Resources Creation of Bengali for SPPAS

Moumita Pakrashi, Brigitte Bigi, Shakuntala Mahanta

Resources Creation of Bengali for SPPAS

Moumita Pakrashi[1], Brigitte Bigi[2], Shakuntala Mahanta[3]

[1]Centre for Linguistic Science and Technology, IIT Guwahati, Assam, India.
moumi176155103@iitg.ac.in
[2]Laboratoire Parole et Langage, CNRS, Aix-Marseille Univ., Aix-en-Provence, France.
brigitte.bigi@cnrs.fr
[3]Department of Humanities and Social Sciences, IIT Guwahati, Assam, India.
smahanta@iitg.ac.in

Abstract

The development of HLT tools inevitably involves the need for language resources. However, only a handful number of languages possess such resources for free. This paper presents the development of speech tools for the Bengali language. Particularly, this paper focuses on developing language resources of a tokenizer, an automatic speech system for predicting the pronunciation of the words and their segmentation in this low-resourced language. The newly created resources have been integrated into SPPAS software tool and distributed under the terms of public licenses.

Keywords: Human Language Technology, Automated Annotation, Less-Resourced Languages, linguistic resources

1. Introduction

The development of Human Language Technologies (HLT) tools is a way to break down language barriers. Only a handful of the approximately 7,000 languages of the world possess the linguistic resources required for implementing HLT technologies (Bigi, 2014). One needs to analyse a considerable amount of speech dataset to achieve reliable and consistent results from phonetic research. Even in languages where a huge speech dataset is available for research, analysing and annotating the data is a primary challenge. The basic problem is that analysing data (in the context of this paper, speech data) is very tedious and time-consuming task, even for a phonetician or a trained expert. The next problem is that differences in research objectives necessitate their corresponding analysis and annotation. In such a situation, automatic annotation of speech data becomes a primary requirement of phonetic research.

Currently, a number of tools are available for the purpose of automatically segmenting and aligning speech data with its corresponding transcription. Speech recognition engines like the open-source CSR-Engine Julius (Lee et al., 2001) or the licensed HTK Toolkit (Young and Young, 1993) can perform the task. Even some wrappers for HTK such as WebMaus - Automatic Segmentation and Labelling of Speech Signals over the Web (Kisler et al., 2017), P2FA - the Penn Phonetics Lab Forced Aligner (Yuan et al., 2008), and others[1] can make the task easier. Most of these tools require varying amounts of expertise in computer science - particularly those that require installing HTK to be able to operate, or they are not available across multiple platforms. But the SPPAS tool runs on multiple platforms, and the incorporation of a new language requires some bare minimum linguistic resources that can be easily handled by linguists (Bigi, 2015). Apart from the automated functions of phonetization, annotation in multiple formats (such as X-SAMPA, IPA), alignment and syllabification; prosodic analysis of utterances can also be performed using the Momel-INTSINT algorithm (Hirst and Espesser, 1993; Hirst, 2011) incorporated within SPPAS.

Bengali is one of the dominant languages in the Indian subcontinent that historically belongs to the Indo-Aryan (IA) family of languages. Spoken by almost 210 million people as their native or second language, it currently holds the seventh position among the world's languages[2]. Bengali is one of the official languages of India and the national language of Bangladesh. It is spoken primarily in Bangladesh and the Eastern Indian states of West Bengal, Tripura, parts of Assam. In this paper, we deal with the Standard Colloquial Bengali (SCB) variety which is spoken mostly in and around Kolkata. The Bengali script or Bangla alphabet (Bengali: বাংলা বর্ণমালা baŋla bɔrnamala) is the alphabet used to write the Bengali language. This writing system of Bengali has its origins in the Brahmi script, which is the source of all modern scripts of Indian languages (Klaiman and Lahiri, 2018).

This paper describes the process of implementing automatic annotation and analysis of Bengali speech using SPPAS software (Bigi, 2015). The SPPAS tool produces automatic segmentation, annotations, and analysis of a speech sound and its corresponding orthographic transcription. With that objective in mind, this paper describes the development of a corpus and

[1]For a list, see https://github.com/pettarin/forced-alignment-tools

[2]https://www.britannica.com/topic/Bengali-language

some language resources for Bengali. Such newly created linguistic resources were integrated into SPPAS for the multi-lingual automatic tokenizer (Bigi, 2014), the multi-lingual automatic speech system for predicting the pronunciation of words (Bigi, 2016) and for their segmentation (Bigi and Meunier, 2018).

2. Corpus description

In order to use standardised speech data of Bengali for creating necessary linguistic resources, we have used an open-source speech database created by Google to develop Text to Speech (TTS) systems. The data consists of audio recordings of short phrases/sentences, a pronunciation lexicon, and a phonology definition of Bengali[3]. All the data have been released under the Creative Commons Attribution 4.0 international license (CC-By-NC-4.0).

This set of data contains audio-transcript pairs. Audio recordings are in WAVE format. The accompanying line index.tsv file has the normalized transcript of the recorded audio and the ID of the corresponding audio file. The audio data was collected from a group of volunteers between the ages of 20 and 35. They were asked to read short sentences, each containing 5 - 20 words. The texts used for recording have been either extracted from Wikipedia or general websites or are declarative sentences created by native speakers of the language. The recording was conducted in a quiet environment: either a sound studio or a quiet room with a soundproof booth. Moreover, all audio files have passed through a QC process to ensure good audio quality, absence of background noise, and match between recorded audio and text transcript.

In order to convert each entry to consist of an audio file and its corresponding text transcript file, we used the "Fill in IPUs" automatic annotation of SPPAS to automatically detect the IPUs of each file. The result is a file indicating the time alignment of each sounding segment. This automatic annotation wasn't verified.

The corpus duration is 7291.82 seconds (2 hours) among 1366 audio files.

3. Phonetic description of Bengali

Bengali is an Eastern Indo-Aryan language. Phonemically, Bengali features 29 consonants and 7 vowels. However, the phonological alternations of Bengali vary greatly due to dialectal differences.

Among the corpus described in the previous section, 76 files were manually time aligned at the phoneme level. It represents 211.64 seconds including 36.9 of silences. Tables 3., and 1 indicate the phonemes both in SAMPA (Wells, 1997) and in the International Phonetic Alphabet, an example of a word and the number of occurrences in the manually aligned corpus. In this small part of the corpus, we observed 29 consonants and 11 vowels. We have also added those phonemes that had nil occurrence in our corpus such as ɖ, ɖʰ and nasal vowels primarily because of their prominent presence in Bengali vocabulary.

SAMPA	IPA	Example	Occ.
b	b	ব্রাত্য (*outcast*)	76
b_h	bʰ	ভদ্র (*polite*)	14
c	c	চকচকে (*shiny*)	14
c_h	cʰ	ছবি (*picture*)	40
d	d	দর (*rate*)	64
d_h	dʰ	ধনী (*the rich*)	10
d‘	ɖ	ডিম (*egg*)	0
d‘_h	ɖʰ	ঢালু (*slope*)	0
g	g	গণতন্ত্র (*democracy*)	30
g_h	gʰ	ঘটনা (*incident*)	12
k	k	কুটির (*cottage*)	140
k_h	kʰ	খেজুর (*dates*)	50
p	p	পার্থক্য (*difference*)	48
p_h	pʰ	ফেরা (*to return*)	10
t	t	তালিকা (*list*)	88
t_h	tʰ	থালা (*plate*)	20
t̀	ʈ	টোকা (*to copy*)	36
t_̀h	ʈʰ	ঠেকানো (*to prevent*)	2
dZ	d͡ʒ	জলন্ত (*burning*)	54
dZ_h	d͡ʒʰ	ঝাণ্ডা (*flag*)	0
f	f	ফান্ড (*fund*)	2
h	h	হাত (*hand*)	29
s	s	শামুক (*snail*)	6
S	ʃ	সহকর্মী (*colleague*)	88
m	m	মাস (*month*)	52
n	n	নিবন্ধ (*essay*)	115
N	ŋ	ধ্বংস (*destruction*)	4
l	l	লিপি (*script*)	78
r	r	রহস্য (*suspense*)	184
ṛ	ɽ	পড়া (*to study*)	4
j	j	হৃদয় (*heart*)	18
w	w	হওয়া (*to happen*)	2

a	a	আদর্শ (*principle*)	263
a~	ã	আঁকা (*to draw*)	2
e	e	এবার (*now*)	270
e~	ẽ	এঁকেবেঁকে (*twisted*)	6
i	i	ইচ্ছা (*wish*)	195
i~	ĩ	ইঁদুর (*rat*)	2
O	ɔ	অংশ (*part*)	70
O~	ɔ̃	পঁচা (*rotten*)	0
o	o	ওজন (*weight*)	194
o~	õ	ওঁৎ (*trap*)	0
u	u	উত্তর (*answer*)	87
u~	ũ	উঁচু (*high*)	2
{	æ	এক (*one*)	16
@	ə	জংশন (*junction*)	0

Table 1: Consonants and vowels of Bengali

[3]https://github.com/google/language-resources/tree/master/bn

4. Creating resources for HLT tools

4.1. Vocabulary, Pronunciation dictionary

The vocabulary list contained approximately 65000 lexical entries of Bengali, including many loan words written in English orthography. The dictionary entries provide a broad phonemic transcription of colloquial Bengali but of the Bangladeshi Standard Bengali. Therefore we *manually corrected* each of the lexical items in the list to suit our required Standard Indian variety of Bengali speech. This created a pronunciation dictionary of Bengali corresponding to the Standard colloquial Bengali speech variety of India.

4.2. Acoustic model

Acoustic models are Hidden Markov models (HMMs) created using the HTK Toolkit (Young and Young, 1993), version 3.4. HMM states are modelled by Gaussian mixture densities whose parameters are estimated using an expectation-maximization procedure. Acoustic models were trained from 16 bits, 16,000 Hz wav files for the corpus. The Mel-frequency cepstrum coefficients (MFCC) along with their first and second derivatives were extracted from the speech in the standard way (MFCC_D_N_Z_0). See (Bigi, 2012) for details. The training procedure is implemented into a Python script included in SPPAS.

The outcome of the training procedure is dependent on both: 1/ the availability of accurately annotated data; and 2/ on good initialization. The initialization of the models creates a prototype for each phoneme using time-aligned data. In the context of this study, this training stage has been switched off: it has been replaced by the use of phoneme prototypes already available in some other languages. The articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent of the underlying language (Schultz and Waibel, 2001). In SPPAS package, 10 acoustic models of the same type - i.e. same HMMs definition and the same MFCC parameters, are freely distributed with a public license so that the phoneme prototypes can be extracted and reused: English, French, Italian, Spanish, Catalan, German, Polish, Mandarin Chinese, Southern Min, Naija. To create an initial model for the Bengali language, most of the prototypes of English language were used, nasal vowels from French language and some from Southern Min, and Polish. The prototypes of noise and laughter were also added to the model to be automatically time-aligned. This approach enabled the acoustic model to be trained with the small amount of Bengali language speech data we collected (Le et al., 2008; Bigi et al., 2021).

5. HLT tools

5.1. Automatic tokenization, phonetization and forced-alignment

In recent years, the SPPAS software tool has been developed to produce annotations automatically. It proposes 23 automatic annotations of audio or video, including the ones for the alignment of recorded speech sounds with its phonetic annotation. The multilingual approaches that are proposed enabled us to make the automatic annotations of SPPAS available for the Bengali language. For this purpose, we created an archive containing the lexicon - a list of words, the pronunciation dictionary in HTK-ASCII format and the acoustic model; and we made it available on the SPPAS website. We also added their description in the SPPAS resources documentation. Figure 1 shows an example of the resulting automatic Text Normalization, Phonetization and Alignment of a Bengali speech segment when using SPPAS 4.7.

5.2. Experiments

Forced-alignment is the task of automatically positioning a sequence of phonemes in relation to a corresponding continuous speech signal. Given a speech utterance along with its phonetic representation, the goal is to generate a time-alignment between the speech signal and the phonetic representation.

Some experiments were conducted to evaluate the accuracy of the phoneme alignments. It was evaluated using the Unit Boundary Positioning Accuracy - UBPA that consists in the evaluation of the delta-times (in percentage) comparing manual phonemes boundaries with the automatically aligned ones. This *Delta* time is estimated on the beginning of each phoneme as: $Delta = T(auto) - T(manual)$ When *Delta* is a positive value, $T(auto) > T(manual)$ means that the automatic boundary is "in late".

SPPAS can perform the alignment either with julius or with hvite. The UBPA we report in this paper is estimated while using Julius CSR engine (Lee and Kawahara, 2009) but the results were also estimated with HVite command of the HTK 3.4 toolkit (Young and Young, 1993). Both are very close, so there's no need in this paper to report them both.

The UBPA was first estimated with a model we didn't train on data. This model is created from the prototypes of the phonemes extracted from models of different other languages. It resulted in an UBPA of 88.82% within a delta of 0.04 seconds, e.g. the automatic boundary is not more than 40 milliseconds before or after the manual one.

Another model was trained by using the prototypes as bootstrap and the transcribed data only. The UBPA of such model is 93.22% within a delta of 0.04 seconds.

Finally, we trained a model by using the prototypes as bootstraps, the transcribed data, and the manually time-aligned data. In order to estimate the UBPA, we used a leave-one-out algorithm by training 79 models and testing it on the single test sentence that had been excluded from training. This resulted in an UBPA of **93.98%** within a delta of 0.04 seconds.

Figure 2 show details about the differences between automatic boundaries and manual ones for vowels (including nasal vowels with almost nil occurrence in our

Figure 1: Example of result of the automatic annotations of Bengali - SPPAS 4.7

corpus); while Figure 3 shows this difference in the case of Bengali consonants. As shown in Figures 2 and 3, the automatic system nearly preserves the manually annotated duration measurements. For example, it allows to see that the end position of the vowel /{/ is correct, but the automatic system detects it lately. However, we can't comment much on nasal vowels because of very few occurrences.

6. Conclusion

This paper presents free linguistic resources created for the Bengali language. These were useful for creating HLT tools for Text Normalization (including a tokenizer), Phonetization and Alignment of automatic annotations for Bengali. These resources have been made available freely since SPPAS version 4.1. Developing an automatic syllabification system for Bengali will be the future focus of our work in this software. It will be based on the existing system already available for French, Italian and Polish (Bigi and Klessa, 2015).

References

Bigi, B., 2012. The SPPAS participation to Evalita 2011. In *Evaluation of Natural Language and Speech Tool for Italian*, volume 7689 of *Lecture Notes in Artificial Intelligence*. Springer Berlin Heidelberg, pages 312–321.

Bigi, B., 2014. A multilingual text normalization approach. *HLT for Computer Science and Linguistics, LNAI 8387*:515–526.

Bigi, B., 2015. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111–112:54–69.

Bigi, B., 2016. A phonetization approach for the forced-alignment task in SPPAS. *Human Language Technology. Challenges for Computer Science and Linguistics, LNAI 9561*:515–526.

Bigi, B. and K. Klessa, 2015. Automatic Syllabification of Polish. In *Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznan, Poland.

Bigi, B. and C. Meunier, 2018. Automatic segmentation of spontaneous speech. *Revista de Estudos da Linguagem. International Thematic Issue: Speech Segmentation*, 26(4).

Bigi, B., A.-S. Oyelere, and B. Caron, 2021. Resources for automated speech segmentation of the african language naija (nigerian pidgin). *Human Language Technology. Challenges for Computer Science and Linguistics, LNAI 12598*:164–173.

Hirst, D.-J., 2011. The analysis by synthesis of speech melody: from data to models. *Journal of Speech Sciences*, 1(1):55–83.

Hirst, D.-J. and R. Espesser, 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15:71–85.

Kisler, T., Reichel U. D., and F. Schiel, 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Klaiman, M.H. and A. Lahiri, 2018. *Bengali*, chapter 23. Routledge, 3rd edition, pages 427–447.

Le, V.B., L. Besacier, S. Seng, B. Bigi, and T.N.D. Do, 2008. Recent advances in automatic speech recognition for vietnamese. In *International Workshop on Spoken Languages Technologies for Under-resourced languages*. Hanoi, Vietnam.

Lee, A. and T. Kawahara, 2009. Recent development of open-source speech recognition engine julius. In *Asia-Pacific Signal and Information Processing Association*. Annual Summit and Conference, International Organizing Committee.

Lee, A., T. Kawahara, and K. Shikano, 2001. Julius - an open source real-time large vocabulary recognition engine. In *European Conference on Speech Communication and Technology, EUROSPEECH*. Aalborg, Denmark.

Schultz, T. and A. Waibel, 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1):31–51.

Wells, J.C., 1997. Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.

Young, S-J. and Sj Young, 1993. *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering.

Yuan, J., M. Liberman, et al., 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

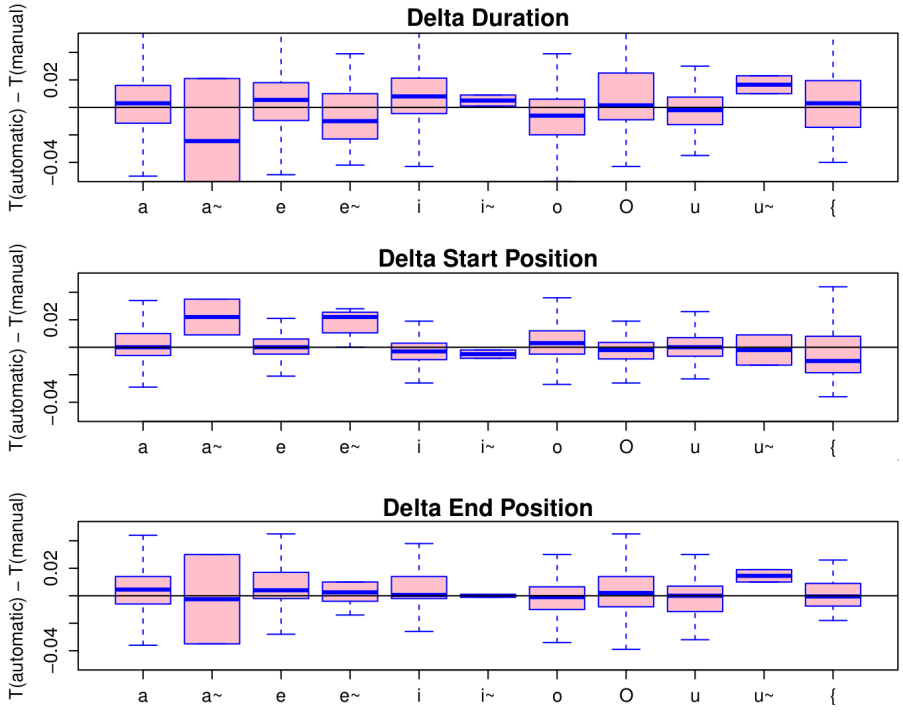

Figure 2: Detailed results of automatic alignment for vowels

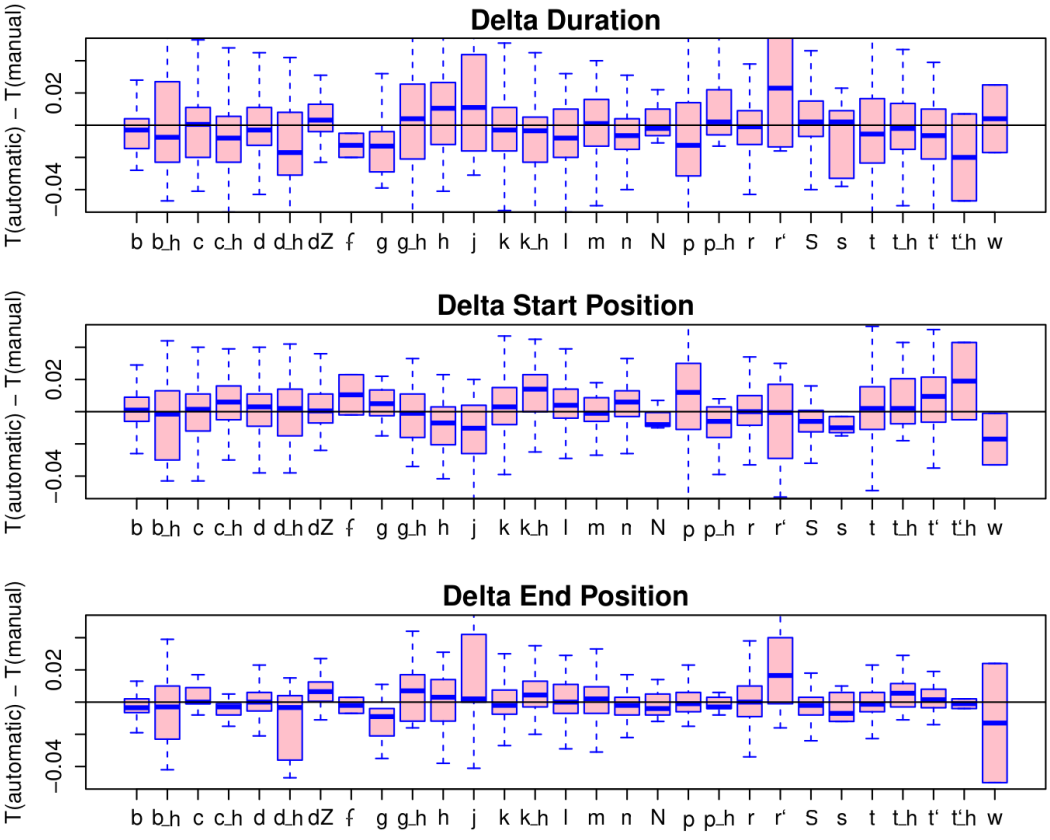

Figure 3: Detailed results of automatic alignment for consonants