary stress are separated for learning. The main purpose is to improve the accuracy of primary stress labeling as much as possible.

(2) ***Adopt the main accent annotation algorithm combining morphological rules and machine learning:*** Because the main stress of some words can be determined completely according to their morphological structure, corresponding morphological rules are formulated for the characteristics of these words, and the correct rate of labeling is generally higher than the labeling rules obtained through machine learning [7]. For words that cannot be distinguished by morphological rules, the main stress labeling rules are generated through machine learning [8,9].

While considering the secondary stress, since it cannot be distinguished by the morphological structure, it will be marked completely by machine learning. The vocabulary used by this system for stress annotation learning and testing has 27,300 words, which basically covers some of the most commonly used words in English.

This work contributed in addressing the challenges faced in language learning and various other voice-based advent by proposing a machine learning-based accent labeling method. The research work mainly focuses on the labeling of primary accents and separating the primary stress from the secondary stress. For this purpose, this article presented a combination of morphological rules and machine learning algorithm using ten-fold cross-validation process. The perceptual study is effective in separating the labeling of primary and secondary accents and provides the promising outcomes in terms of average tagging accuracy rate for both primary and secondary stresses. This work provides a feasible solution for accent labeling of unregistered words using the concept of machine learning.

The remainder of this article is organized as follows: Section 2 presents the existing literature in the filed of accent labeling and conversion; Section 3 discussed the construction of main accent labeling rules based on morphological structure followed by the formulation of Machine learning-based sub-stress labeling rules in Section 4; Section 4 presents the results and discussion part; and Section 5 provides the concluding remarks for this work including the future research directions.

2 Literature review

Cruz-Benito, J. et al. used machine learning algorithms (trained using data from more than 3000 users previously involved in the questionnaire survey) to build prediction models, extract the most relevant factors describing the model, and cluster users according to the similar characteristics of users and these factors. Based on these groups and their performance in the system, the researchers generated heuristic rules between different versions of web forms to guide users to provide them with the most adequate versions (modifying the user interface and user experience). The results are as follows [10]: Fan et al. proposed a quantitative association rule mining (QARM) method based on machine learning. The method consists of sliding window partition (SWP) and random forest (RF), which associate KPI with KQI. Discrete continuous attributes into Boolean values using SWP are used to mine its association rules. Early warning intervals and early warning points are obtained by SWP. Use RF feature importance to measure the association between KPI and KQI. The priority of the correlation strength between all KPI and each KQI can be obtained by RF. The warning point and the associated intensity priority are selected as the optimal solution. The actual data of telecom operators are tested and the results confirm the feasibility and accuracy of the method [11]. Tania et al. investigated an intelligent image-based system for automated paper-based colorimetric testing in real time, providing a proof-of-concept for dry chemical or microfluidic, stable and semiquantitative analysis using larger datasets with different conditions. The proposed image processing framework is based on color channel separation, global thresholding, morphological manipulation, and target detection. A server-based convolutional neural network framework is deployed on mobile platform for image classification using induced transfer learning on mobile platform [12].

The two accents were experimented by Chen et al. [13], and the results obtained for 12 accents provide overall accuracy of 57.12 and 13.32% using the machine learning Support Vector Machine (SVM) and Linear Predictive Coding (LPC) algorithms, respectively. Three different foreign accents were tested by Watanaprakornkul et al. [14] using the SVM and Gaussian Mixture Model (GMM) approaches. The feature extraction method along with machine learning-based classification yields 41.18 and 37.5% accuracy values for SVM and GMM methods, respectively. The hidden markov models with directed acyclic graphs and SVM mechanism were utilized by Tang and Ghorbani [15], which suggests that the SVM-based methodology is effective in classifying the different accents. Novich and Trevino [16] presented a study on accent classification using the concept of neural networks and the extracted features classify the vowels of speech utterances. Later, Behravan et al. [17] presented an approach using the I-vector to separate the eigen-voice from the eigen-channel and then this is compared with the GMM model recognizer. The accent of speech as well as speech gender was identified using the GMM model by Chen et al. [18]. The speaker understanding of accent for a singing language identification is presented in ref. [19] which affects the singing melody transcription accuracy [20,21]. The accent detection had been performed in ref. [22] using the combination of two parallel deep neural networks (DNN and RNN). They have characterized the difference between the prosodic and articulation characteristics. A unique approach has been presented using the acoustic feature extraction and machine learning algorithms [23]. The different combinations of feature descriptors have been utilized for this study to classify the aspects of five different national and international languages. A feed-forward neural network has been utilized for the conversion of weight matrix between the target and the source accents [24]. This system accepts the Mel Frequency Cepstral Coefficient (MFCC) features in one accent and outputs these features in another accent using the activation function outputs of the neural network. The photonic similarity feature was used to propose a frame matching-based method to output the maximum likelihood response of the system [25]. This method was useful in doing the evaluation based on acoustic quality and speaker identity of the native speaker. Authors in ref. [26] proposed a method based on Deep Bidirectional Long Short-Term Memory-based Recurrent Neural Networks for mapping the acoustic features of target speech. This method is useful in aperiodic conversion into the target speech without using the parallel data. A recurrent neural network (RNN) was trained by the authors in ref. [27] to develop an automated speech recognition engine for the identification of acoustic features of the target person. This approach observed that the outcomes of both the singing voice and the source singing were found similar using this RNN-based approach. The authors used a cycle-consistent adversarial network for the image translation using nonparallel data to extract the speech features [28]. The approach was found superior in finding the inter-gender conversion of the target voice. The voice conversion in ref. [29] was accomplished by using the same cycle-consistent adversarial network in order to process the phoneme PosteriorGrams sequences. The target speech features are evenly extracted and synthesized using this method.

The literature suggests that there is still a possibility of improvement in the machine learning aspect of accent labeling which is being addressed in this research work.

3 The construction of main accent labeling rules based on morphological structure

The main stress of many words in English can be determined with the help of its morphological structure [30–32], which is mainly determined by the suffix (Suffix), and the specific situations are divided into two categories:

(1) ***The morphological structure directly determines the position of the main stress***. There are some words in English. The position of the accent can basically be determined according to the ending of the word. The following rules can be defined for such words: Suffix=SufStr Syll Count>=m=>P=Rn where: Suffix=SufStr indicates that the ending of a word is SufStr, and Syll Count>=m indicates that the number of syllables of the word must be greater than or Equal to m, P=Rn means that the main accent is located at the nth syllable from the bottom (numbering starts from 0). Therefore, this type of rule indicates that if the ending of a word is SufStr, and the number of syllables is greater than or equal to m, then the main stress is located at the last nth syllable. For two-syllable or multisyllable words ending with ion, the main stress is mostly located in the penultimate syllable, such as capitalization [,kæ_pi_t_lai_‘zei_n]. According to statistics, there are 1102 double-syllable or multisyllable words with ending ion in the lexicon, and the main stress of 1099 words is located in the penultimate syllable. For such words, the following rules can be formulated to determine the main stress:

    Suffix=ionSyll Count>=2=>P=R1

    There are 71 suffixes with this function, such as ical, ian, ity, ial, eous, ious, ament, ator, etc., for which a total of 71 morphological rules have been defined in order to distinguish this group from other rules. The morphological rules are called α_rules.

(2) ***The main stress of the derivative word formed by the root and the specific ending is consistent with the main stress of the root word.***

    The proportion of words that can determine the position of the main stress according to the ending of the word in English is limited. Among the 27,300 words, there are only 4035 such words, accounting for only 14.8%. But in addition, there are some words with specific endings. Although the position of the main stress cannot be determined by the ending, the position of the main stress is consistent with the main stress of the root word. The most typical of this kind of situation is the inflected form of some words, such as the past tense of the verb acclaim and the past participle form acclaimed[_ʼkleimd]. The main stress of acclaim[_ʼkleim] is consistent with the main stress of acclaim[_ʼkleim] in the second syllable. In addition, there are comparative and superlative forms of adjectives and adverbs, adverb forms composed of adjectives + ly, and noun forms composed of verbs + ment.

The following rules can be formulated for this type of situation (for the sake of distinction, this set of morphological rules is called β-rules in this article):

Suffix=SufStr
Existing(Stem)=TRUE
SylCount=m=>P=nSuffix=SufStr
Existing(Stem)=TRUE SylCount>=m=>P=Stem Accent Pos

Among them: Suffix=SufStr indicates that the ending of a word is SufStr, Stem indicates the root word after removing the ending of a word, and the function Existing is used to judge whether the root word exists in the thesaurus; SyllCount=m (or >=m) indicates the number of syllables of a word is m (or greater than or equal to m), P=n indicates that the position of the main stress is located in the nth syllable (numbering from 0), and Stem AccentPos indicates the position of the main stress of the root word. Therefore, this type of rule indicates that if the suffix of a word is SufStr, the root of a word after removing the suffix is still a word and exists in the thesaurus; and the number of syllables is m (or greater than or equal to m), then its main. The stress is located in the nth syllable (or consistent with the main stress of the root word) [33].

For example, for the case where "adjective + ment constitutes a noun form," the following rules are defined:

Suffix=mentStem=Word_4
SylCount=1=>P=0
Suffix=mentStem=Word_4
SylCount>1=>P=StemAccentPos

The first rule targets words such as pavement ['peiv_mnt], whose root word is pave. After searching, pave exists in the lexicon. Since it is a monosyllable word, the main stress of pavement is located on the first syllable. The second rule targets words such as measurement ['me-mnt], whose root is measure, search for the word in the lexicon, and find the phonetic transcription of measure as ['me-], so the main accent position of measurement is consistent with measure, which is located on the first syllable. A total of 117β-rules were formulated to apply machine learning on main accent labeling rules.

For words that cannot be distinguished by their morphological structure, they can be labeled with machine learning methods. Jian na considers primary stress and secondary stress together in his algorithm and extracts attributes for a word's syllable for learning, which increases the complexity of learning, and the learning effect is not ideal. In this system, the main stress and secondary stress are marked separately for learning, and it is assumed that a word has only one main stress. Take a three-syllable word as an example. Its main stress may appear in three positions. If the main stress appears in the first syllable, the word is marked as 0, and when it appears in the second syllable, it is marked as 1, and so on [34,35].

When learning, the attributes will be extracted for the entire word. Since the main stress can appear in multiple positions, for a two-syllable or multisyllable word, we don't know which syllable or which component of the syllable is paired. It is more important to determine the position of the main stress, so to be safe, all syllables of a word are taken into consideration. It is generally believed that syllable (Syllable) is composed of two parts: Onset and Rhyme. The rhyme is composed of the core (Nucleus) and the ending (Coda).

Therefore, if the initial, core and final sounds of all syllables of a word are extracted as attributes. For a two-syllable word balance[ʹbæ_lns], there are 6 attributes that can be extracted, as shown in Table 1.

Each attribute is identified by the first letter of its corresponding English name plus the sequence number of the corresponding syllable in the word. For example, C_0 indicates the end of the first syllable.

Table 1: Extract attributes for each syllable of balance

O_0	N_0	C_0	O_1	N_1	C_1
B	æ	0	1	ə	ns

If a certain attribute does not have corresponding content, such as the first, and the syllable has no ending, then the value of C_0 will be 0.

The distribution of words with different numbers of syllables is extremely uneven. Through statistics on the lexicon used in this system, it is found that the distribution of two-syllable and multisyllable words shows that the smaller the number of syllables, the larger the proportion. For example, the proportion of two syllables is 38.90%, the proportion of 3 syllables is 26.55%, and the proportion of 6 syllables is only 0.6%. Excluding single-syllable words, if all other words are grouped according to the number of syllables, the fewer the number of syllables, the fewer attributes in the instance, which can greatly simplify the learning process and improve learning effect.

This paper uses the transformation-based learning method (Transformation-Based Learning) to carry out the main stress annotation learning. This method is a machine learning method proposed by EricBrill, which has been obtained in terms of part-of-speech tagging and block segmentation. However, we have not yet seen any successful application of English accent labeling. Since accent labeling is a typical labeling problem, the conversion learning method will be a very suitable method to solve this problem.

The learning by this method is an error-driven greedy search process (Error-driven Greedy Searching). In each round of learning, the labeled phonetic symbol stream is compared with the correct phonetic symbol stream, and the wrong annotations are converted into Type conversion. The conversion template is defined in advance. For example, when learning the main accent labeling of two-syllable words, each instance contains six attributes. Since it is not known in advance which attribute is more important, any combination of these 6 attributes is a conversion template; so the number of possible templates are:

C16 + C26 + C36 + C46 + C56 + C66 = 26_1 = 63

For example, the conversion template “CON1” is a combination formed by the end of the first syllable and the core of the second syllable. The learning process based on the conversion learning method is shown in Figure 1.

It can be seen from Figure 1 that there are two main steps in the whole learning process, namely, initial labeling and labeling learning. The initial labels are subjected to labeling learning by applying various rules for candidate conversion of phonetic transcription. The rules are evaluated and the final score is calculated based on the comparison of the applied rules to the rules evaluated. Initial labeling refers to initial labeling of words in the unlabeled phonetic transcription stream as a certain label. Taking two-syllable words as an example, since the main stress of 8403 words in all 10,620 two-syllable words is on the first syllable, the initial labeling of two-syllable words is to mark all the two-syllable words in the learning corpus as 0, that is, the main stress is on the first syllable. In the subsequent tagging learning process, first for each

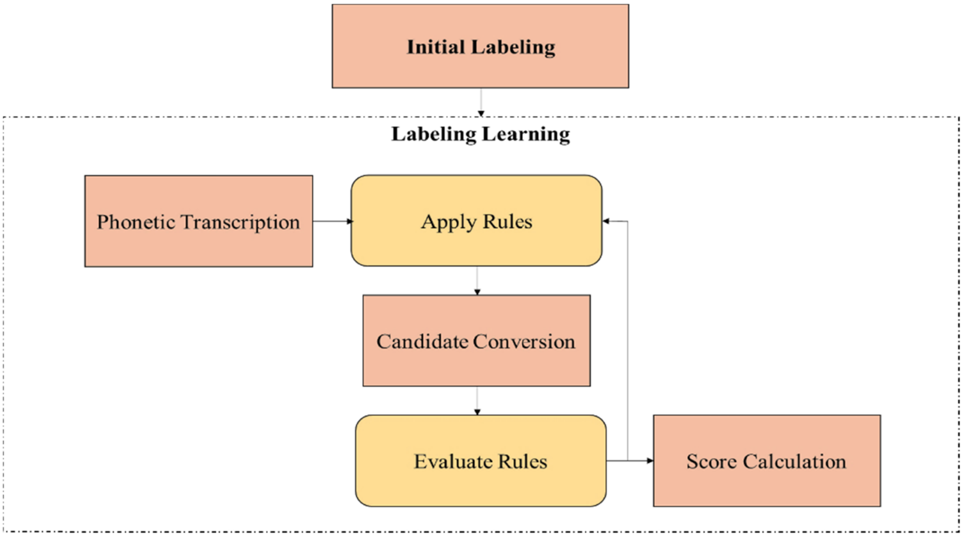

Figure 1: The learning process based on the conversion learning method.

conversion template, the phonetic transcription after initial annotation is compared with the correct phonetic transcription, and the candidate conversion is generated for the words with the main accent incorrectly labeled.

Take two-syllable words as an example. For example, for the conversion template “O_0”, it is found that the first syllable of 828 incorrectly labeled two-syllable words is empty, and there are 259 incorrectly labeled two-syllable words. The first sound of the first syllable of the syllable word is dd, and the candidate conversion formulas are generated respectively:

“O_0=0=>P=1” and “O_0=d=>P=1”

For this conversion template, this round of learning generates a total of 45 candidate conversions.

Then calculate its score for each candidate conversion formula. The score is calculated by applying this candidate conversion formula to the entire learning corpus. If a word that was originally marked incorrectly, is marked correctly, then 1 point is awarded. If the original correct word is marked incorrectly, 1 pointis deducted, and the conversion with the highest score wins. After calculation, among the 45 candidate conversion formulas, “N_0=0=>P=1” has the highest score, which is 84, so it is the winning conversion formula of the conversion template “O_0”.

Obviously, in the first round of learning, each conversion template is likely to produce a conversion with the highest score, up to 63, and then compare these conversions, and pick the conversion with the highest score as this round. In the first round, among the 63 conversion formulas with the highest scores generated by all 63 conversion templates, the conversion formula “N0=2=>P= 1” has the highest score, which is 687, hence, the first round of learning. The labeling rule is “N0=2=>P=1”, which means that if the syllable core of the first syllable of a two-syllable word is YES, it will be labeled 1, that is, the main stress is on the second syllable.

Then go to the second round of learning; first apply the rule “N0=2 =>P=1” just learned to the learning corpus, that is, mark all the two-syllable words whose core of the first syllable is a vowel sound as 1. And then repeat the above learning process until there is no candidate conversion with a score greater than 0, and the learning ends. The learning process of the main stress annotation of the entire two-syllable word is shown in Figure 2.

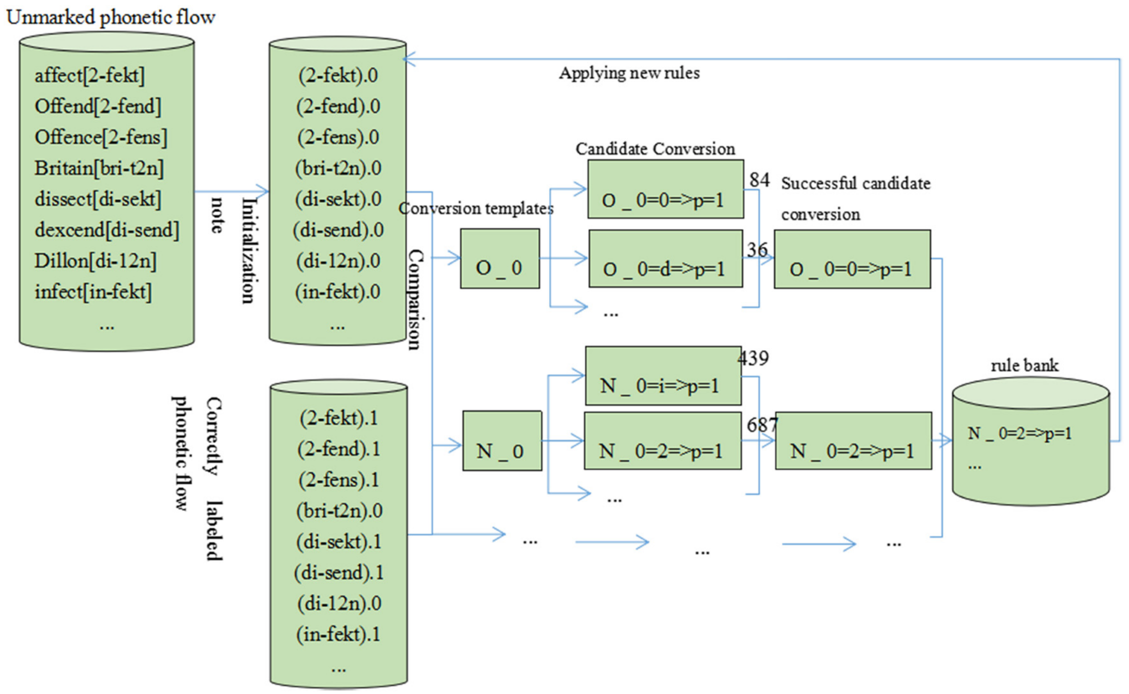

Figure 2: Learning process of two-syllable word accent marking.

4 Machine learning of sub-stress labeling rules

The secondary accent annotation learning also adopts the conversion-based learning method. Since some words have no secondary stress, some have secondary stress, and some even have more than one, it is difficult to directly use words as learning objects like primary stress when performing secondary stress labeling, but syllables must be used as learning objects. If it is the second stressed syllable, it is marked as 1, otherwise it is marked as 0. For a syllable in a two-syllable word or a multisyllable word, the following attributes will be collected:

SyllCount: The number of syllables of the word.

PrimNo: The number of the main stressed syllable (if the first syllable is the main stressed syllable, the value is 0, and so on).

CurrentNo: The number of the current syllable (if it is the first syllable, it is 0).

O0, N0, C0: respectively represent the initial, core, and end of the current syllable. A*i*(*i* = _1,_2,1,2): respectively represent the accent marks of the first syllable from the left of the current syllable, the second syllable from the left, the first syllable from the right, and the second syllable from the right (if it is the main accent, then it is P, secondary stress is S, and nonstress is N).

Taking the 4th syllable '*uu*' in arteriosclerosis[a:_,ti2_ri_u_,skli2_'ru_sis] as an example, the collected attributes are shown in Table 2. The specific learning process is similar to the learning process of the main accent labeling rules.

5 Results and discussion

Since the primary stress and the secondary stress are labeled separately, and the primary stress is labeled using an algorithm based on a combination of morphological rules and machine learning methods, multiple rule tests must be performed.

5.1 Test of main accent marking rules

The test results obtained by the main accent marking rule using the proposed method are provided in the following section for both; morphological rule-based and machine learning rule-based testing.

5.1.1 Test of morphological rules

Regarding the morphological rules, the 22,621 two-syllable or multisyllable words removed from the entire thesaurus were used as test objects. The test results of the two sets of morphological rules are shown in Table 3. Among them, "the number of applicable words," taking the α_rule as an example, means that among 22,621 words, there are 4035 words that can use the α_rule to determine the position of the main stress, of which 3969 words have the main stress discrimination correct.

Table 2: Attributes extracted for syllables

SyllCount	PrimNo.	CurrentNo.	O_0	N_0	C_0	A_-1	A_-2	A_1	A_2
7	5	3	0	$\overset{u}{2}$	0	N	S	S	P

Table 3: Test results of morphological rules

Rule	Number of rules	Number of words applicable	Proportion of total test word base (%)	Principal heavy duty discriminating positive the exact number of words (%)	Principal heavy duty discriminating correct rate (%)
α-rule	71	4035	17.8	3969	98.4
β-rule	117	7405	31.1	7317	98.8

5.1.2 Testing of machine learning rules

For the main stress labeling rules obtained by machine learning, because there are too few optional samples for words with six syllables and above in the lexicon, when learning these words in groups, all words are used as learning samples. For two-syllable words to five-syllable words, K-Fold Cross-Validation is used for rule testing. Let K be equal to 10. Take two-syllable words as an example, that is, 10,620 words Divide into 10 equal parts, each with 1062 words, for a total of 10 rounds of verification. In each round of verification, nine groups of words are used as the learning set, and the remaining 10%, or 1062, are used as the test set. The results of ten rounds of verification are shown in Table 4. It can be seen from Table 4 that the smaller the number of syllables, the higher the accuracy of the main stress labeling. The average main stress labeling accuracy of two-syllable words can reach 93.56%, while for five-syllable words, it drops to 81.72%. As the number of syllables is smaller, the proportion is greater. Therefore, the test result also proves that the decision to group words with different syllables for learning was correct. Comparing Tables 3 and 4, it can be seen that the correct rate of the main stress labeling of the morphological rules is much higher than that of the rules obtained by machine learning, and the two sets of morphological rules can cover nearly 50% of the words in the lexicon. This shows the importance of morphological rules to the main accent annotation.

5.2 Test of the second accent marking rule

For su-_stress annotation, 22,621 two-syllable or multisyllable words are used for 10 rounds of rule testing using K iterations of cross-validation. Each round uses 90% of the words for learning, and the remaining 10% is used for testing. The test results are shown in Table 5.

5.3 Comprehensive test

Finally, 22,621 two-syllable or multisyllable words are used in K iterations to cross-validate the comprehensive test of the primary and secondary stress labeling rules. In the test, first use the α_rule and β_rule to

Table 4: The ten rounds of cross-validation results of main stress machine learning rules

	Group study test			
	Average number of words	Average number of rules	Average number of words	Correct rate of average key tone marking (%)
Disyllable	9558	120	1062	93.56
Three-syllable words	6523	198	725	91.78
Quadrisyllable	3088	99	343	85.65
Five-syllable words	1028	37	114	81.72

Table 5: Results of ten rounds of cross-validation for accent labeling rules

<table>
<tr><th colspan="6">Study test</th></tr>
<tr><th>Average number of words</th><th>Average number of syllables</th><th>Average number of rules</th><th>Average number of words</th><th>Average number of syllables</th><th>Average stress marking correctness (%)</th></tr>
<tr><td>20,359</td><td>57,163</td><td>722</td><td>2262</td><td>6353</td><td>89.3</td></tr>
</table>

Table 6: The ten rounds of comprehensive cross-validation results

<table>
<tr><th rowspan="3">Average number of words tested</th><th colspan="7">Main tone marking</th></tr>
<tr><th colspan="2">α-rule</th><th colspan="2">β-rule</th><th colspan="2">Machine learning rules</th><th rowspan="2">Marking correctness (%)</th></tr>
<tr><th>Number of words identified</th><th>Marking correctness (%)</th><th>Number of words identified</th><th>Marking correctness (%)</th><th>Number of words identified</th><th>Marking correctness (%)</th></tr>
<tr><td>2262</td><td>401</td><td>98.6</td><td>707</td><td>98.3</td><td>1154</td><td>90.7</td><td>94.4</td></tr>
<tr><th colspan="8">Secondary stress markers</th></tr>
<tr><th colspan="5">Machine learning rules</th><th colspan="3" rowspan="2">Correctness of total marking (%)</th></tr>
<tr><th colspan="3">Number of words identified</th><th colspan="2">Marking correctness (%)</th></tr>
<tr><td colspan="3">2262</td><td colspan="2">86.9</td><td colspan="3">83.6</td></tr>
</table>

label the main stress, and then use the machine learning rules for the main stress labeling for the words that cannot be labeled with the morphological rule. After the main stress labeling is completed, use the machine learning method and the learning rules are labeled with secondary stress. The results of 10 rounds of verification are shown in Table 6.

It can be seen that the overall labeling accuracy is 83.6% and the labeling accuracy rate of the main accent can reach 94.4%. This approach is effective in providing the accurate main accent labeling and a satisfactory performance is achieved for the overall labeling, thereby yielding a feasible English conversation system.

6 Conclusion

This paper proposes an accent labeling algorithm based on morphological rules and machine learning. This algorithm separates the labeling of primary and secondary accents. The main accent labeling uses an accent labeling algorithm that combines morphological rules and machine learning. As for the secondary stress, since the position of the secondary stress cannot basically be distinguished by the morphological structure of the word, it can only be marked by machine learning. The proposed methods achieve the main accent labeling accuracy of 94.4% and the overall labeling accuracy of 83.6%, providing an effective and feasible English conversation system with a good overall performance.

The main accent annotation achieved satisfactory results; however, the accent labeling of compound words is somewhat complicated restricting the further improvement of the accuracy of labeling. The future work of this research work will focus on putting the compound accent annotation together with general words for accent labeling, and specialized labeling algorithms will be tested.

Conflict of interest: Authors state no conflict of interest.

References

[1] Vojir S, Zeman V, Kuchar J, Kliegr T. Easyminer.eu: web framework for interpretable machine learning based on rules and frequent itemsets. Knowl Based Syst. 2018;150(JUN 15):111–5.

[2] Zhao F, Chen Y, Hou Y, He X. Segmentation of blood vessels using rule-based and machine- learning-based methods: a review. Multimed Syst. 2019;25(2):109–18.

[3] Rodellar J, Alférez S, Acevedo A, Molina A, Merino A. Image processing and machine learning in the morphological analysis of blood cells. Int J Lab Hematol. 2018;40:46–53.

[4] Kaur G, Bhardwaj N, Singh PK. An analytic review on image enhancement techniques based on soft computing approach. In: Urooj S, Virmani J, editors. Sensors and image processing. Advances in intelligent systems and computing. Vol. 651, Singapore: Springer; 2018. doi: 10.1007/978-981-10-6614-6_26.

[5] Sharma A, Tomar R, Chilamkurti N, Kim BG. Blockchain based smart contracts for internet of medical things in e-healthcare. Electronics. 2020;9(10):1609.

[6] Rabbani A, Babaei M. Hybrid pore-network and lattice-boltzmann permeability modelling accelerated by machine learning. Adv Water Resour. 2019;126(APR):116–28.

[7] Ly C, Olsen AM, Schwerdt IJ, Porter R, Sentz K, Mcdonald LW, et al. A new approach for quantifying morphological features of u3o8 for nuclear forensics using a deep learning model. J Nucl Mater. 2019;517:128–37.

[8] Tanwar S, Bhatia Q, Patel P, Kumari A, Singh PK, Hong W. Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access. 2020;8:474–88. doi: 10.1109/ACCESS.2019.2961372.

[9] Zhao Y, Ren W, Li Z. An accent marking algorithm of english conversion system based on morphological rules. Int J Emerg Technol Learn. 2021;16(1):234–46.

[10] Cruz-Benito J, Vázquez-Ingelmo A, Sánchez-Prieto JC, Therón R, García-Pealvo FJ, Martín-González M. Enabling adaptability in web forms based on user characteristics detection through a/b testing and machine learning. IEEE Access. 2018;6:2251–65.

[11] Fan G, Shi W, Guo L, Zeng J, Gui G. Machine learning based quantitative association rule mining method for evaluating cellular network performance. IEEE Access. 2019;7:1.

[12] Tania MH, Lwin KT, Shabut AM, Najlah M, Chin J, Hossain MA. Intelligent image-based colourimetric tests using machine learning framework for lateral flow assays. Expert Syst Appl. 2020;139:112843.1–22.

[13] Stantic D, Jo J. Accent identification by clustering and scoring formants. Int J Comput Syst Eng. 2012;6(3):379–84.

[14] Kumar P, Chandra M. Speaker identification using Gaussian mixture models. MIT IJECE. 2011;1(1):27–30.

[15] Tang H, Ghorbani AA. Accent classification using support vector machine and hidden markov model. Conference of the Canadian Society for Computational Studies of Intelligence. Berlin, Heidelberg: Springer; 2003 June. p. 629–31.

[16] Novich S, Trevino A. Introduction to accent classification with neural networks. Rice University; 2010.

[17] Behravan H, Hautamäki V, Kinnunen T. Foreign accent detection from spoken Finnish using i-vectors. Lyon, France: International Speech Communication Association (ISCA)-INTERSPEECH, Vol. 2013, 2013 Aug. p. 14.

[18] Chen T, Huang C, Chang E, Wang J. On the use of Gaussian mixture model for speaker variability analysis. Seventh International Conference on Spoken Language Processing. Germany: Causal Productions Pty; 2002.

[19] Tsai WH, Wang HM. Towards automatic identification of singing language in popular music recordings. Barcelona, Spain: International Society for Music Information Retrieval; 2004 Oct.

[20] Kruspe AM, Fraunhofer IDMT. Keyword spotting in a-capella singing. Taipei, Taiwan: International Society for Music Information Retrieval, Vol. 14, 2014 Oct. p. 271–6.

[21] Nichols E, Morris D, Basu S, Raphael C. Relationships between lyrics and melody in popular music. New York City, USA: International Society for Music Information Retrieval; 2016.

[22] Jiao Y, Tu M, Berisha V, Liss JM. Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. San Francisco, CA, USA: International Speech Communication Association; 2016 Sept. p. 2388–92.

[23] Kim S, Jung H. A study on the utilization of speech recognition technology in foreign language learning applications-focusing on English and French speech. J Digit Contents Soc. 2018;19(4):621–30.

[24] Bird JJ, Wanner E, Ekárt A, Faria DR. Accent classification in human speech biometrics for native and non-native english speakers. Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments. New York, NY, United States: Association for Computing Machinery; 2019 June. p. 554–60.

[25] Zhao G, Sonsaat S, Levis J, Chukharev-Hudilainen E, Gutierrez-Osuna R. Accent conversion using phonetic posteriorgrams. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). United States: Institute of Electrical and Electronics Engineers Inc.; 2018 April. p. 5314–8.

[26] Sun L, Li K, Wang H, Kang S, Meng H. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, WA, USA: Institute of Electrical and Electronics Engineers; 2016 July. p. 1–6.

[27] Chen X, Chu W, Guo J, Xu N. Singing voice conversion with non-parallel data. 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). San Jose, California, USA: Brazilian Ministry of Education; 2019 March. p. 292–6.

[28] Fang F, Yamagishi J, Echizen I, Lorenzo-Trueba J. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). United States: Institute of Electrical and Electronics Engineers Inc.; 2018 April. p. 5279–83.

[29] Yeh CC, Hsu PC, Chou JC, Lee HY, Lee LS. Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posteriorgram sequences. In 2018 IEEE Spoken Language Technology Workshop (SLT). United States: IEEE; 2018 Dec. p. 274–81.

[30] Upadhyay R, Lui S. Foreign English accent classification using deep belief networks. 2018 IEEE 12th International Conference on Semantic Computing (ICSC). United States: IEEE; 2018 Jan. p. 290–3.

[31] Parikh P, Velhal K, Potdar S, Sikligar A, Karani R. English language accent classification and conversion using machine learning. Singapore: Springer; 2020. Available at SSRN 3600748.

[32] Zhang A, Ni C. Pitch accent prediction using ensemble machine learning. 2009 Second International Conference on Intelligent Computation Technology and Automation. Vol. 1, NW Washington, DC, United States: IEEE Computer Society 1730 Massachusetts Ave.; 2009, October. p. 444–7.

[33] Feng N, Sun B. On simulating one-trial learning using morphological neural networks. Cognit Syst Res. 2018;53:61–70.

[34] Sutton J, Mahajan R, Akbilgic O, Kamaleswaran R. Physonline: an open source machine learning pipeline for real-time analysis of streaming physiological waveform. IEEE J Biomed Health Inform. 2018;99:1.

[35] Lou Z, Ren Y. Investigating issues with machine learning for accent classification. J Phys Conf Ser. 2021;1738(1):012111. IOP Publishing.