Accuracy of Natural Language Processing Algorithms in Representing Stressed Words in Closed Captioning

Prawit Boonmee
Faculty of Informatics
Burapha University
Chonburi, Thailand
prawit@buu.ac.th

Prajaks Jitngernmadan
Faculty of Informatics
Burapha University
Chonburi, Thailand
prajaks@buu.ac.th

Kanuengnij Kubola
Faculty of Informatics
Burapha University
Chonburi, Thailand
kubola@go.buu.ac.th

Kemtong Sinwongsuwat
Faculty of Liberal Arts
Prince of Songkla University
Songkla, Thailand
kemtong.s@psu.ac.th

Pradya Prempraneerach
Faculty of Engineering
Thammasat University
Bangkok, Thailand
ppradya@engr.tu.ac.th

***Abstract*— For English as a Foreign Language (EFL) learners whose native tongue is not a stress-timed language, it is a difficult task to master word stress in speaking. Despite technological advances and years of learning English in school, many Thai EFL learners, in particular, are struggling with correct stress placement in English words, especially when conversing in English. With knowledge of natural language processing and English linguistics, this study creates closed captions to provide additional visual information on stress placement to help these learners making themselves better understood in English conversation. This also helps those who are deaf or hard of hearing to recognize highlighted audio narratives properly. The purpose of this study is to determine the accuracy of word stress representation in closed captions via human auditory analysis and perception. The results show that the current approach delivers 74.19% of representation accuracy.**

Keywords*— *natural language processing, closed captioning, algorithms, English as a foreign language, conversation analysis (CA)

I. INTRODUCTION

Unlike many Asian languages including Thai, English is a stress-timed language in which different stress placement in words can result in different meanings and parts of speech. For instance, if one stresses on the second syllable of the word *reCORD*, it becomes a verb; however, when stress is placed on the first syllable *RECord*, it turns into the noun with meaning changes. Incorrect stress placement can easily lead to misunderstanding problems in oral English communication Therefore, it is important for EFL learners to master correct stress in English pronunciation to be fluently and effectively engaged in English conversation.

Noted in K. Tantiwich and K. Sinwongsuwat [1], stress placement was among the most frequently found problems in Thai university students' English use in conversation including deletion of the final sound in inflectional endings and of /l/ and /r/ in consonant clusters. It is therefore likely that these learners will encounter communicating and understanding problems when engaged in conversations with English speakers.

With increasingly sophisticated multimedia and advanced educational technology, internet platforms have been recommended such as YouTube, TED Talk, and YouGLISH, allowing both learners and teachers to easily access various media including video clips of naturally spoken English with transcripts and closed captions. Closed-captioning, originally developed for deaf and hard-of-hearing people, has in fact been widely used to teach the target language to L2 learners, not only exposing them to the world of English speakers, but helping them become familiar with global varieties of English with different accents. Closed captions (CC) provide learners with a better control of the verbal element of audiovisual material and crucial support for their informal and autonomous learning.

When watching videos, one has a chance to turn on CC if it is available. However, the CC is normally presented in the form of sentence strings. When it comes to learning prosodic features of spoken language such as stress and intonation, a video watcher cannot tell from the text which part of a word or sentence has to be stressed. Therefore, the enrichment of the text of the CC will help the video watchers, especially English learners, to get more of such information as word stressing, giving them a better chance to acquire English pronunciation. The closed captions enhanced with stress features in videos can especially be used to help English learners better notice where to stress each word.

One way to convey word stress in English speech is to use capitalized letters. Stressed syllables are capitalized in vocabulary building so one knows how to pronounce the English words and how words get stressed in sentences or natural utterances. This makes it simple especially for EFL learners to practice and memorize English word stress. Our research question accordingly is: How accurately do natural language processing algorithms we have developed represent stressed words in closed captioning?

In this work, the concept of ARPABET [2] is implemented. The ARPABET is another way to represent phonetic alphabets that are similar to the International Phonetic Alphabet (IPA) or a set of symbols for phonetic transcriptions. The IPA was created in 1888 and has

undergone several changes since then. It aimed at trying to link a symbol to each sound in all of the known languages in the world. P. Ladefoged provides more details about the IPA symbols [3]. The problem with the IPA is that it makes extensive use of letters that are not usually available on computers. Therefore, associating IPA to ASCII symbols as the ARPABET was proposed to ease the use of computers. The ARPABET is a selection of symbols used within the Advanced Research Projects Agency Speech Understanding Research (ARPA SUR) project [3]. There are two representations in ARPABET as follows: 1) one embraces only one character and includes lower-case letters. 2) The second applies only to upper-cased letters and is known as "2-characters."

Pronunciation dictionaries are widely available and used for both speech recognition and synthesis, including the CMU dictionary for English and CELEX dictionaries for English, German, and Dutch [5].

The CMU Pronunciation Dictionary (also known as CMUdict) [6] is an open-source articulating lexicon initially developed by Carnegie Mellon University (CMU) for utilization in speech recognition research. CMUdict gives a mapping of orthographic and phonetic for English words in their North American pronunciations. It contains over 134,000 words and their elocutions. CMUdict is being effectively kept up to date and extended. It has mappings from words to their elocutions within the ARPABET phoneme set, a standard for English elocution. The current phoneme set contains 39 phonemes, vowels carry a lexical stretch marker: 0 — No stress, 1 — Primary stress, and 2 — Auxiliary stress.

Another research done by employing ARPABET is SoundSpelling, which converts English text into SoundSpelling notation [6]. They generate the information on syllables and sound intensity used to create teaching and learning materials for Japanese students who learn English in High School.

As a test and representative platform, the YouTube platform is selected. It is well-known for video-sharing that can be reached at www.youtube.com. It allows users to post videos, watch existing videos, and split videos. Short movie clips, TV show clips, music videos, and video blogging are all examples of content information. The majority of video clips on YouTube are short (1-10) minute snippets filmed by members of the general public. Recent files, most seen files, and most voted files are the three categories in which clips are categorized and rated. The most important feature used in this work is the closed caption feature (CC) [7]. There are also the psychological and cognitive benefits of using videos with captioning. The study reveals that improved captioning has a more positive impact on the motivation of intermediate-high and advanced German learners. The findings provide valuable insight and enthusiasm for the effective use of Youtube videos in curriculum building of German Courses [8]. A user can type in, upload, or use the auto-translate function for the transcription of a video clip. This versatility allows us to equip a video clip with a manipulated CC according to the ARPABET notations mentioned above.

This work presents how to employ phonological studies in closed captions to improve English learners' prosody. A digit immediately following a vowel indicates stress. Auxiliary symbols in 1- and 2-letter codes are identical. A space separates parts in 2-letter notation [2]. The accuracy of the notations comparing to the real pronunciation is evaluated based on linguistic experts.

II. Method and Procedure

As linguists have defined different levels of analysis for natural language, this work studies, in particular, prosody that involves rhythm and intonation of language.

Regarding pronunciation, CMUdict uses the expectation-maximization (EM) algorithm in order to produce reasonable pronunciations for unseen words when building letter to sound (LTS) rules from a word list in a language and the algorithm is shown in [8].

The proposed approach is to turn regular words of a sentence into emphasized words with special notations. For the next step, this information will be embedded into Youtube's closed captions and can be then accessed via YouTube player's options. By using this approach, it reduces the expense of developing a new application for learning proper English pronunciation. Furthermore, the English learners can notice and memorize the notations and they will last longer in the long-term memory processing.

The procedure overview includes following steps:

1. **Input**: In this step, the transcripts of a video will be read as text files.
2. **Processing**: In this step, each word of the transcripts will be validating and syllabifying using CMUdict.
3. **ARPABET**: In this step, the syllabifying words will be then converted from ARPABET to stressed words.
4. **Output**: In this step, the final transcripts containing the stressed words will be then put in the closed captions for the video clips.

Figure 1 depicts the procedure overview of the approach.

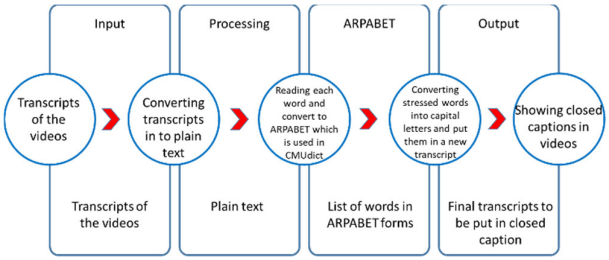

Fig. 1. The overall procedure to get closed captions with stressed words.

The word stressing can be divided into 3 parts, as follows:

1. **ARPABET number 0**: It is an unstressed word, e.g. ['AH0', 'V'] → of.
2. **ARPABET number 1**: It is the most stressed tone (Primary Stress), e.g. ['HH', 'AH0', 'L', 'OW1'] → helLO
3. **ARPABET number 2**: It is secondary (Secondary Stress), e.g. ['AE1', 'D', 'R', 'EH2', 'S'] → ADDRESS

For testing and evaluating purpose, a program had been created. This special program is written in Python and makes use of the following tools: 1) docx2txt, which is used to convert Microsoft Office documents (Words) to text format so that data can be imported into a work environment, 2)

NumPy, it is a mathematical calculator that works with matrices, 3) Pandas, it is Python's primary data manipulation tool, and it can be used in conjunction with other packages like SKLearn to prepare data prior to modeling, 4) NLTK is used to do natural language processing, and 5) FinnSyll, it is a syllable splitting Python library. According to the syllabary principle, it is more practical to work on the highlighted forms of the word before adding words and uploading them to YouTube.

Here are the steps involved in order to build closed captions:

1. The transcripts of the English lesson video are read in and converted into plain text.
2. Each word is read and converted to ARPABET which is used in CMUdict.
3. Stressed words are converted into capital letters and put in a new transcript.
4. Put new transcripts in Youtube's closed captions.

Figure 2 shows the transcript of an English conversation video. The words are separated and converting according to the ARPABET phonetic concepts using CMUdict framework. In the highlighted words, the word "sorry" is converted as follows:

sorry ----- > ['S', 'AA1', 'R', 'IY0'] ----- > SOrry

That means, the English learner has to stress the SO in the word "sorry" in terms of pronunciation correctness.

Another example is the word "neighbor", which can be converted as follows:

neighbor ----- > ['N', 'EY1', 'B', 'ER0'] ----- > NEighbor

That means, the English learner has to stress the NE in the word "neighbor" in terms of pronunciation correctness.

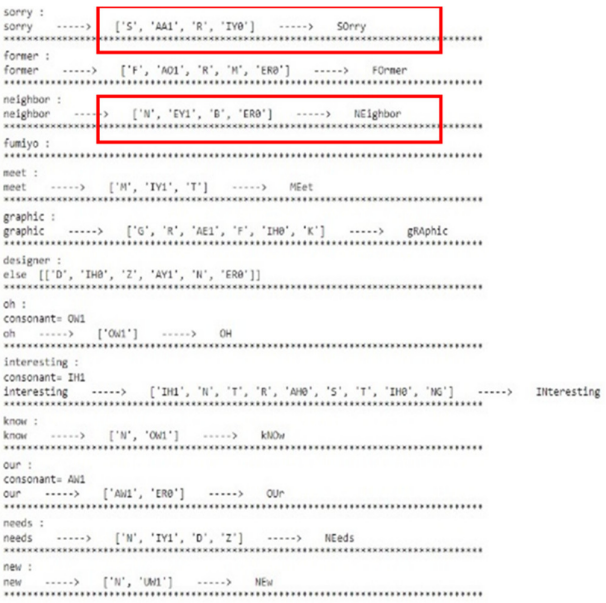

Fig. 2. Converting original words to stressed forms.

III. Results and Discussion

One of the results of the special program is the stressed word transcripts that can be imported into the YouTube video closed caption system. The Figure 3 and 4 show the example videos for English learning with closed caption options activated.

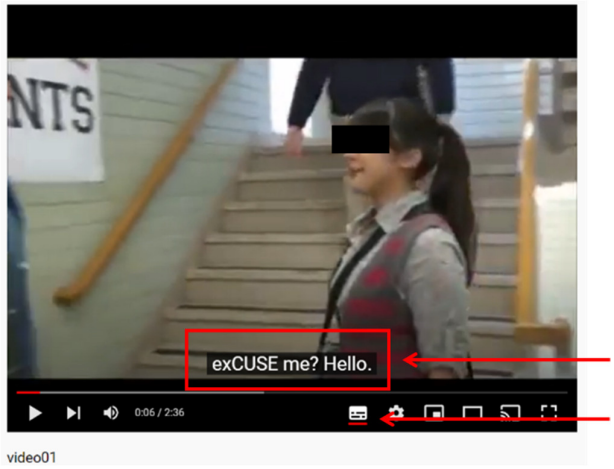

Fig. 3. A resulting video with enhanced closed caption 1 [8]

Figure 3 and Figure 4 indicate the conversation in the videos with the manipulated transcripts in the closed captions. To see this option, a YouTube user must turn on the CC options in the option menu (bottom right area of the player). With the enriched transcripts, the English learners have more information and they can create their own way of learning and memorizing how to stress a particular word in a sentence or a conversation.

In Figure 3, the sentence "exCUSE me? Hello." is capitalized where the learners have to stress the word or part of the word.

It is also the same as in Figure 4. The sentence "Sue. It's nice TO SEE you." is also indicated where to stress the words correctly.

Fig. 4. A resulting video with enhanced closed caption 2 [9]

The manipulated transcript of a video looks like Figure 5. It is the example of the edited transcript of the English conversational video in the Figure 4. In this representation, an evaluation method is also proposed. The manipulated

transcript shows the words in red that do not align with the speakers' emphasis/stress from auditory evaluation by the linguists. On the other hand, the words highlighted in yellow do align with the speakers' stresses of the conversations.

Sue: bob?
Bob: sue. It's nice TO SEE YOU.
Sue: It's great TO SEE YOU, too. HOW are YOU?
Bob: I'm just fine. sue. THIS IS FUmio suZUki.
Sue: HI. FUmio.
Bob: FUmio IS A chef at the best greek REStauRANT IN town. FUmio. THIS IS MY old NEIGHbor, sue MURPhy.
Sue: old
Bob: SORry. FORmer NEIGHbor.
Fumiyo: It's nice TO meet YOU.
Sue: It's nice TO meet YOU, too.
Bob: sue IS A GRAPhic deSIGner.
Fumiyo: OH! That's INteresting.
YOU know our REStauRANT needs A new MEnu deSIGN.
Sue: OH REally.
Fumiyo: DO YOU DO THAT sort of thing?
Sue: YES. I've deSIGNED LOGOS and MEnus FOR SEveral RESTauRANTS IN town.
Fumiyo: Let's get toGETher and talk Over some iDEas.
Sue: great!! Here's MY CARD. I. am with deSIGN quest.
Fumiyo: thanks. I. don't HAVE MY CARD with me but I CAN give YOU MY phone NUMber at the RESTauRANT.
Sue: OH, YES DO. OKAY.
Fumiyo: It's 646-555-1330.
Sue: OKAY, 646-555-1330. and YOUR NAME IS FUmio. could YOU sPELL THAT FOR me?
Fumiyo: SURE. Hmm. FUmio F-U-M-I-Y-O. suZUki S-U-Z-U-K-I.
Sue: thanks. are YOU JApaNESE.
Fumiyo: YES, I, am. .
Sue: and YOU are A chef at A greek RESTauRANT.
Fumiyo: greek COOking IS MY SPEcialty.
Sue: It's A long STOry.

Fig. 5. Correspondence with the speakers' emphasis/stress from aural assessment

Figure 5 shows the manipulated transcripts that is stressed correct and incorrectly approved by linguistic experts. The sentence "Bob: sue. It's nice TO SEE YOU." indicates "sue" as wrong stressing because in the conversation video, Bob was surprised to see Sue on the street. The created program apparently pickups only the phonetic pronunciation with no relation of emotional responses of the speakers. However, the results of the evaluation done by linguistic experts are still usable, especially for the English learners who are self-learners and are at the beginning of English learning. The Table 1 shows the example of the manipulated words with the represented capitalized letters of the stressing. In addition, the Table 2 shows the evaluation results and accuracy of the first version of this special software.

TABLE I. EXAMPLE OF MANIPULATED WORDS THAT REPRESENT SOUNDS

#	Mono VS. Multisyllabic words that represent sounds	Correspondence with the speaker's emphasis/stress from the aural assessment	
		agreeable	disagreeable
1	SUe	/	
2	BOb	/	
3	NIce	/	
4	TO		/
5	SEe	/	
6	YOu	/	
7	gREat	/	
8	too		/
9	ARe	/	
10	FIne	/	
11	IS	/	
12	suZUki	/	
13	AT		/
14	BEst	/	
15	gREek	/	
16	REstaurant	/	
17	TOwn	/	
18	MY		/
19	OLd		/
20	NEIGHbor	/	

Table 1 shows the first 20 words that were manipulated with the approach presented above. Where the word should be stressed in the pronunciation, it is notated with capitalized letters. The word "Suzuki", for example, the correct stressing of the word would be "suZuki". Therefore, the evaluation of this word pronunciation suggested by our special software is correct. In the same way of evaluation, the word "too" is suggested as "too" for stressing in the pronunciation, which is not correct. The correct stressing of the pronunciation would be "TOO". The results are then collected and analyzed in the descriptive statistics. Table 2 depicts the evaluation summary of the approach.

TABLE II. ACCURACY OF THE APPROACH

Words	Evaluation Summary		
	Agreeable	Disagreeable	Accuracy (%)
62	46	16	74.19

The results we have achieved in this work are the new transcripts for the video clips with words highlighted (in capitalized letters) by their different stresses. This can be used in the closed captions (CC) options of a video player. It can help English learners practice pronunciations more easily with enhanced closed captions from our work, especially for beginners. Furthermore, a deaf person can learn and see how the words should be stressed in a conversation. This might convey the emotional response of the conversation to a deaf learner, too. However, the accuracy rate shows that it is still a

number of things to consider in case of improving the correctness. Several factors such as stress shift and the program can only handle the stress of only multiple syllable words but is unable to manage sentence stress. As for future work, we may need to work on acoustic analysis for better results. We plan to study further intonation, one-syllable stress and sentence stress.

IV. CONCLUSION

In this work, we showed the way that the stressing of English words or sentences can be done using CMUdict and ARPABET method. The contribution of this work can benefit EFL learners and teachers at any levels. The stress or emphasis of a word or part of it is represented using capitalized letters. The pronunciation of that words can also be assumed from the text right away. In a use case with YouTube English conversational video clip, these manipulated words can be used in the closed captions of the video. With this enhanced text, a learner has an opportunity to extend the way of learning and memorizing the word pronunciation.

The accuracy rate of the approach is evaluated by linguistic experts. They evaluated the correspondence of the manipulated words with the speaker's emphasis/stress from the aural assessment of the example conversational video clips. The agreeable rate is 74.19 %and the disagreeable rate is 25.81 .%This reveals that actual conversation has several factors that reflect real phonologies .The proposed program needs an improvement to better capture this phenomenon .The results show that only employing CMUdict alone cannot cover all cases of prosodies. Acoustic analysis may need to be taken into account for better results.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Porames Khamkaew and Mr.Patipan Pakkerate for helping with Python programming. We are also grateful to have access to English conversation lessons created in the project funded via Prince of Songkla University by the 2019 Thailand Research Fund for Integrated Research and Innovation (Project ID: LIA6201041S). Finally, we would like to thank the Faculty of Informatics, Burapha University for supporting this research.

REFERENCES

[1] K. Tantiwich and K. Sinwongsuwat. Thai university students' problems of language use in English conversation. LEARN Journal, 14(2), 598-626, 2021

[2] Wikipedia, ARPABET. [Online]. Available: https://en.wikipedia.org/w/index.php?title=ARPABET&oldid=1062602312 (accessed: May 29 2022).

[3] P. Ladefoged and K. Johnstone, A Course in phonetics. Stamford CT: Cengage Learning, 2015.

[4] D. Jurafsky and J. H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed. Upper Saddle River N.J.: Pearson Prentice Hall, 2009.

[5] k. Lenzo, The CMU Pronouncing Dictionary. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=Night (accessed: May 29 2022).

[6] K. Ogi and K. Nakatsuka, Sound Spelling - CLR Phonetics Lab. [Online]. Available: http://clrlab1.u-aizu.ac.jp/soundspelling_e.html (accessed: May 29 2022).

[7] YouTube, Add subtitles and captions - YouTube Help. [Online]. Available: https://support.google.com/youtube/answer/2734796?hl=en (accessed: May 29 2022).

[8] P. Yang, "The Cognitive and Psychological Effects of YouTube Video Captions and Subtitles on Higher-Level German Language Learners," in Technology and the Psychology of Second Language Learners and Users, Freiermuth, M. R. Freiermuth, and Scott, Eds., 1st ed., [S.l.]: Springer International Publishing, 2020, pp. 83–112.

[9] Alan W. Black, Kevin A. Lenzo, and Vincent Pagel, "Issues in building general letter to sound rules," in SSW, 1998.

[10] S. Borem, Welcome! Unit 1 Intro Interchange Video. [Online]. Available: https://www.youtube.com/watch?v=JD2Umh3QJIE (accessed: May 29 2022).

[11] S. Borem, English Lesson 1 WV1 - It's nice to meet you. [Online]. Available: https://www.youtube.com/watch?v=9Ve7dCKeo4c (accessed: May 29 2022).