The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Building an Automatic Speech Recognition System in Sora Language Using Data Collected for Acoustic Phonetic Studies

Kishalay Chakraborty[1], *Luke Horo*[2], *Priyankoo Sarmah*[2,3]

[1]DFM InfoAnalytics Private Limited, India
[2]Department of Humanities and Social Sciences
[3]Center for Linguistic Science and Technology
Indian Institute of Technology Guwahati, India

kishalaychakraborty1@gmail.com,{luke,priyankoo}@iitg.ac.in

Abstract

This paper reports the building of a limited vocabulary speech recognition system for Sora without a speech database designed specifically for automatic speech recognition (ASR). The system was built using speech data collected in field for acoustic phonetic analysis of the Sora language in Assam, India. As Sora is an under resourced language, the speech database is small and contains recordings of single words. Thus, the system is trained without a language model. There is no available ASR for Sora language and hence, this is the first attempt to build one for the language which may lead to the development of a more robust ASR for the language. The ASR shows better recognition rate with Subspace Gaussian Mixture Model (SGMM) and Deep Neural Network (DNN) frameworks. While the system performs with adequate accuracy, as expected, phonetically similar words are often misrecognized.

Index Terms: speech recognition, Sora language, under-resourced languages,

1. Introduction

The North-East Indian states of India are rich in linguistic and cultural diversity. More than 200 languages are spoken in this are that belong to several language families, such as the Tibeto-Burman, Indo-European and Austro-Asiatic. While, a diverse population in the area speak these languages, there are not sufficient documentation done for these languages. Needless to say, the amount of technology development in the languages of North East India is also minimal as almost all the languages spoken in the region are under-resource languages. Recently, there are attempts at building speech technology tools in a few languages of the area. For example, an ASR system built for agricultural commodity price inquiry using voice input in Assamese using HMM-GMM modeling achieved word error rate(WER) of approx 16% [1]. Another limited vocabulary Assamese ASR system built using Zero Crossing Rate (ZCR), Short Time Energy (STE), and Mel Frequency Cepstral Coefficient (MFCC) features obtained 99% and 93% accuracy for speaker dependent and speaker independent systems, respectively [2]. ASR system for Assamese numerals designed with Artificial Neural Network(ANN) obtained 90% accuracy [3]. A Phone Recognition System (PRS) built for Mizo, another under-resourced language, using HMM-GMM, SGMM and DNN shows better result in SGMM and DNN Compared to HMM-GMM approaches. The PRS showed an error rate of 13.9% for DNN using language model [4]. Recently, a digit recognition system built for Mizo, with augmented training data and trained with prosodix features, demonstrated an accuracy of 100% [5].

While the rate of speech technology development in the languages of North -East India is slow, there is an utmost need for it due to the challenging terrain and frequent natural disasters in the area. However, as with the under-resource languages, the challenge is the unavailability of speech data corpus in the languages. On the other hand, there are several linguistic documentation on the low resourced languages of the area, such as in the Tai Languages [6], Chokri [7], Sora [8] etc. Most of these works have collected speech data from several languages for acoustic phonetic descriptions of the languages. Hence, some resources are created for these under-resourced languages, even though they may not have been collected keeping automatic speech recognition as a final goal. Hence, the data collected for acoustic phonetic analysis have the following conditions:

- Limited amount of speech data
- Speech data collected in isolation
- Speech data collected in sentence frames
- Recorded in noiseless or quiet environment
- Data recorded with high sampling rates

These conditions mentioned above usually are not suitable for ASR development, as a robust ASR system demands large amount of speech data of connected speech collected in various noise conditions. However, for several low-resource or no-resource languages, such as Sora, the data collected for acoustic phonetic analysis by linguists is the only speech resource available. Considering such limitations, and considering the need for building ASR systems for low-resource language communities for last mile connectivity, it is imperative that an attempt is made to build such systems. This work details such an attempt to build the first ASR system for Sora from data collected for an acoustic phonetic analysis of the Sora language as part of a doctoral dissertation [8].

The paper is divided into the following parts, Section 2 gives an outline of Sora language, Section 3 details the organization and origin of the database. Section 4 describes the development of the ASR and the experiments conducted, Section 4 discusses the results and finally, Section 6 concludes the paper.

2. The Sora language

Sora belongs to the South Munda sub-branch of the Austroasiatic language family [9]. It is mainly spoken by the Sora tribe living in parts of Orissa and Andhra Pradesh in Eastern India and in Assam of North-East India [10]. It is spoken in Assam as some Sora speaking population migrated from Orissa

10.21437/SLTU.2018-50

to Assam in the 19th century as indentured tea labourers [11]. Currently, there are 5900 Sora individuals living in Assam [12]. Linguistic analysis of Assam Sora has suggested that the migrated language has adequately preserved its linguistic characteristics even after 100 years of its migration from Orissa and living surrounded by diverse linguistic groups [13, 8]. Significantly, while the Sora language of Orissa has some linguistic resources available, the Sora language of Assam does not have any available resources. However, as far as we know, speech data corpus is not available for for the language in any form.

Data in this work represents the Sora language as it is spoken in Assam. The Sora language of Assam exists only in oral form, though Some native speakers have attempted to write the language in Assamese and Roman scripts. However, the orthography is still not standardized. Also, linguistic description of the language in Assam is very limited. Extensive descriptions of Sora in Assam include works that mainly describe the phonetic properties of the language and are based on acoustic analysis [14, 8]. The following subsections summarize some of the major phonetic properties of the Sora language of Assam, drawing from previous studies.

2.1. Phoneme Inventory

Phoneme inventory of the Sora language of Assam includes 24 speech sounds. Table 1 presents a list of phonemes in Sora in International Phonetic Alphabet (IPA) script and its ASCII correspondences. Additionally, a broad phonemic categorization of the Sora phonemes is presented in Table 2.

Table 1: *Inventory of Sora Phonetic units*

Phonetic units in IPA	Phonetic units in ASCII
a	a
e	e
i	i
ə	E
o	o
u	u
b	b
d	d
ɡ	g
p	p
t	t
k	k
m	m
n	n
ɲ	nn
ŋ	N
j	j
r	r
l	l
w	w
ɽ	R
s	s
ʤ	z
ʔ	Q

Phonetic units of the Sora language of Assam have co-relation with the phoneme inventory of its parent language namely, Munda sub-family and Austroasiatic language family. Also, typical characteristics of some phonetic units have even direct co-relation with the Munda sub-family. For instance, asymmetry between voiceless dental stop /t/ and voiced alveolar stop /d/ found in Sora is also a typical feature in Munda sub-family of the Austroasiatic language family.

Table 2: *Categories of Sora Phones*

Category	Phones
Vowels	a, e, i, E, o, u
Obstruents	p, b, t, d, k, g, s, z, Q
Sonorants	m, n, nn, N, j, r, l, w, R

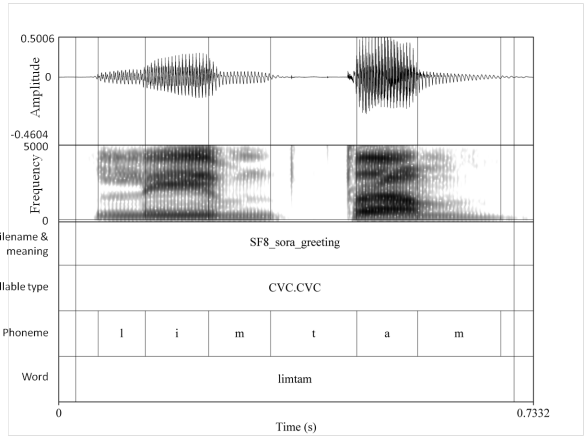

Figure 1: *Example of transcription of a Sora utterance in the corpus*

2.2. Syllable Properties

Basic syllable structure of the Sora language of Assam is (C)V(C) and a word length is minimally disyllabic. This implies that, shortest meaningful word in the language is usually disyllabic. Evidences suggest that, monosyllabic words are rare in the Sora language of Assam and a few monosyllabic words that occur sometimes are also bimoraic. Additionally, acoustic analysis provided evidence that, Sora has iambic stress pattern [14]. There are evidences that, in a disyllabic word, the second syllable is longer, louder and pitched higher then the first syllable. Thus, similar to the phoneme inventory, syllable properties also show a co-relation between Sora of Assam and its parent language, as, a week strong word prosody is typical to the Austroasiatic language family.

3. Organization of the Speech database

The speech database used in this work is created using the data collected for acoustic phonetic analysis of the Sora language from Assam, India [8]. The data is recorded using a linear PCM recorder with a unidirectional head-worn microphone in the field. The speech samples are collected as small sentences and later split into individual words with excess silence areas manually removed for phonetic analysis. The individual words split for phonetic analysis later served as inputs in the Sora ASR system as the split data contained phonetic transcriptions, while the small sentences did not. Word level transcription and corresponding phone sequences are created manually as the part of Acoustic Phonetic analysis [8]. An example of transcription of a Sora utterance in the corpus is illustrated in Figure 1.

The stereo channel audio files are originally recorded with 44100 Hz sampling frequency and 32 bit/sample resolution.

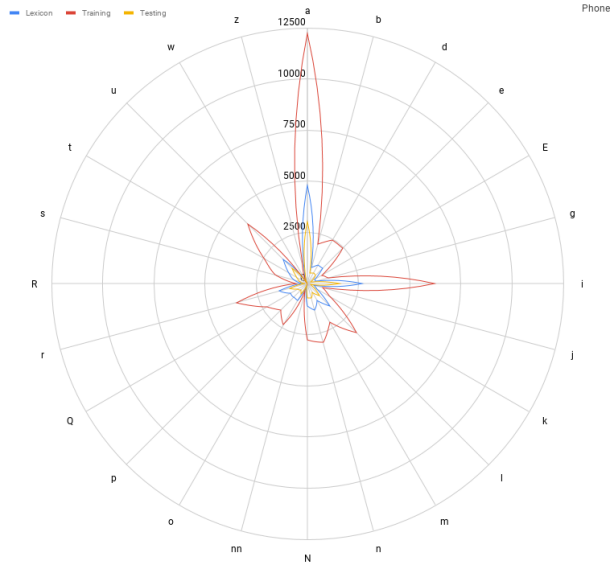

Figure 2: *Occurrences of individual phones in the database*

The database consists of 20 individual speakers and a total of $11,569$ utterances. The overview of the dataset is provided in Table 3. The files are annotated using the speaker and utterance information. Later, the stereo files were converted to mono with a sampling rate of 16000 Hz and a bit-rate of 16 bit/sample resolution.

As the database is not designed for automatic speech recognition, the frequency of occurrence of individual phones in the database is not uniform. For example, the phone /a/ has appeared $12,207$ times whereas, the phone /w/ appeared only 64 times. The variation in the occurrence of phonemes is illustrated in the Figure 2. The entire databases is randomly split in two parts in 8:2 ratio for training and testing, respectively, in a Round Robin fashion so that biases resulting from unbalanced data can be reduced.

Table 3: *Overview of speech database*

Details	Quantity
Total audio length	≈ 146.3 min
Training audio length	≈ 117 min
Testing audio length	≈ 9.3 min
Total no of Utterance files	11569
No. of Utterance files for training	9256
No. of Utterance files for testing	2313
No. of Total unique words	3499
No. of unique words for training	3253
No. of unique words for testing	1606
No. of unique phones	24

4. Development of an ASR for Sora

The speech recognition system is developed using the state-of-the-art Kaldi Speech Recognition toolkit [15]. The acoustic models were created using GMM-HMM,SGMM-HMM and DNN-HMM systems. GMM-HMM system is a baseline system for ASR, while SGMM-HMM is an advanced version of GMM-HMM with higher accuracy rate. DNN-HMM is a state-of-the-art technique using Deep Neural Network, but this technique requires huge amount of data for significant increase in accuracy.

4.1. Feature extraction

Mel-Frequency Cepstral Coefficients (MFCC) features are used in developing the ASR system. The MFCC features are calculated from the audio file split into 25ms long frameswith a frame shift of 10ms, using Hamming window.

4.2. GMM based modeling

In this step, static modeling scheme Gaussian Mixture Modeling (GMM) and dynamic modeling scheme Hidden Markov Model are used. This is the most common method used in speech recognition. In this method, models are implemented both with and without context information.

4.2.1. Monophone modeling

In monophone training, acoustic models of individual 24 phonemes are calculated without any context information. The training is based on the HMM models, created using the phone sequence of individual lexicon entry. The accuracy of recognition in this system is less due to the lack of context information.

4.2.2. Triphone modeling

To introduce context information, three different triphone models are trained. Tri1 works on the basis of decision tree-based state tying. In Tri2 the dynamic features are captured using the static MFCC vector from neighbouring phones, split in time. Tri3 introduces the speaker information using feature space Maximum Likelihood Linear Regression (fMLLR). The recognition rate of this system is much higher compared to monophone models.

4.3. SGMM based modeling

In Subspace Gaussian Mixture Model or SGMM, the HMM states share the same GMM structure. The number of Gaussians is same in each state. This method gives better results in smaller databases[16].

4.4. DNN based modeling

DNN is a feedforward network with multiple hidden layers. Acoustic properties of human speech varies due to many factors like accent, dialect etc. It is difficult for other systems to consider the effects of all factors and take accurate decision. DNN has the capability to discover and learn complex feature structures from very large amount of data set[17].

5. Results and discussion

The system shows the best result of 13.23%WER with SGMM and 13.92% with DNN. Word error rate (WER) of different training methods are shown in Table 4. The system is not speaker independent as the speaker set is same in both training and testing.

The analysis of the misrecognized words reveals two types of errors namely substitution errors and insertion errors. In substitution errors, one word is misrecognized as another word. Such errors occur mostly due to the confusion between phonetically similar words. The top 10 substitution errors are shown in Table 5. In insertion error, one word is misrecognized as

multiple words. This type of errors is found mainly in long words where the system confuses between words and two or three smaller words with similar phone sequences. The list of top 10 insertion errors is shown in Table 6.

Table 4: *Testing Result in Different Algorithms*

Algorithms	Word Error Rate
Monophone	25.38%
Tri1	15.56%
Tri2	14.53%
Tri3	14.57%
SGMM	13.23%
DNN	13.92%

Table 5: *List of top 10 substitution errors and there occurrence*

Expected output	Detected word	Occurrence
jira	ira	3
aQdia	addia	2
aiN	ajiN	2
alam	alaN	2
are	arre	2
asatuQ	assatu	2
aseiziN	asiziN	2
asub	asup	2
lpitinapmeai	itinammeailp	2
moQo	moQ	2

Table 6: *List of top 10 insertion errors*

Expected output	Detected words
uraolsiQruNruNsuN	uRaaN siQruN usuN
amannoriNnennorilp	aman uRiQ Nnen usi
addEiirguQnailp	aQda irgana iQ
aundruiQdoNlp	ondri oN
madoduleQdatiNlp	madoi aQdur edatin
amanitinlp	aman itin
kodingizalp	kolin giza
lpalalidaQa	alali ida
aNguNlai	aNguN ai
lpumeNjirnaumeNtEda	umeN jirna umeN tEda

6. Conclusion

In this work, an ASR system has been developed using the database not specifically designed for speech recognition. The goal of this work is to design and test ASR system for a low resource language, Sora. The database used in this work has only $\approx$ 146 minutes of audio. The database is not uniform in terms of the distribution of phonemes. The system is trained using isolated words and without a language model. The observations from this work indicates that the system gets confused between phonetically similar words. This problem may be so

lved using a robust language model. The results demonstrate that an efficient ASR system can be built, bootstrapping on data that is not specifically collected for such a system. It also demonstrates the potential of using linguist collected speech data for acoustic phonetic analysis to build ASR systems for under-resourced languages.

7. References

[1] S. Shahnawazuddin, D. Thotappa, B. D. Sarma, A. Deka, S. R. M. Prasanna, and R. Sinha, "Assamese spoken query system to access the price of agricultural commodities," in *2013 National Conference on Communications (NCC)*, Feb 2013, pp. 1–5.

[2] B. Medhi and P. H. Talukdar, "Isolated assamese speech recognition using artificial neural network," in *2015 International Symposium on Advanced Computing and Communication (ISACC)*, Sept 2015, pp. 141–148.

[3] M. P. Sarma and K. K. Sarma, "Assamese numeral speech recognition using multiple features and cooperative lvq-architectures," 2013.

[4] A. Dey, W. Lalhminghlui, P. Sarmah, K. Samudravijaya, S. R. Mahadeva Prasanna, R. Sinha, and S. Nirmala, "Mizo phone recognition system," 12 2017.

[5] B. D. Sarma, A. Dey, W. Lalhminghlui, P. Gogoi, P. Sarmah, and S. M. Prasanna, "Robust mizo digit recognition using data augmentation and tonal information," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 621–625.

[6] S. Morey *et al.*, "The Tai languages of Assam: a grammar and texts." 2005.

[7] B. Bielenberg and Z. Nienu, "Chokri (Phek dialect): phonetics and phonology," *Linguistics of the Tibeto-Burman Area*, vol. 24, no. 2, pp. 85–122, 2001.

[8] L. Horo, "An acoustic phonetic description of assam sora," Ph.D. dissertation, Indian Institute of Technology Guwahati, 2018.

[9] G. D. Anderson, "A new classification of the Munda languages: Evidence from comparative verb morphology," in *209th meeting of the American Oriental Society, Baltimore, MD*, 1999.

[10] R. S. G. V. Ramamurti, *A manual of the Sora (or Savara) language*. Superintendent, Government Press, Madras: Published under the authority of the Director of Public Instruction, 1931.

[11] R. K. Kar, *Savaras of Mancotta*. New Delhi: Cosmo Publications, 1981.

[12] Registrar General of India, *Distribution of the 99 Non-Scheduled Langauges- India/ States/ Union Territories-2011 Census*. New Delhi: Office of the Registrar General and Census Commissioner, India, 2011.

[13] P. Sarmah and L. Horo, "Language maintenance and shift: The Assam Sora perspective," in *LAMAS, Indonesia*, 2015.

[14] L. Horo and P. Sarmah, "Acoustic analysis of vowels in Assam Sora," *North East Indian Linguistics*, vol. 7, pp. 69–88, 2015.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[16] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture modela structured model for speech recognition," *Computer Speech Language*, vol. 25, no. 2, pp. 404 – 439, 2011, language and speech issues in the engineering of companionable dialogue systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523081000063X

[17] D. Fohr, O. Mella, and I. Illina, "New Paradigm in Speech Recognition: Deep Neural Networks," in *IEEE International Conference on Information Systems and Economic Intelligence*, Marrakech, Morocco, Apr. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01484447