

Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech

Yin May Oo[†], Theeraphol Wattanavekin, Chenfang Li[†], Pasindu De Silva,
Supheakmunkol Sarin, Knot Pipatsrisawat,
Martin Jansche[†], Oddur Kjartansson, Alexander Gutkin

Google Research

Singapore, United States and United Kingdom

{mungkol,thammaknot,oddur,agutkin}@google.com

Abstract

This paper introduces an open-source crowd-sourced multi-speaker speech corpus along with the comprehensive set of finite-state transducer (FST) grammars for performing text normalization for the Burmese (Myanmar) language. We also introduce the open-source finite-state grammars for performing grapheme-to-phoneme (G2P) conversion for Burmese. These three components are necessary (but not sufficient) for building a high-quality text-to-speech (TTS) system for Burmese, a tonal Southeast Asian language from the Sino-Tibetan family which presents several linguistic challenges. We describe the corpus acquisition process and provide the details of our finite state-based approach to Burmese text normalization and G2P. Our experiments involve building a multi-speaker TTS system based on long short term memory (LSTM) recurrent neural network (RNN) models, which were previously shown to perform well for other languages in a low-resource setting. Our results indicate that the data and grammars that we are announcing are sufficient to build reasonably high-quality models comparable to other systems. We hope these resources will facilitate speech and language research on the Burmese language, which is considered by many to be low-resource due to the limited availability of free linguistic data.

Keywords: speech corpora, finite-state grammars, low-resource, text-to-speech, open-source, Burmese

1. Introduction

The Burmese (or Myanmar) language belongs to the Sino-Tibetan language family. It is the largest non-Chinese language from that family (Jenny and Tun, 2016) with the total number of speakers in Myanmar and abroad is estimated at around 43 million: 33 million native (L1) speakers with the additional population of 10 million second-language (L2) speakers (SIL International, 2019). Due to the relative scarcity of the available language resources Burmese is often considered to be an under-resourced language (Goldhahn et al., 2016).

With regard to speech resources, and in particular to text-to-speech (TTS), this scarcity is especially pronounced. Unlike the automatic speech recognition (ASR), TTS corpora for state-of-the-art or commercial systems has traditionally been recorded in professional studios by dedicated voice actors. This process is both expensive (good voices are hard to find) and time consuming (considerable effort is spent on supervising the recordings and maintaining a steady high audio quality). With the advance of new approaches that range from utilizing the found data for under-resourced languages (Baljekar, 2018; Cooper, 2019), crowd-sourcing (Gutkin et al., 2016) to multi-speaker and multilingual sharing (Li and Zen, 2016; Chen et al., 2019; Nachmani and Wolf, 2019), building TTS systems for under-resourced languages has become simpler (Wibawa et al., 2018; Prakash et al., 2019). However, the availability of a Burmese TTS corpora that is free for all is still an issue: the corpora reported in the literature are usually proprietary. On the other hand, the recent multilingual open-source datasets (Black, 2019; Zen et al., 2019) do not include Burmese.

Two further crucial components in the TTS pipeline include text normalization and pronunciation modeling. Text normalization is the process of converting non-standard words (NSWs), such as numbers and abbreviations, into standard words so that their pronunciations can be derived by consulting the pronunciation component (Sproat et al., 2001). Pronunciations, usually provided in terms of phonemes, are obtained either by direct dictionary lookup or estimated from the orthography using machine learning methods (Bisani and Ney, 2008).

This paper introduces three open-source components developed for Burmese TTS: a free speech dataset, text normalization grammars and a grapheme-to-phoneme (G2P) conversion grammar. The last component is especially useful in lexicon development for bootstrapping the pronunciations for unknown words that can later be checked and edited, if necessary, by human annotators.

This paper is organized as follows: We briefly review the related research in the next section. Section 3 presents basic linguistic detail on Burmese relevant to this work. A linguistic front-end of the Burmese pipeline is presented in Section 4. In particular, that section describes the open-source grammar components for text normalization (Section 4.2) and grapheme-to-phoneme conversion (Section 4.4). The details of the Burmese speech corpus that we are releasing are provided in Section 5. Experiments that use this corpus to build multi-speaker Burmese TTS system are described in Section 6. Section 7 concludes the paper.

2. Related Work

Despite the under-resourced status of the language, Burmese speech research is steadily becoming a burgeoning field with an increasing number of applications within both ASR (Mon et al., 2017; Chit and Khaing, 2018; Mon et

[†]The author contributed to this paper while at Google.

al., 2019) and TTS (Win and Takara, 2011; Soe and Thida, 2013b; Hlaing et al., 2018). Providing a comprehensive overview of the applications is outside the scope of this paper, instead we specifically deal with the research directly related to the language resources that we developed.

TTS Corpora The corpora used for developing Burmese TTS applications have predominantly been developed in-house: The diphone-based concatenative synthesis system reported in (Soe and Thida, 2013b; Soe and Thida, 2013a) uses an in-house diphone database developed at University of Computer Studies (UCS) in Mandalay (Myanmar). For the first statistical parametric Burmese system based on Hidden Markov Models (HMMs) developed by Thu et al. (2015c), the high-quality in-house Burmese speech dataset was developed jointly by UCS in Yangon (Myanmar) and NICT in Kyoto (Japan). The neural network-based TTS systems described in (Hlaing et al., 2018; Hlaing et al., 2019) employ an in-house Phonetically Balanced Corpus (PBC) which was constructed from the Myanmar portion of a Basic Travel Expression Corpus (BTEC) originally created for Japanese (Kikui et al., 2003). A small phoneme database consisting of 133 segments was recorded by Hlaing and Thida (2018) for their low-footprint phoneme-based synthesizer, a database which is too small for building practical state-of-the-art models. All the above corpora were recorded in professional studios and, to the best of our knowledge, are not in the public domain.

Text Normalization Our text normalization grammars for Burmese are based on the finite-state transducer grammar framework called Thrax (Tai et al., 2011; Roark et al., 2012). The release of these grammars (Google, 2018b) continues the line of work aimed at open-sourcing text normalization grammars for low-resource languages (Sodimana et al., 2018). These grammars have origins in our internal text normalization framework (Ebden and Sproat, 2015; Ritchie et al., 2019). While text normalization state-of-the-art is moving towards trendier machine learning methods (Bornás and Mateos, 2019; Zhang et al., 2019; Mansfield et al., 2019; Gokcen et al., 2019), we believe that the finite-state methods are still extremely useful in low-resource scenarios, such as Burmese, where one does not have access to sufficient amounts of training data to build sophisticated models (Nikulásdóttir and Guðnason, 2019). One alternative approach in a low-resource scenario is to apply a minimally supervised method for inducing the grammars (Ng et al., 2017). This, however, may not be necessary for Burmese because finding a native speaker with the necessary linguistic knowledge for developing such grammars is not that hard.

The only other comprehensive text normalization framework for Burmese was developed by Hlaing et al. (2017). The authors also base their framework on Thrax and provide details of text normalization grammars for various classes of non-standard words (NSWs) in their paper. The sets of NSW types they support and our types are different. Both frameworks support numbers, digits, dates, times, currencies and ranges. Hlaing et al. (2017) also support sport scores and national identification numbers (NRCs), while we offer support for decimals, fractions, letter sequences,

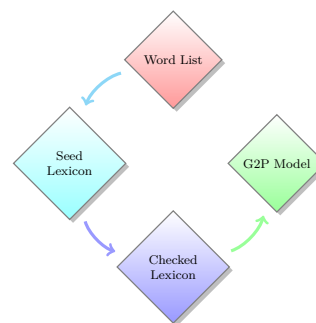


Figure 1: Possible approach to creating a lexicon.

measures, emoticons and telephone numbers. Since the grammars described in (Hlaing et al., 2017) are not available in public domain, it is difficult to compare the two implementations of the overlapping types and their grammatic coverage.

G2P Grapheme-to-phoneme (G2P) conversion models form an integral part of TTS pipelines. A G2P component provides pronunciations for words not found in the pronunciation dictionary. Modern G2P systems, including the ones developed for Burmese (Thu et al., 2015a; Thu et al., 2015b; Thu et al., 2016), employ machine learning methods (Bisani and Ney, 2008; Novak et al., 2012), which require grapheme-phoneme pairs for training. The graphemes can be obtained from the available dictionaries, such as the standard data from Myanmar Language Commission (MLC) (Nyunt et al., 1993), or scraped from the web. Given the orthography, in the absence of statistical model pronunciation rules are required to generate the corresponding pronunciation. A simplified diagram of this process is shown in Figure 1. This paper introduces a resource, which is a weighted finite-state transducer grammar based on Thrax, for generating Burmese pronunciations from Unicode orthography. While the authors of (Thu et al., 2015a; Thu et al., 2016) published some of the details of their phoneme conversion methods and shared the resulting pronunciation dictionaries¹, they did not provide software for generating such dictionaries, a gap that this work tries to fill (Google, 2018a). This step corresponds to transition between the word list and the seed lexicon in Figure 1.

Another interesting alternative is Epitran, a multilingual open-source G2P system based on context-sensitive rewrite rules (Mortensen et al., 2018). The set of 61 languages supported by Epitran includes Burmese, but beyond the medial consonants support the set of phonological rules is rather limited at present and for this language the system currently functions more like the transliterator into IPA (International Phonetic Association, 1999).

3. The Script and Phonology of Burmese

The Writing System The Burmese language uses an *abugida* (or alphasyllabary) writing system. Burmese script belongs to the family of Brahmic scripts, where every consonant can function as a standalone syllable with an inherent vowel sound /a/ (Roop, 1972; Wheatley, 1996) articulated with a creaky tone in full open syllables. The inherent

¹Available at <https://github.com/ye-kyaw-thu/myG2P>.

vowel can be changed to other vowels using diacritic marks placed around the consonant. Burmese script has 33 letters representing basic consonants. In addition, Myanmar Letter A (အ), which has properties of both consonant and vowel, and Myanmar Letter Great Sa (ခ), a special form of Myanmar Letter Sa (ဆ), are considered as consonants (Hosken, 2007).

Certain consonants can be combined together, or *stacked*, to represent consonant clusters using *virama*, the invisible character for syllable chaining. The further signs include a set of free vowel letters that can appear in syllable initial position, dependent vowel and tone diacritics, and virama, or *asat* (meaning “to kill” in Burmese) (Ding et al., 2016), for suppressing the inherent vowel in certain syllable codas. There are also two punctuation marks which in their usage are remotely related to the function of comma and full stop in Western scripts (Okell et al., 2010; Jenny and Tun, 2016). Burmese is written left to right. Similar to other Southeast Asian languages, such as Thai, Lao, and Khmer, Burmese writing system does not use whitespace to separate words. In fact, the concept of words is not always clearly defined and is often subjective (Wheatley, 1990; Jenny and Tun, 2016). Word segmentation is thus a critical first step in a Burmese natural language processing pipeline. Without the ability to identify words, none of the conventional natural language processing methods are applicable. Dealing with word segmentation in Burmese is uniquely challenging in a few ways. First of all, the most popular Burmese font used by most websites is the Zawgyi font, which is not Unicode compatible (Liao, 2017; Arnaudo, 2019). Therefore, text extracted from online sources will contain a mixture of Zawgyi and Unicode encodings. Secondly, Burmese has a relatively complex writing system. A word in Burmese may consist of as few as one character (e.g., a consonant) or may be a combination of consonants and many diacritic marks. Moreover, diacritic marks in Burmese words have specific order (Hosken, 2007) which is often violated in practice when typing. The resulting visual representation is correct despite the wrong underlying character sequence.

Phonology Following Wheatley (1987) and Jenny and Tun (2016), Burmese has 34 consonant phonemes (roughly corresponding to the respective basic consonant script symbols, e.g. “ဆ” /t^h/) occurring either alone, in syllable-initial positions or as part of consonantal clusters. The stops are represented by 15 phonemes: voiced, voiceless aspirated and voiceless unaspirated, with aspiration providing a crucial phonological contrast. The group of fricatives consists of five alveolar, palatal and glottal sounds. There are eight nasal sounds (labial, alveolar, palatal and velar) and six approximants with labial, dental, alveolar and palatal places of articulation. The number of consonantal clusters is relatively small similar to other Sino-Tibetan languages.

Burmese has seven basic vowels, one neutral vowel (or *schwa*) and four diphthongs which only occur in nasalized and stopped syllables (Green, 2005). Burmese tones are characterized by their pitch, contour, length and phona-tion type (creaky and breathy voice). The combination of these features make tonal contrasts (Jenny and Tun, 2016). Green (2005), following Wheatley (1987), describes four-way tone contrasts in major syllables (low, high, creaky and

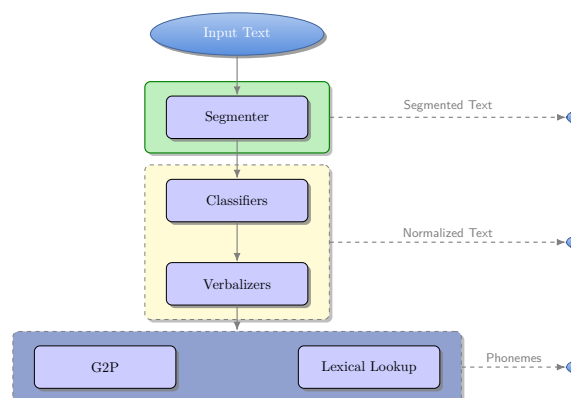


Figure 2: Simplified depiction of the linguistic pipeline.

checked) with only three tones possible in syllables with nasal rhymes (low, high and creaky). Different tones denote different meanings for syllables with the same phonemic structure.

We provide more details on Burmese phonology in Section 4.3 that describes the phoneme inventory used in our pipeline.

4. Linguistic Front-End

This section provides an overview of the core components in the TTS linguistic pipeline that precede the acoustic model. A simplified depiction of the pipeline is shown in Figure 2 (with some initial script conversion and normalization steps omitted).

4.1. Word Segmentation: Corpus and Models

To build our in-house segmentation corpus² we crawled popular Burmese websites and extracted sentences by splitting text using the Burmese full-stop symbol (။, U+104B). We tried to cover many topics by targeting diverse websites covering categories such as sports, entertainment, legal, medical, governmental affairs and so on.

We carefully selected the websites that mostly contained Unicode. Nevertheless, the sentence extraction process yielded many sentences with Zawgyi encoding. We applied a pattern matching algorithm to identify these sentences and transformed them to Unicode (these steps are not shown in the overall diagram in Figure 2). We removed all the punctuation, special characters, Latin characters, Latin numbers as well as zero-width joiner (ZWJ, U+200D) from the text. As mentioned above, the Burmese writing system is fairly complex, where a syllable can contain up to nine Unicode characters. As a result, it is common for the text to contain various types of typographical mistakes. Therefore, we also ran a special tool to remove common patterns containing diacritic combinations that were invalid. Other spelling mistakes discovered during annotation were fixed manually.

Segmentation Guidelines As mentioned in Section 3, the Burmese word segmentation task inherently contains a fair amount of ambiguity and subjectivity. We created

²Similar to other Burmese segmentation corpora, like the one used by Ding et al. (2016), our corpus and the resulting segmentation models described in this section are currently not in public domain. We plan to address this shortcoming in future work.

Sentence:	ရန်ကုန်မြို့တွင်နေသမည်။
English translation:	It does not rain in Taunggyi city.
Segmented:	ရန်ကုန် မြို့ တွင် နေ သာ မည်
Literal translation:	Taunggyi city at rain does-not-fall
Sentence:	တောင်ကြီးမြို့တွင်မိုးမရွာပါ။
English translation:	It does not rain in Taunggyi city.
Segmented:	တောင်ကြီး မြို့ တွင် မိုး မရွာပါ
Literal translation:	Taunggyi city at rain does-not-fall

Table 1: Segmentation examples.

Algorithm	Precision	Recall	F-score
CRF	95.74	96.57	96.15
NN	94.96	96.32	95.64

Table 2: Performance of segmentation models.

the segmentation guidelines that explained how Burmese text should be segmented in different contexts in order to standardize the process. The guidelines served as a reference for all the human annotators during the segmentation process. In general, we aimed to segment the sentences at word boundaries. However, since the concept of “words” in Burmese is not always clear-cut, there were many ambiguous situations. As the corpus was primarily aimed for use as part of a text-to-speech system, we paid special attention to the pronunciations of the resulting segments. In particular, we avoided segmenting a compound word if the pronunciations of the resulting segments were different from the pronunciation of the compound. For example, stem verbs can have certain suffixes to form adjectives, adverbs, or other forms. The pronunciation of the words would be wrong, if they were segmented down to stem-verb and morpheme level. Therefore, the annotators were instructed to keep the words meaningful and articulable. If a segmentation would cause the pronunciation of the subsequent segments, when pronounced separately, to be different from the true pronunciation in that context, then we do not make that segmentation. For example, the word သွားပွတ်တံ (toothbrush), which is pronounced as /θə . bʊʔ . tən/, consists of 3 smaller meaningful units: သွား (teeth) ပွတ် (brush) တံ (stick). However, if this word is segmented into three segments, their pronunciations would become /θwá/, /pʊʔ/, /tən/, which sound unnatural. Another example is နွားမ (cow), which is pronounced as /nə . mə/. However, if this word is segmented into smaller meaningful units, which are နွား (ox, gender-neutral) and မ (female), it will be pronounced as /nwá/ and /mə/, which is unnatural. In both of these cases, we do not segment the inputs.

A few examples of full Burmese sentences and their segmentation illustrating our segmentation principles are shown in Table 1. In the second example, မရွာပါ (“does-not-fall”) is kept together as one segment, because segmenting it into smaller units would alter its pronunciation.

Post-processing After segmentation, we filtered out those entries with more than 40 words from our corpus, because they tend to be run-on sentences or sentences with

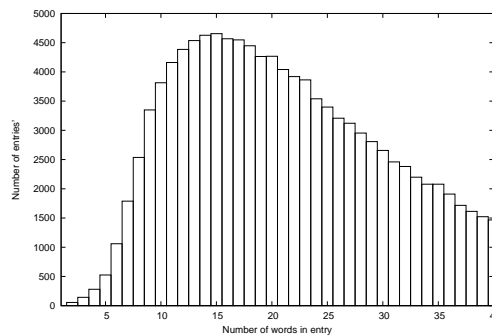


Figure 3: A histogram of sentence counts (y axis) vs. the sentence lengths (x axis).

Class	Description	Example inputs
ABBREVIATION	Abbreviations	Dr., Mr., Ms.
ADDRESS	Address expressions	34 Main st.
CARDINAL	Normal numbers	3479, 90,581, ၁၀
CONNECTOR	Ranges, ratios	1996-1997
DATE	Date expressions	3/5/2018, 2018-01-02
DECIMAL	Decimal point numbers	234.79
DIGIT	Digit sequences	123-4, ၀၀၈
ELECTRONIC	Email addresses and URLs	hello@test.org
EMOTICON	Emoticons/Emojis	:-), 8-)
FRACTION	Mathematical fraction	$\frac{2}{5}$, ၁/၁၂၅
LSEQ	Letter sequences	UN, အ.လ.က
MEASURE	Unit quantities	10 km, 30 sq.m.
MONEY	Currency quantities	\$2.5, 1.2ကျပ်
TELEPHONE	Phone numbers	+(94) 123 4567
TIME	Time expressions	၁၂:၄၅, 12:45, 4.33pm
VERBATIM	Special symbol names	$\Delta X + \Delta Y$

Table 3: Semiotic classes for Burmese.

missing sentence-final punctuation. Overall, our segmentation corpus consists of 110,947 segmented entries. These entries consist of 2,322,084 segments after segmentation, which amount to 59,312 unique words in our corpus. The entries vary in length, ranging from 2 words to 40 words, with an average of 20.93 words per sentence. Figure 3 shows a histogram of the length distribution of these entries.

Models To gauge the performance of popular segmentation methods using the collected data, we evaluated two standard approaches to segmentation: conditional random fields (CRF) (Lafferty et al., 2001), popular for Burmese (Pa et al., 2015; Ding et al., 2016; Phyu and Hashimoto, 2017; Ma and Yang, 2018), and feed-forward neural networks (NN) (Botha et al., 2017). Feed forward neural networks are particularly suited for low-resource platforms, such as mobile devices, due to their relatively low footprint and computational complexity. We trained both types of models after splitting the corpus into a training set (95%) and a test set (5%) using an ad-hoc partition. No cross-validation was performed. The performance of both algorithms on the test set is shown in Table 2. We chose the CRF as the better performing model to be included in our TTS pipeline.

4.2. Text Normalization

A popular approach to designing text normalization pipelines is to form it with two main components: a classifier and a verbalizer (Ebden and Sproat, 2015). These modules occupy the second stage in our pipeline (Figure 2) following the word segmenter. Our Thrax grammars adhere to this approach (Google, 2018b).

Classified NSW markup	Verbalization
cardinal integer:-19	အနုတ် ဆယ့် ကိုး
connector type:range	မှ
date day:4 month:၅ style:2	လေး ဇွန် လ
decimal integer_part:4 fractional_part:5	လေး ဒသမ ငါး
digit digit:9	ကိုး
concept concept:0_o	စိတ်ရှက်မျက်နှာ
fraction numerator:4 denominator:5	ပိုင်းဝေ လေး ပိုင်းခြေ ငါး
letters letters:အလက	အ လ က
measure integer:7 units:second	ခုနစ် စက္ကန့်
money fractional_part:98 currency:gbp	ကိုး ဆယ့် ရှစ် ပဲနီ
time hours:4 minutes:0	လေး နာရီ
verbatim verbatim:✓	နှစ်ထပ်ကိန်းရင်း

Table 4: Some example verbalizations.

```

1 # Download Google Language Resources repository.
2 git clone https://github.com/google/language-resources.git
3 cd language-resources
4 # Build classification grammars and tests.
5 bazel build //my/textnorm/classifier/...
6 # Build verbalization grammars and tests.
7 bazel build //my/textnorm/verbalizer/...
8 # Run the unit tests.
9 bazel test //my/textnorm/...

```

Table 5: Setting up text normalization.

The classifier component identifies and classifies the non-standard words (NSW), a candidate set whose elements belong to out-of-vocabulary tokens, in the segmented input text (Sproat et al., 2001). Each NSW belongs to a certain *semiotic class* (Taylor, 2009). These classes are often organized into taxonomies which aim to cover most of the practical use cases (Sproat et al., 2001; Ebden and Sproat, 2015; van Esch and Sproat, 2017). The semiotic classes supported by our grammars are shown in Table 3 along with the example inputs. Note that the classifier accepts both English and Burmese expressions³. Once classified, the NSW is converted to a simple structured markup. The verbalizer component is responsible for converting the NSW markup from a particular class to its corresponding word spelling, as shown in Table 4.

Prerequisites The grammars⁴ reside in the Google Internationalization Language Resources repository, along with other resources for low-resource languages (Google, 2016). The Bazel build system (Google, 2019) is required to compile the grammars which depend on Thrax (Roark et al., 2012). Bazel is a flexible build system that is able to pull further dependencies of Thrax, such as OpenFst finite-state transducer framework (Allauzen et al., 2007), from their respective remote repositories. The sequence of steps involving downloading the repository (line 2), compiling the Burmese grammars (lines 5 and 7) and running the unit tests (line 9) is shown in Table 5.

Using the Grammars The grammars compile into finite state archive (FAR) files which contain collections of rules expressed as weighted FSTs (Tai et al., 2011), each corresponding to a particular semiotic class. Overall there

³At the moment the classifier only supports inputs in English for addresses, abbreviations and measures.

⁴Available at <https://github.com/google/language-resources/tree/master/my/textnorm>.

```

1 # Compile Thrax driver.
2 bazel build @thrax//:thraxrewrite-tester
3 # Check classification.
4 bazel-bin/external/thrax/thraxrewrite-tester \
5 --far=bazel-bin/my/textnorm/classifier/classify.far \
6 --rules=CLASSIFY
7 ...
8 # Check verbalization.
9 bazel-bin/external/thrax/thraxrewrite-tester \
10 --far=bazel-bin/my/textnorm/verbalizer/verbalize.far \
11 --rules=ALL
12 ...

```

Table 6: Command-line use of text normalization.

	Voiced	Labial	Dental	Alveolar	Palatal	Velar	Glottal
Stop	✓	p p ^h b		t t ^h d		k k ^h g	ʔ
Affricate	✓				tʃ tʃ ^h dʒ		
Fricative	✓		θ ð	s s ^h z	ʃ		h
Nasal	✓	ɱ m		ɲ n	ɲ ɲ	ŋ	
Liquid	✓			l l̥			
Glide	✓	w w			j		

Table 7: Burmese consonants.

are 5 Burmese-specific individual classification grammars (e.g., `date.far`) and further 10 prebuilt language-agnostic grammars available in `universal_rules.far`. The grammars rely on the language-agnostic cardinal and ordinal number grammar in `number_names_rules.far`. All the grammars are combined into a single master classification grammar `classify.far` exposing the main rule (FST) `CLASSIFY` for performing the classification. The verbalizer component contains 14 Burmese-specific grammars in individual FAR files, combined into a single master verbalization grammar in `verbalize.far` that exposes the main verbalization rule `ALL`. Table 6 shows the use of the Thrax command-line driver for verifying individual rewrites for the classifier and verbalizer grammars using an interactive prompt.

System Integration The text normalization grammars described above can be integrated into a broader TTS pipeline using the Sparrowhawk text normalization framework (Google, 2015) which is an open-source version of Google text normalization (Ebden and Sproat, 2015). While we have not integrated Sparrowhawk with Burmese, the integration is pretty straightforward and can be based on existing integrations for other low-resource languages (such as Sinhala and Khmer⁵) described by Sodimana et al. (2018) that reside in the same language resource repository.

4.3. Phoneme Inventory

The representation of thirty-four Burmese consonants is shown in Table 7. The voiced liquid /l/, sometimes denoted /r/, is rare and serves as an optional variant of a palatal

⁵Available at <https://github.com/google/language-resources/tree/master/km/sparrowhawk>.

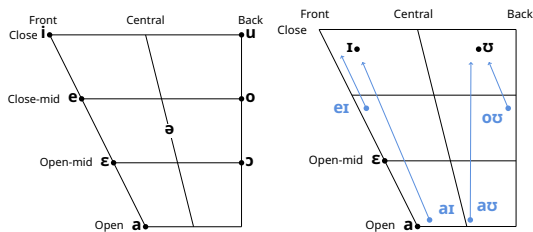


Figure 4: Chart of Burmese monophthong and diphthong distributions for orthographically open (left) and closed (right) syllables.

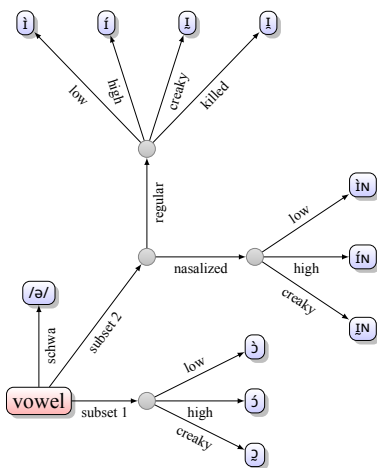


Figure 5: Very simplified depiction of Burmese tones.

glide /j/ in some loanwords of Pali origin (Watkins, 2001) or in modern loanwords (Chang, 2009; Gruber, 2011). The glide /w/ is rare, along with the voiced fricative /ð/ (Green, 2005). The main difference between our representation and the literature is our choice of affricates: voiceless alveolo-palatals (/tʃ/ and /tʃʰ/) instead of the more standard voiceless postalveolars (/tʃ/ and /tʃʰ/), and voiced alveolo-palatal /dʒ/ instead of voiced postalveolar /dʒ/. Although not part of the consonant inventory, a special symbol /N/, a placeless nasal, is often used to denote either a nasal consonant or simply a nasal quality on the vowel syllabic nucleus (Green, 2005; Gruber, 2011).

Our choice of vowel inventory follows the asymmetrical distribution of vowels provided by Watkins (2000). We divide the inventory into two sets. The first set, shown on the left-hand side of Figure 4, contains vowels that occur in orthographically open syllables and are not nasalized (Watkins, 2001). This set consists of eight vowels that include the schwa (/ə/). Some representations do not assign the schwa phonemic status because it occurs as an allophone (Chang, 2009). The second set, shown on the right-hand side of Figure 4, occurs either as nasalized or in the syllables closed by a glottal stop. This set consists of four vowels, two of which (/ɛ/ and /a/) are shared with the first set, although /ɛ/ can only appear in non-nasalized syllables. Burmese has four diphthongs (shown in lighter, blue, color) which also belong to the second set. The phonotactics require that these must be closed by either a glottal stop or a placeless nasal (Green, 2005).

As mentioned in Section 3, Burmese has four tones: *low*

```

1 # Download Google Language Resources repository.
2 git clone https://github.com/google/language-resources.git
3 cd language-resources
4 # Build G2P grammars and helper tool.
5 bazel build -c opt my:g2p
6 # Run G2P on command line.
7 echo "ကော့" | bazel-bin/my/g2p
8 → k á

```

Table 8: Downloading and building G2P.





(/à/), *high* (/á/), *creaky* (/ǎ/) and *checked* (/ǎ/) (Green, 2005). The checked tone is sometimes referred to as *killed* (Watkins, 2000) or *glottal* (Chang, 2009). Burmese tones are a complex phenomena that can be explained only with reference to several aspects of the language’s phonological structure, including the features of pitch, phonation type and vowel quality, among others (Gruber, 2011). According to Watkins (2001), the *low*, *high* and *creaky* tones may be described in terms of modification of the vowel feature alone. The *killed* tone modifies vowels from the second subset, always includes a syllable-final glottal stop and is not possible in syllables with nasal rhymes (Green, 2005). A very simplified diagram of this tonal system is shown in Figure 5. Because the tones potentially transform each vowel into up to four phonemes, we end up with the phoneme inventory consisting of 88 symbols: 34 consonants and 54 vowels.

4.4. Grapheme-to-Phoneme Conversion

Similar to text normalization, the G2P conversion grammar (Google, 2018a) resides in the Google Internationalization Language Resources repository (Google, 2016) and can be built using Bazel. The grammar (*burmese.grm*) is based on the Thrax grammar language and compiles into a weighted finite-state transducer with graphemes on the input tape and phonemes on the output tape. The grapheme and phoneme alphabets are specified in the *grapheme.syms* and *phoneme.syms* symbol table files collocated with the grammar file. The phoneme alphabet contains the phoneme inventory described in Section 4.3.

The simple sequence of commands required to fetch, compile and use the G2P grammar on the command-line is shown in Table 8. The grammar compiles into a weighted FST (in OpenFst format) with 195 states and 2966 arcs, with input and output cycles (139 input and 263 output epsilons, respectively). The FST accepts Unicode characters from the Myanmar block as defined in Unicode Standard (The Unicode Consortium, 2015) and outputs phonemes represented in IPA.

The transducer complexity relates to the non-trivial structure of the corresponding grammar which involves composition of several stages of processing which include

- Orthographic processing that involves re-ordering of medial consonant markers, if required, unifying variant spellings, normalizing the vowels (e.g., “” → “”) and unstacking complex stacked consonants (e.g., *kinzi*: ).
- Core G2P rewrites, such as turning the historical palatal stops into alveolar fricatives (e.g., “” → /sʰ/).

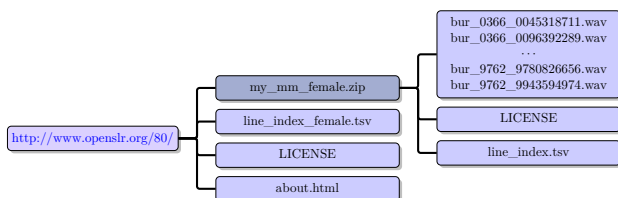


Figure 6: Structure of the Burmese speech corpus.

- Transformations that depend on the position in a syllable, i.e. are coda or nucleus-specific (e.g., handling nucleus-final “ သ် ”).
- Phonological transformations, such as turning palatal affricates into allophones of the velars (e.g., $/k^h j/ \rightarrow /t^c^h/$).
- Pronunciation exceptions that arise from various graphemic combinations, handling the special markers (such as locative “ ၌ ” and possessive “ ၏ ”), and so on.

5. Overview of the Corpus

In this section we provide an overview of the free multi-speaker Burmese speech dataset that we built (Burmese Corpus, 2019).

Distribution and Licensing The corpus is open-sourced under “Creative Commons Attribution-ShareAlike” (CC BY-SA 4.0) license (Creative Commons, 2019) and hosted on the Open Speech and Language Resources (OpenSLR) repository (Povey, 2019). The OpenSLR identifier for the corpus is SLR80, the International Standard Language Resource Number (ISLNR) (Mapelli et al., 2016) is 999-939-436-742-0⁶.

The corpus structure is shown in Figure 6. Collections of audio and the corresponding transcriptions are stored in a separate compressed archive for each gender. Only the female recordings are released at present. Transcriptions are stored in a *line index* file, which contains a tab-separated list of pairs consisting of the audio file names and the corresponding hand-curated transcriptions that were segmented at the word level. The name of each utterance consists of three parts: the symbolic dataset name (bur), the four-digit speaker ID and the 10-digit hash. The 48 kHz single-channel audio files are provided in 16 bit linear PCM RIFF format.

The Recording Process The speakers were all volunteer participants aged between 25 and 35. The recording took place in Yangon, Myanmar, in a rented studio. Using multiple speakers for the recording allowed us to obtain more data without putting too much burden on each of the volunteers, who were not professional voice talents. All the speakers were native speakers of Burmese. We recorded the audio with an ASUS Zenbook UX305CA fanless laptop, a Neumann KM 184 microphone and a Blue Icicle XLR-USB A/D converter.

The audio was recorded using our web-based recording software. Each speaker was assigned about 150–200 sentences to read. The tool recorded each sentence at 48 kHz (16 bits per sample). We also used in-house software for quality control where reviewers could check the recording

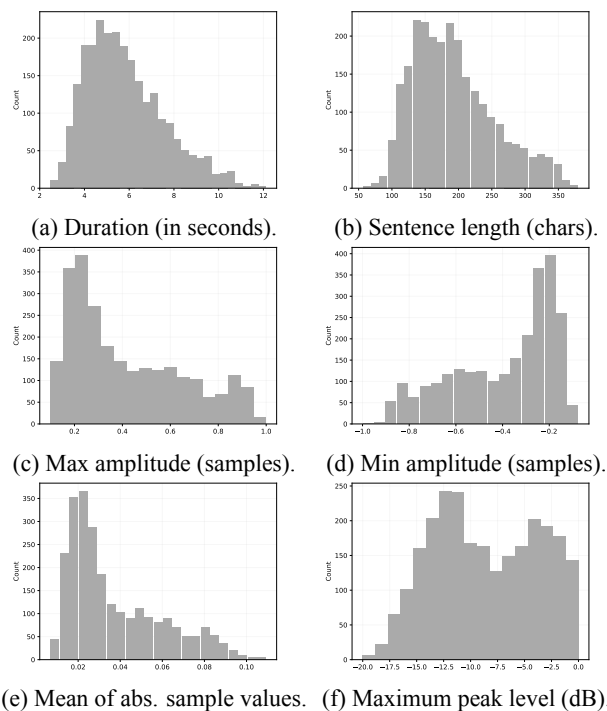


Figure 7: Speech corpus properties.

against the recording script and provide additional comments when necessary.

Since the speakers were not professional voice talents, their recordings could contain problematic artifacts such as unexpected pauses, spurious sounds (like coughing or clearing the throat) and breathy speech. As a result, it was very important to conduct quality control (QC) of the recorded audio data. All recordings went through a quality control process performed by trained native speakers to ensure that each recording (1) matched the corresponding script (2) had consistent volume (3) was noise-free (free of background noise, mouth clicks, and breathing sounds) and (4) consisted of fluent speech without unnatural pauses or mispronunciations. The reviewers could use the QC tool to edit the transcriptions to match the recording. Utterances not meeting the above criteria were discarded.

Basic Statistics The corpus consists of 2,528 utterances from 20 female speakers with the corresponding manually segmented transcriptions. Transcriptions contain 9,941 unique words and 22,443 words in total. Some of the basic properties of the corpus are shown in Figure 7. The durations of the audio utterances (a) are between 2.5 and 12 seconds, with the majority of the utterances having durations between 4 to 7.5 seconds. The sentence lengths (b) are between 55 and 360 Unicode characters, the majority of the sentences have between 120 and 230 characters. The arithmetic means of the absolute values of the audio samples (e) for the majority of the utterances are between 0.01 and 0.03. These values indicate the presence of a small direct current (DC) component in the signals, also known as DC offset, which can be removed using signal processing techniques (Gu and Yu, 2000; Karimi-Ghartemani et al., 2011). We also measured the maximum peak levels for the recordings (f) to ensure that the audio does not clip and to gauge the loudness levels. The spread in the maximum peak level

⁶Available at <http://www.openslr.org/80/>.

values as well as minimum and maximum amplitude values in subfigures (c) and (d) reflects the fact that the dataset has multiple speakers and has been recorded under varying conditions.

6. Experiments

Model Architecture Details We used the long short term memory recurrent neural network (LSTM-RNN) acoustic model configuration originally proposed by Zen and Sak (2015) to build an acoustic model using the corpus described in the previous section. LSTM-RNNs are designed to model temporal sequences and long-term dependencies within them (Hochreiter and Schmidhuber, 1997). To determine the phoneme time boundaries, prior to training the models the audio was force-aligned with the corresponding transcriptions (Young et al., 2006).

Two unidirectional LSTM-RNNs for duration and acoustic parameter prediction are used in tandem in a streaming fashion. Given the input features, the goal of the duration LSTM-RNN is to predict the duration (in frames) of the phoneme in question. This prediction, together with the input features, is then provided to the acoustic model which predicts smooth acoustic vocoder parameter trajectories. The smoothing of transitions between consecutive acoustic frames is achieved in the acoustic model by using recurrent units in the output layer.

The input features used by both the duration and the acoustic models consist of one-hot linguistic features that describe the utterance including the phonemes, syllable counts, distinctive features and so on. An additional important feature that we use is a one-hot speaker identity feature. When using a model trained on multiple speakers, this feature is instrumental in forcing the consistent speaker characteristics on the output of the model. In other words, it forces the voice to sound like the requested speaker.

The original audio was downsampled to 24 kHz. Then mel-cepstral coefficients (Fukada et al., 1992), logarithmic fundamental frequency ($\log F_0$) values (interpolated in the unvoiced regions), voiced/unvoiced decision (boolean value) (Yu and Young, 2011), and 7-band aperiodicities were extracted every 5 ms, similar to (Zen et al., 2016). These values form the output features for the acoustic LSTM-RNN and serve as input vocoder parameters (Agiomyrghiannakis, 2015). The output features for the duration LSTM-RNN are phoneme durations (in seconds). The input features for both the duration and the acoustic LSTM-RNN are linguistic features. Both the input and output features were normalized to zero mean and unit variance. At synthesis time, the acoustic parameters were synthesized using the Vocaine vocoding algorithm (Agiomyrghiannakis, 2015).

The architecture of the acoustic LSTM-RNN consists of a 1×128 ReLU layer (Zeiler et al., 2013) followed by 3×160 -cell LSTM with recurrent projection (LSTMP) layers (Sak et al., 2014) with 64 recurrent projection units and a linear recurrent output layer (Zen and Sak, 2015). The architecture of the duration LSTM-RNN consists of a 1×128 ReLU layer followed by a single 160-cell LSTMP layer with a feed-forward output layer with linear activation. Acoustic and duration networks were trained using ϵ -contaminated

Gaussian loss function (Zen et al., 2016) with exponentially decaying learning rate of 2^{-6} and a batch size of 4.

Results and Discussion We chose five out-of-domain sentences in order to establish the best sounding speaker for the next stage of experiments. The best sounding speaker with identity 4409 (speaker identities are encoded in the audio file names described in Section 5) was chosen by consensus from five evaluators.

We then performed subjective evaluation of the voice resulting from applying the best speaker identity feature by Mean Opinion Score (MOS) testing (Streijl et al., 2016). A system pipeline includes all the open-source components and data described in this paper. A set of 100 sentences was used that are neither in the training data, nor in the five-sentence set used for best speaker selection. Each sentence was rated by five different native speakers who were asked to rate each utterance on a 5-point scale (1: worst, 5: best). The resulting MOS score (with the corresponding 95% confidence interval) is 3.63 ± 0.10 , which ranks as reasonable on MOS scale and is overall competitive with the results of similar MOS evaluations reported in (Thu et al., 2015c; Hlaing et al., 2018). It is interesting to note that our system outperforms a somewhat similar LSTM-RNN configuration but fares worse when compared to the hybrid LSTM-RNN configuration recently reported by Hlaing et al. (2019). In their tests these two configurations scored 3.26 ± 0.35 and 4.01 ± 0.20 MOS, respectively. The differences in performance can be attributed to many factors, such as quality of the data and phoneme forced aligner algorithms, neural network parameters, linguistic pipeline differences and so on.

7. Conclusions and Future Work

In this paper we presented three open-source components for building Burmese text-to-speech pipelines: a free multi-speaker dataset of Burmese speech with the corresponding transcriptions (to the best of our knowledge, this is the first such dataset open to all with no restrictions) and the open-source finite-state transducer grammars for performing text normalization and grapheme-to-phoneme conversion. We showed that by using this data and algorithms in a reasonably standard LSTM-RNN pipeline well suited for low-resource scenarios, the resulting model scores competitively when compared to other systems reported in the literature. We hope the corpus and the grammars described in this paper will contribute to the burgeoning field of Burmese speech and language research and help advance state-of-the-art for other significantly more low-resource languages of Tibeto-Burman family.

As part of future experiments, it will interesting to see our Burmese corpus combined with other language corpora from broader Sino-Tibetan family to train a multilingual Sino-Tibetan model. This will open up venues for interesting investigations, such as testing whether transfer learning improves certain aspects of the Burmese speech synthesis (such as tone quality).

8. Acknowledgments

The authors would like to thank Richard Sproat, Rob Clark and the anonymous reviewers for many useful suggestions.

9. Bibliographical References

- Agiomyrgiannakis, Y. (2015). VOCAINE the vocoder and applications in speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4230–4234, Brisbane, Australia, April. IEEE.
- Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.
- Arnaudo, D. (2019). Bridging the Deepest Digital Divides: A History and Survey of Digital Media in Myanmar. In Aswin Punathambekar et al., editors, *Global Digital Cultures: Perspectives from South Asia*, pages 96–125. University of Michigan Press.
- Baljekar, P. (2018). *Speech Synthesis from Found Data*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
- Bisani, M. and Ney, H. (2008). Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech communication*, 50(5):434–451.
- Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.
- Bornás, A. J. and Mateos, G. G. (2019). A Character-Level Approach to the Text Normalization Problem Based on a New Causal Encoder. *arXiv preprint arXiv:1903.02642*.
- Botha, J. A., Pitler, E., Ma, J., Bakalov, A., Salcianu, A., Weiss, D., McDonald, R., and Petrov, S. (2017). Natural Language Processing with Small Feed-forward Networks. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2879–2885, Copenhagen, Denmark.
- Chang, C. B. (2009). English loanword adaptation in Burmese. *Journal of the Southeast Asian Linguistics Society*, 1:77–94.
- Chen, Y.-J., Tu, T., Yeh, C., and Lee, H.-Y. (2019). End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech 2019*, pages 2075–2079, Graz, Austria.
- Chit, Y. W. and Khaing, S. S. (2018). Myanmar Continuous Speech Recognition System Using Fuzzy Logic Classification in Speech Segmentation. In *Proc. of the Intl. Conference on Intelligent Information Technology*, pages 14–17, Hanoi, Vietnam. ACM.
- Cooper, E. L. (2019). *Text-to-Speech Synthesis Using Found Data for Low-Resource Languages*. Ph.D. thesis, Columbia University, New York.
- Creative Commons. (2019). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). <http://creativecommons.org/licenses/by-sa/4.0/deed.en>.
- Ding, C., Thu, Y. K., Utiyama, M., and Sumita, E. (2016). Word Segmentation for Burmese (Myanmar). *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 15(4):22.
- Ebden, P. and Sproat, R. (2015). The Kestrel TTS text normalization system. *Natural Language Engineering*, 21:333–353.
- Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 137–140. IEEE.
- Gokcen, A., Zhang, H., and Sproat, R. (2019). Dual Encoder Classifier Models as Constraints in Neural Text Normalization. pages 4489–4493, Graz, Austria.
- Goldhahn, D., Sumalvico, M., and Quasthoff, U. (2016). Corpus collection for under-resourced languages with more than one million speakers. In *Proc. of Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (CCURL)*, pages 67–73.
- Google. (2015). Sparrowhawk Text Normalization System. <https://github.com/google/sparrowhawk>.
- Google. (2016). Google Internationalization Language Resources. <https://github.com/google/language-resources>.
- Google. (2018a). Finite-state grammars for Burmese grapheme-to-phoneme conversion. <https://github.com/google/language-resources/my/burmese.grm>.
- Google. (2018b). Text normalization finite-state transducer grammars for Burmese. <https://github.com/google/language-resources/my/textnorm>.
- Google. (2019). Bazel. <http://bazel.build>. [Online], Accessed: 2019-10-2.
- Green, A. D. (2005). Word, foot, and syllable structure in Burmese. *Studies in Burmese linguistics*, 570:1–24.
- Gruber, J. F. (2011). *An articulatory, acoustic, and auditory study of Burmese tone*. Ph.D. thesis, Georgetown University.
- Gu, J.-C. and Yu, S.-L. (2000). Removal of DC offset in current and voltage signals using a novel Fourier filter algorithm. *IEEE Transactions on Power Delivery*, 15(1):73–79.
- Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. In *10th edition of the Language Resources and Evaluation Conference (LREC)*, pages 2005–2010, Portorož, Slovenia, May.
- Hlaing, C. S. and Thida, A. (2018). Phoneme based Myanmar text to speech system. *International Journal of Advanced Computer Research*, 8(34):47–58.
- Hlaing, A. M., Pa, W. P., and Thu, Y. K. (2017). Myanmar Number Normalization for Text-to-Speech. In *Proc. of 15th International Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–274, Yangon, Myanmar, August. Springer.
- Hlaing, A. M., Pa, W. P., and Thu, Y. K. (2018). DNN Based Myanmar Speech Synthesis. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 142–146, Gurugram, India.
- Hlaing, A. M., Pa, W. P., and Thu, Y. K. (2019). Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN. In *Proc. 10th ISCA Speech Synthesis Workshop (SSW)*, pages 189–193, Vienna, Austria.

- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Hosken, M. (2007). Representing Myanmar in Unicode: Details and Examples. Version 3, SIL International and Payap University Linguistics Institute, Chiang Mai, Thailand.
- International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- Jenny, M. and Tun, S. S. H. (2016). *Burmese: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Routledge.
- Karimi-Ghartemani, M., Khajehoddin, S. A., Jain, P. K., Bakhshai, A., and Mojiri, M. (2011). Addressing DC component in PLL and notch filter algorithms. *IEEE Transactions on Power Electronics*, 27(1):78–86.
- Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S. (2003). Creating corpora for speech-to-speech translation. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech)*, pages 381–384, Geneva, Switzerland.
- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Li, B. and Zen, H. (2016). Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis. In *Proc. of Interspeech*, pages 2468–2472, San Francisco.
- Liao, H.-T. (2017). Encoding for Access: How Zawgyi Success Impedes Full Participation in Digital Myanmar. *SIGCAS Comput. Soc.*, 46(4):18–24, January.
- Ma, C. and Yang, J. (2018). Burmese Word Segmentation Method and Implementation Based on CRF. In *Proc. of 2018 International Conference on Asian Language Processing (IALP)*, pages 340–344, Bandung, Indonesia.
- Mansfield, C., Sun, M., Liu, Y., Gandhe, A., and Hoffmeister, B. (2019). Neural Text Normalization with Subword Units. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 190–196.
- Mapelli, V., Popescu, V., Liu, L., and Choukri, K. (2016). Language Resource Citation: the ISLRN Dissemination and Further Developments. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1610–1613, Portorož, Slovenia, May. ELRA.
- Mon, A. N., Pa, W. P., and Thu, Y. K. (2017). Building HMM-SGMM continuous automatic speech recognition on Myanmar Web news. Dubai, United Arab Emirates. Proc. of 15th International Conference on Computer Applications (ICCA 2017).
- Mon, A. N., Pa, W. P., and Thu, Y. K. (2019). UCSY-SC1: A Myanmar Speech Corpus for Automatic Speech Recognition. *International Journal of Electrical & Computer Engineering*, 9(4):3194–3202.
- Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epi-tran: Precision G2P for many languages. In *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2710–2714, 7-12 May 2018, Miyazaki, Japan.
- Nachmani, E. and Wolf, L. (2019). Unsupervised Polyglot Text-to-Speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7055–7059. IEEE.
- Ng, A. H., Gorman, K., and Sproat, R. (2017). Minimally Supervised Written-to-spoken Text Normalization. In *Proc. of 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 665–670, Okinawa, Japan. IEEE.
- Nikulásdóttir, A. B. and Guðnason, J. (2019). Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case. pages 4455–4459, Graz, Austria.
- Novak, J. R., Dixon, P. R., Minematsu, N., Hirose, K., Hori, C., and Kashioka, H. (2012). Improving WFST-based G2P Conversion with Alignment Constraints and RNNLM N-best Rescoring. In *Proc. Interspeech*, pages 2526–2529, Portland, USA.
- Nyunt, U. B., Gyi, U. H., Kyan, D., Htun, T., Kaung, U. T., Ha, U. T., Shwe, H., Fatt, H., Than, D. M., Htut, T., Pe, W., Mae, H., Yin, H., et al. (1993). Myanmar-English Dictionary. *Myanmar Language Commission, Ministry of Education, Yangon, Myanmar*.
- Okell, J., Tun, U., and Swe, D. K. M. (2010). *Burmese (Myanmar): An Introduction to the Script*. Northern Illinois University Press.
- Pa, W. P., Thu, Y. K., Finch, A., and Sumita, E. (2015). Word Boundary Identification for Myanmar Text Using Conditional Random Fields. In *Proc. 9th International Conference on Genetic and Evolutionary Computing (GEC)*, pages 447–456, Yangon, Myanmar. Springer.
- Phyu, M. L. and Hashimoto, K. (2017). Burmese word segmentation with Character Clustering and CRFs. In *Proc. of 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6, Nakhon Si Thammarat, Thailand. IEEE.
- Povey, D. (2019). Open SLR. <http://www.openslr.org/resources.php>. Accessed: 2019-03-30.
- Prakash, A., Thomas, A. L., Umesh, S., and Murthy, H. A. (2019). Building Multilingual End-to-End Speech Synthesizers for Indian Languages. In *Proc. of 10th ISCA Speech Synthesis Workshop (SSW'10)*, pages 194–199, Vienna, Austria.
- Ritchie, S., Sproat, R., Gorman, K., van Esch, D., Schallhart, C., Bampounis, N., Brard, B., Mortensen, J. F., Holt, M., and Mahon, E. (2019). Unified Verbalization for Speech Recognition & Synthesis Across Languages. pages 3530–3534, Graz, Austria.
- Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J., and Tai, T. (2012). The OpenGrm Open-source Finite-state Grammar Software Libraries. In *Proc. of the Association for Computational Linguistics (ACL) 2012 Sys-*

- tem Demonstrations, ACL '12, pages 61–66, Stroudsburg, PA, USA. ACL.
- Roop, D. H. (1972). *An introduction to the Burmese writing system*. Yale Linguistic Series. Yale University Press, New Haven.
- Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. of Interspeech*, pages 338–342, Singapore, September. ISCA.
- SIL International. (2019). Ethnologue. <https://www.ethnologue.com>. Accessed: 2019-03-25.
- Sodimana, K., De Silva, P., Sproat, R., Wattanavekin, T., Gutkin, A., and Pipatsrisawat, K. (2018). Text Normalization for Bangla, Khmer, Nepali, Javanese, Sinhala and Sundanese Text-to-Speech Systems. In *Proc. of 6th Intl. Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, pages 147–151, Gurugram, India.
- Soe, E. P. P. and Thida, A. (2013a). Diphone-Concatenation Speech Synthesis for Myanmar Language. *International Journal of Science, Engineering and Technology Research (IJSETR)*, 2(5):1078–1087.
- Soe, E. P. P. and Thida, A. (2013b). Text-to-speech synthesis for Myanmar language. *International Journal of Scientific & Engineering Research*, 4(6):1509–1518.
- Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of Non-Standard Words. *Computer Speech and Language*, 15(3):287–333, July.
- Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.
- Tai, T., Skut, W., and Sproat, R. (2011). Thrax: An Open Source Grammar Compiler Built on OpenFst. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, volume 12, Waikoloa Resort, Hawaii.
- Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.
- The Unicode Consortium. (2015). Unicode standard, version 8.0.0. Mountain View, CA, <http://www.unicode.org/versions/Unicode8.0.0>.
- Thu, Y. K., Pa, W. P., Finch, A., Hlaing, A. M., Naing, H. M. S., Sumita, E., and Hori, C. (2015a). Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion. In *Proc. The 13th International Conference on Computer Applications (ICCA2015)*, pages 161–167.
- Thu, Y. K., Pa, W. P., Finch, A., Ni, J., Sumita, E., and Hori, C. (2015b). The Application of Phrase Based Statistical Machine Translation Techniques to Myanmar Grapheme to Phoneme Conversion. In *Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 238–250. Springer.
- Thu, Y. K., Pa, W. P., Ni, J., Shiga, Y., Finch, A. M., Hori, C., Kawai, H., and Sumita, E. (2015c). HMM Based Myanmar Text to Speech System. In *Proc. of Interspeech*, pages 2237–2241, Dresden, Germany. ISCA.
- Thu, Y. K., Pa, W. P., Sagisaka, Y., and Iwahashi, N. (2016). Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary. In *Proc. of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 11–22, Osaka, Japan.
- van Esch, D. and Sproat, R. (2017). An Expanded Taxonomy of Semiotic Classes for Text Normalization. In *Proc. of Interspeech*, pages 4016–4020. Stockholm, Sweden.
- Watkins, J. (2000). Notes on creaky and killed tone in Burmese. *School of Oriental and African Studies (SOAS) Working papers in Linguistics and Phonetics*, 10:139–149.
- Watkins, J. W. (2001). Illustrations of the IPA: Burmese. *Journal of the International Phonetic Association*, 31(2):291–295.
- Wheatley, J. K. (1987). Burmese. In Bernard Comrie, editor, *The World's Major Languages*, pages 834–854. Oxford University Press, New York.
- Wheatley, J. K. (1990). Burmese. In Bernard Comrie, editor, *The Major Languages of South-East Asia*, pages 106–126. Routledge, London.
- Wheatley, J. K. (1996). The World's Writing Systems. In Peter T Daniels et al., editors, *Burmese Writing*, pages 450–455. Oxford University Press.
- Wibawa, J. A. E., Sarin, S., Li, C. F., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., and Ha, L. (2018). Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1610–1614, 7-12 May 2018, Miyazaki, Japan.
- Win, K. Y. and Takara, T. (2011). Myanmar text-to-speech system with rule-based tone synthesis. *Acoustical Science and Technology*, 32(5):174–181.
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., G. M., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book*. Cambridge University Engineering Department.
- Yu, K. and Young, S. (2011). Continuous F0 modeling for HMM based statistical parametric speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1071–1079.
- Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P. c., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G. (2013). On rectified linear units for speech processing. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 3517–3521, Vancouver, Canada, May. IEEE.
- Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia, April. IEEE.
- Zen, H., Ajiomyrgiannakis, Y., Egberts, N., Henderson, F., and Szczepaniak, P. (2016). Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. In *Proc. Interspeech 2016*, pages 2273–2277, San Francisco, September.
- Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia,

- Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. of Interspeech 2019*, pages 1526–1530, Graz, Austria.
- Zhang, H., Sproat, R., Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., and Roark, B. (2019). Neural Models of Text Normalization for Speech Applications. *Computational Linguistics*, 45(2):293–337.

10. Language Resource References

- Burmese Corpus. (2019). *Crowd-sourced high-quality Burmese speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), <http://www.openslr.org/80>, Google crowd-sourced speech and language resources, 1.0, ISLRN 999-939-436-742-0.