Automatic Speech Recognition of Galo

Karter Nyodu and Samudra Vijaya

Abstract The development of spoken language systems for the tribal languages of India is very beneficial to society. The details of the implementation of automatic speech recognition for Galo language, spoken in the northeast Indian state of Arunachal Pradesh, are presented here. A multi-speaker speech database of continuously spoken Galo sentences was specifically created for this purpose. The speech recognition system was implemented using Kaldi, a public domain software toolkit. The automatic speech recognition system recognizes Galo sentences spoken continuously by new speakers with an accuracy of about 80%.

Keywords ASR · Galo · Speech database · Kaldi · Speech-to-text

1 Introduction

Arunachal Pradesh is one of the states in the northeastern region of India. Being home to a large number of tribes and subtribes, a large number of languages are spoken in the state. Galo, a language of the Tani branch of the Tibeto-Burman language family, is spoken by the people belonging to the Galo tribe of Arunachal Pradesh. Galo is one of 12 tribal languages of Arunachal Pradesh, listed in the 'The Scheduled Castes and Scheduled Tribes Lists' published by the census of India [1]. Although 29,246 Indians stated Galo as their native language in 2011 [2], it is in the UNESCO list of 'vulnerable' languages [3]. Figure 1 shows the map of Arunachal Pradesh state wherein the primary area of Galo speakers is shaded [4].

Even though 7 of the 99 major, non-scheduled languages of India belong to the state of Arunachal Pradesh [2], study and technology development of the languages of the state are limited. A speech database of English, Hindi, and local language

K. Nyodu (✉) · S. Vijaya
Centre for Linguistic Science and Technology, Indian Institute of Technology Guwahati, Guwahati, India
e-mail: karter@iitg.ac.in

S. Vijaya
e-mail: samudravijaya@gmail.com

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 686, https://doi.org/10.1007/978-981-15-7031-5_63

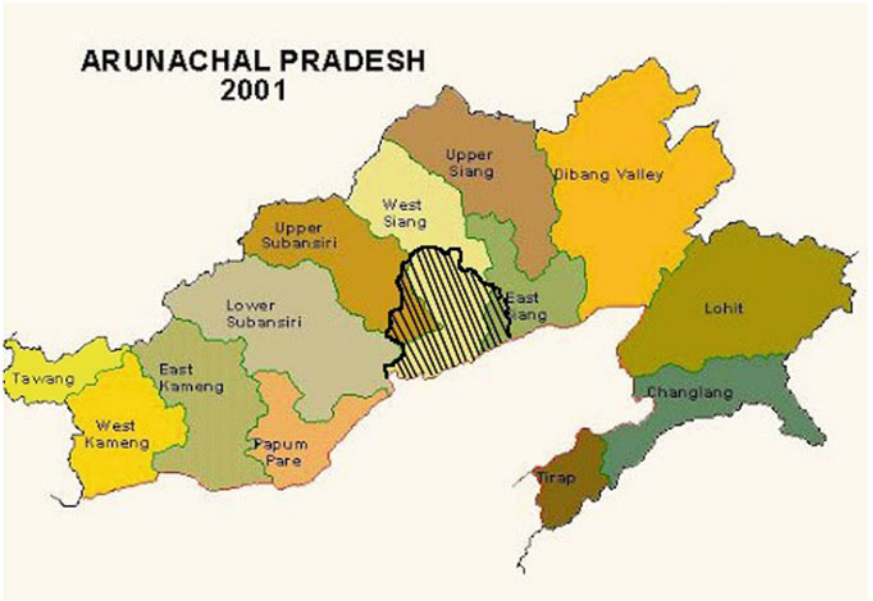

Fig. 1 Map of Arunachal Pradesh where the Galo speaking area is shaded. *Sources* [4, 5]

of Arunachal was created and used for speaker identification [6]. The same authors prepared a similar, but a larger database of speech from 200 speakers. This Arunchali language speech database was used for identifying the language of the input speech as one of the English, Hindi, Adi, Nyishi, Galo, and Apatani [7].

Very little research work on the Galo language has been reported in the literature. A book introducing the Galo language was written in 1963 [8]. A descriptive grammar of the Lare dialect of Galo was the theme of a recent Ph.D. thesis work [5]. The sole work on spoken Galo is an investigation of the acoustic features of Galo phonemes using formant frequencies and cepstral coefficients as features [9]. However, no spoken language system for Galo, whether it is speech-to-text or text-to-speech, has been implemented so far. The primary objective of this research work is to implement a speech-to-text system for Galo. A secondary objective was to create spoken language resources necessary for achieving the primary objective. Specifically, the following are the outcomes of the current research work:

1. A Galo speech database consisting of sentences read by many native Galo speakers.
2. Transcription of the speech data using the Galo script, which is a modified Roman Script.
3. A multi-speaker, continuous speech recognition system for Galo language.

The organization of the paper is as follows. The details of the spoken language resources developed for Galo language are given in Sect. 2. The implementation of ASR systems using various acoustic models and evaluation methodology is described

Table 1 Two examples of dialect-dependent transcription of words using modified Roman Script

Dialect		Gloss
Lare	Pugo	
‘aci	‘asi	brother
inrv	inye	to go

in Sect. 3. The results and discussion are also presented in Sect. 3. Section 4 draws some concluding remarks.

2 Spoken Language Resources

This section describes the steps for the development of linguistics resources for training and evaluating the Galo Automatic Speech Recognition (ASR) system. The subsections contain detailed descriptions of text and speech corpora.

2.1 Text Corpus

The text corpus contains a total of 200 short sentences, selected from the Galo-English dictionary [10]. The text corpus consists of 721 unique words. Galo script, which is a variety of modified Roman Script (MRS) [10], was used for writing the text.

M.W. Post, in his Ph.D. thesis, lists 6 dialects of Galo: lare, pugo, tai, gensi, karko, and zirdo [5]. Dialect-dependent variations in lexical terms were taken into account while writing the transcriptions of recorded speech. Table 1 shows two such examples of dialect-dependent variations.

2.2 Recording of Speech Data

Thirty-five Galo speakers were asked to read sets of 30–50 Galo sentences written in modified Roman Script. Speech data were collected at users’ locations using laptop, PC, and an earphone. The speech data were recorded at the sampling rate of 44.1 kHz, 16-bit, mono and were stored in wav format. The statistics of the number of speakers and the speech files is given in Table 2.

The set of 35 Galo speakers belonged to two broad categories of dialects: Lare and Pugo. The dialect-dependent statistics of the speech corpus is given in Table 3. The lexical transcriptions of the speech files were dialect-dependent.

Table 2 Statistics of the Galo speech corpora

Number of	Male	Female	Total
Speakers	20	15	35
Speech files	850	650	1500

Table 3 Distribution of speakers and files according to dialect

Number of	Dialect		Total
	Lare	Pugo	
Speakers	22	13	35
Speech files	930	570	1500

2.3 Pronunciation Dictionary

A pronunciation dictionary was created manually according to the format specified by Kaldi [11], the software toolkit used for the implementation of ASR systems. The entries in the dictionary specify the pronunciation of each word in the text corpus in terms of the phones or phone-like acoustic units of the language.

The Galo script [5] lists 7 vowels and 17 consonants. The script does not seem to distinguish between long and short vowels. In addition, diphthongs are also used in spoken Galo. Further, geminated consonants are present in the spoken Galo language. In order to have acoustic models of these acoustic–phonetic variations, we use a list of 19 (7 short + 7 long + diphthongs) vowel-like labels and 31 (17 single + 14 geminate) consonant-like labels. The labels follow the ILSL12 [12] convention, augmented with notations to mark tones of the language. Table 4 shows a few entries in the pronunciation dictionary.

Table 4 First two columns of the table show typical entries in the pronunciation dictionary

Word	Label sequence	Gloss
panam	p a n a m	To cut
paanam	p aa n a m	To hover
kai	k ai	Big
alo	a l o	Salt
allo	a ll o	Tomorrow

It illustrates the ILSL12 labels used to prescribe the pronunciation of Galo words involving long vowels, diphthongs, and geminated consonants

3 Implementation, Evaluation, and Results

The Kaldi software toolkit was used for the implementation of the ASR system. In this section, the details of the implementation, the evaluation methodology, the results, and the discussion are presented.

3.1 Feature Extraction

The default setting of the Kaldi toolkit was used for feature extraction. Thirteen mel-frequency cepstral coefficients (MFCC) were computed from every speech frame of 25 ms duration at a frame rate of 100/sec. Further, the first and second derivatives of 13 MFCCs were computed. As a result, a 39-dimensional feature vector was obtained.

3.2 Acoustic Modeling

For training the acoustics model, the default setting of the Kaldi toolkit was used. Six types of hidden Markov models (HMM) were trained. Each state of HMM is characterized by either a Gaussian mixture model (GMM) or a subspace GMM or a deep neural network (DNN). A brief description of each of these models is given below.

The simplest acoustic model (called **Mono**) models the context-independent phone-like units, also known as monophones. The second model **(Tri1)** is the first of a series of HMMs that model context-dependent phones (**tri**phone). The third acoustic model **(Tri2)** uses linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT). The fourth acoustic model **(Tri3)** uses the speaker adaptive training and feature-space maximum likelihood linear regression (fMLLR). The fifth acoustic model (**SGMM-HMM**) employs subspace GMMs instead of GMMs. The last model is the **DNN-HMM**-based model which uses the posterior probability given by a DNN to compute the state-dependent likelihood of feature vector.

3.3 Language Model

Bigram language model was used to model the syntax. The parameters of this model are estimated using the transcription of the train data. IRSTLM software was used to train the language model [13].

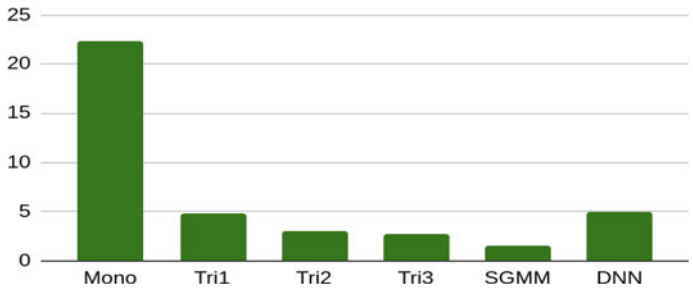

Fig. 2 WER (%) of the Galo ASR system for different types of acoustic models. Test data are the same as training data (entire speech corpus)

3.4 Results and Analysis

This section presents the performance of the Galo speech recognition system using various acoustic models. The word error rate (WER) is used as a measure of the performance of the ASR; lower the WER, the better the system is. The WER is computed as

$$\text{WER}(\%) = 100(D + S + I)/N$$

where D is the number of deletion errors, S is the number of substitution errors, I is the number of insertion errors, and N is the total number of words present in the reference transcription.

3.4.1 Performance on Training Data

The first evaluation was carried out using the same data for both training and testing. Here, the entire speech data were used for this purpose. The word error rates of the ASR system with different acoustic models are shown in Fig. 2.

The word error rate for the monophone model is 22%. The WER decreases drastically by using context-independent phone models (Tri*). The WER is the lowest (2%) for SGMM-HMM model. The WER of the DNN-HMM-based ASR system is slightly higher than that of the SGMM model. This is possibly due to the small size of the speech corpus, as DNN demands a large amount of speech data to adequately train a large number of parameters.

3.4.2 Performance on Test Data

In order to estimate the performance of the ASR system with respect to unseen speech data, a threefold cross-validation methodology was adopted. Accordingly, the entire

Table 5 Statistics of the three subsets of speech files used for threefold cross-validation of ASR

Number of	Fold 1		Fold 2		Fold 3	
	Train	Test	Train	Test	Train	Test
Utterances	990	510	1000	500	1010	490
Male speakers	13	6	12	6	12	7
Female speakers	10	6	11	5	11	5
Total	23	12	24	11	23	12

speech corpus was divided into three threefolds (subsets). These subsets were divided in such a way that each subset had an approximately equal number of speech files from both the female and male speakers. One subset was reserved/labeled as the test set. The system was trained with the remaining two subsets. The WER of the system with respect to the unseen test data is computed. This process is repeated for all three sets. Such a threefold evaluation is carried out for all the six acoustic models. The characteristics of the three subsets employed in our experiments are shown in Table 5.

The WERs of six types of acoustics models in threefold evaluation experiments are listed in Table 6. The WER of the ASR systems is around 20% for unseen data in all threefolds. This value of WER for unseen test data is an order of magnitude higher than that for the training data. Also, the difference between the WER of the context-independent (monophone) model and the best (tri3) model is negligible. Even though triphone models are more powerful, their potential is yet to be exploited due to lack of adequate amount of training data. The WER increases from 18% to 26% when SGMM-HMM acoustic model is used. A similar increasing trend in WER is observed when a DNN is used instead of a GMM.

The WERs of the ASR systems using six types of acoustic models are shown in the form of a bar chart in Fig. 3.

Table 6 WERs of various types of acoustic models in threefold experiments

	WER (%)			
Model	Fold 1	Fold 2	Fold 3	Average
Mono	19	21	17	19
tri1	18	23	17	19
tri2	21	25	19	22
tri3	17	20	15	18
SGMM	25	29	22	26
DNN-HMM	21	23	20	22

The WERs are around 20%

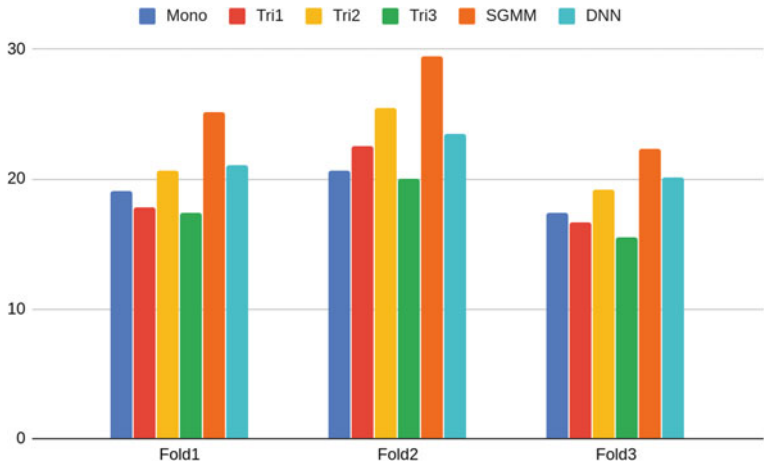

Fig. 3 WERs of various ASR systems in a threefold cross-validation experiment

4 Concluding Remarks

An automatic speech recognition system was implemented for Galo, a zero-resource language, spoken by Galo tribals in the state of Arunachal Pradesh. The system was trained using a preliminary multi-speaker speech database. The effectiveness of different types of acoustic models was investigated using a threefold cross-validation methodology. While the recognition accuracy of the ASR system is good for training data, the accuracy decreases significantly for all six types of acoustic models. This is due to the limited amount of speech data that could be collected in this initial effort. Future work includes expansion of the spoken language corpus and investigation of the utility of using prosodic features for machine recognition of this tonal language.

References

1. List of notified scheduled tribes, census of India. Online: http://censusindia.gov.in/Tables_Published/SCST/ST%20Lists.pdf
2. Office of the Registrar General and Census Commissioner India (2011) Statement-1 part-b languages not specified in the eighth schedule (non-scheduled languages). Online: http://censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf
3. Moseley C (ed.) (2010) Atlas of the world's languages in danger, 3rd edn. Paris, UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/en/atlasmap/language-id-988.html
4. Office of the Registrar General and Census Commissioner, India, State Maps. http://censusindia.gov.in/maps/State_Maps/StateMaps_links/arunachal01.html
5. Mark William Post (2007) A grammar of Galo. PhD Thesis, La Trobe University

6. Bhattacharjee U, Sarmah K (2012) A multilingual speech database for speaker recognition. In: 2012 IEEE international conference on signal processing, computing and control. https://doi.org/10.1109/ispcc.2012.6224374
7. Bhattacharjee U, Sarmah K (2013) Language identification system using MFCC and prosodic features. In: 2013 international conference on intelligent systems and signal processing (ISSP). https://doi.org/10.1109/issp.2013.6526901
8. Das Gupta SK (1963) An introduction to the Gallong language. Shillong, Northeast Frontier Agency
9. Sora M, Talukdar J, Talukdar PH (January, 2013) Formant frequency and Cepstral method estimation of Galo phonemes using acoustical cues. Int J Inf Electron Eng 3(1)
10. Rwbaa I, Post MW, Rwbaa I, Xodu M, Bagra K, Rwbaa B, Rwbaa T, Ado N, Keenaa D (2009) Galo-English dictionary with English-Galo index
11. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K (December, 2011) The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. IEEE signal processing society. Online: https://publications.idiap.ch/downloads/papers/2012/PoveyASRU20-112011.pdf
12. Indian Language speech sound label set (ILSL12). Online: https://www.iitm.ac.in/donlab/tts/downloads/cls/clsv2.1.6.pdf
13. Federico M, Bertoldi N, Cettolo M (2008) IRSTLM: an open source toolkit for handling large scale language models. Proceedings of Interspeech, pp 1618–1621