Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 167–171
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Acoustic-Phonetic Approach for ASR of Less Resourced Languages Using Monolingual and Cross Lingual Information

Shweta Bansal, Shweta Sinha, Shyam S. Agrawal
KIIT College of Engineering, Amity University, KIIT College of Engineering
Gurugram, India
bansalshwe@gmail.com, shwetakant.sinha@gmail.com, ss_agrawal@hotmail.com

Abstract

The exploration of speech processing for endangered languages has substantially increased in the past epoch of time. In this paper, we present the acoustic-phonetic approach for automatic speech recognition (ASR) using monolingual and cross-lingual information with application to under-resourced Indian languages, Punjabi, Nepali and Hindi. The challenging task while developing the ASR was the collection of the acoustic corpus for under-resourced languages. We have described here, in brief, the strategies used for designing the corpus and also highlighted the issues pertaining while collecting data for these languages. The bootstrap GMM-UBM based approach is used, which integrates pronunciation lexicon, language model and acoustic-phonetic model. Mel Frequency Cepstral Coefficients were used for extracting the acoustic signal features for training in monolingual and cross-lingual settings. The experimental result shows the overall performance of ASR for cross lingual and monolingual. The phone substitution plays key role for the cross lingual as well as monolingual recognition. The result obtained by cross-lingual recognition compared with other baseline system and it has been found that the performance of the recognition system is based on phonemic units. The recognition rate of cross-lingual generally declines as compared with themonolingual.

Keywords : Cross-lingual, Mono-lingual, ASR

1. Introduction

Due to an increase in the demand for the speech recognition systems in various languages, the advancement of multilingual systems also increases. For developing the robust multilingual system, it is good to combine the phonetic inventory of the multiple languages to be identified into a single universal acoustic model because of the subsequent merits:

- ❖ The complication of the system gets reduced due to decrease in the size of the parameters by combining the parameters across the languages.
- ❖ The recognition of the new language is possible in the fast and efficient manner even if the existing quantity of training data is not sufficient((Schultz et al.,2013).

For merging the acoustic models of the different languages require the clarity of the speech sounds of a particular language. Former multilingual recognition systems with shared acoustic-phonetic models were restricted to context-independent modeling ((Stuker et al., 2003). In the case of monolingual, it is already verified that the recognition rate has been increased by context-dependent modeling (Partha Lal and Simon King, 2012). We used here context-dependent model to construct the robust and efficient multilingual models and develop a common system which shares their parameters by applying the clustering procedure based on decision tree and analyze the subsequent decision tree.

For conducting the experiments, we have created our multilingual database which is briefly described in the first section of this paper. In the second section, we explain the procedure to design the monolingual system. The experimental sections give results for the monolingual and crosslingual tests based on the systems designed.

2. Design for Text and Acoustic Corpus

For designing ASR in virtue of under-resourced languages, text and acoustic data collection is a predominantly difficult task. The corpus has been designed in following phases:

A. Extraction of Phonetically rich sentences

The process starts with collection of text corpus. This has been done by crawling the web for text corpus. As the corpus on web contains lots of clatter, it required cleaning and filtering. Once the clean corpus is built then phonetically rich sentences are extracted from it. Finally, text prompt sheet for each language were designed. Each prompt sheet consists of 300 meaningful phonetically rich sentences. The sentence length in the text corpus varies from 5 to 12 words.

B. Cleaning and filtering of corpus

Identification of improper syntax e.g. existence of invalid bigrams/character combinations has been done. Sentences with foreign word are filtered so as to have a good quality monolingual corpus. Inadequate sized sentences and words are identified and removed. Duplicate sentences along with duplicate punctuations are also removed.

C. Collection of Speech data

The corpus is recorded using 100 native speakers (60 male and 40 female) for each of the three languages. The age group of all the native speakers were 18 to 55 years and had at least 10 years of formal education in their respective language. Each speaker has to read 300 continuous sentences in one session. The total number of utterances in the corpus is (300 sentences × 3 languages) × 100 speakers = 9000. Nearly 1.20 h of read speech samples are obtained from each speaker. We apportion the recorded dataset into training and testing sets, with an 80-20 split. All recording was done using a single microphone in the office environment. The recorded signals were sampled at 16 kHz using the software GoldWave and are represented as 16 bit number.

LANGUAGES	UTTERANCES	SPEAKERS	
		MALE	FEMALE
HINDI	3000	60	40
PUNJABI	3000	60	40
NEPALI	3000	60	40

Table1: Detail of the Collected Speech Corpus

3. Phone Switching

The phonetic units of every language have peculiar characteristics. The correlation among the acoustic inventory of the prescribed languages must be discovered for performing the monolingual and cross-lingual speech recognition (T. Schultz and A. Waibel,1997). This section will describe the approaches to determine resemblance among sounds of the different languages. In monolingual and cross-lingual speech recognition, usually phonemes are used for the representation of the words. A phoneme may be realized by different phones, for example the phones /sh/ and /s/ can be represented by the same phoneme. The relation between the phones and phonemes of a language differs across languages. For example in Nepali, no difference between /kʂ /,/ ʂ/,/ ʃ/, / ʃrə/ and /s/ and they would belong to the same phoneme class in that language. In other languages, however they represent each a phoneme class on their own. As Punjabi is a tonal language, it was observed that Punjabi speakers used to pronounce the phonemes with tones which change the perception of that particular phoneme. It has been observed that the sound of some phonemes changed according to the positions of phoneme by Punjabi speakers. For example, 'ਘ'/g^h/ is heard as 'ਘ'/g^h / only when it is in the initial position of the word, however, sound as 'ਗ'/g/ when /g^h/ lies at middle and final positions of any word. Similarly sound of /b^h / in Punjabi changes according to the position in a particular word. /bh/sounds like /b/ when occurring at the initial and final position of any word. This type of phone switching can affect the recognition rate of a particular language.

4. Monolingual Speech Recognition

For this work, we have developed three monolingual speech recognition baseline systems for Hindi, Punjabi and Nepali by applying the bootstrap HMM technique for initializing the acoustic models of the mentioned languages. the resultant monolingual system comprises of an entirely continuous 3-state HMM system for each involved language (R.K. Aggarwal, M. Dave,2012). The obtained monolingual system for each language is context-dependent and each HMM state having 1000 polyphone models. Modeling of each state is done by the use of a common codebook which consists of 32 Gaussian mixture distributions along with 24-dimensional feature space. The features of the acoustic signal were extracted by using Mel Frequency Cepstral Coefficients for monolingual speech recognition. The input speech sampled at 16 kHz was used to calculate the first and second derivative of power and 16 cepstra and then the process of Mean subtraction is employed. The word error rates (WER in percentage) and sentence error rate (SER) obtained by each of the monolingual system is shown in table 2. The performance for the Punjabi system is experienced the lower accuracy rate due to the use of morpheme-based units.

LANGUAGES	HINDI	NEPALI	PUNJABI
WER	12%	10.9%	20.6%
SER	8.1%	7.1%	11.2%

Table 2: WER & SER (in %) for Hindi Nepali & Punjabi

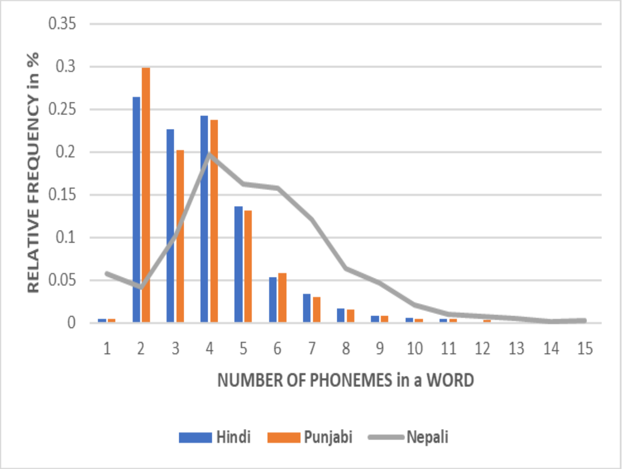

Figure 1: Plot of Number of Phonemes in a word vs Relative Frequency (%)

The total number of phonemes in a word along with their relative frequency in the training corpus was calculated as shown in figure 1. It has been observed from the plot that the Nepali language tends to have long words with phoneme 5,6,7, 10 or 15 which might make it easier to differentiate Nepali words with each other and results in a high accuracy rate. It is also be seen that in Punjabi data 30% of the words having only two phonemes which result in high confusability in recognition of these words. Therefore, the recognition rate of Punjabi is poor than others as shown in table 3.

LANGUAGES	HINDI-HINDI	NEPALI-NEPALI	PUNJABI-PUNJABI
Recognition Rate	88.05%	89.12%	79.4%

Table 3: Recognition rate of Hindi, Nepali & Punjabi

5. Cross-lingual Speech Recognition

In the experiment of Cross-lingual speech recognition, one language(L1) is used for training purpose and unknown language(L2) is used for testing (B. G. Nagaraja and H. S. Jayanna,2012). In this work, we need to integrate the acoustic models of the same sounds across languages into a common phone set. Moreover, the material which was used for the training of L1 is used for the estimation of parameters for developing the recognizer for the second language(L2). The effect of L1 on the recognizer of L2 is shown in Figure 2.

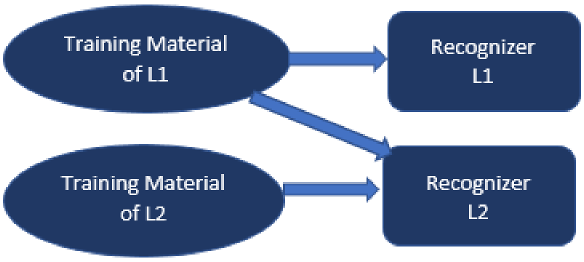

Figure 2: Setup for Cross – lingual recognition

All the sounds of each language were classified based on phonetic information and documented by using Indian Language Speech sound label set (ILSL) popularly known as “common phoneme set” as shown in Table 5. The common phone set comprises of 80 different phonemes including silence. On the basis of these 80 phonemes in the prescribed set, we developed cross-lingual recognizer for three languages, in which we share language and acoustic models across languages. Each phoneme gets initialized by the single mixture of 16 Gaussian distribution.

Following three cases were considered for performing the cross-lingual experiment for speech recognition:

- ❖ Training with Hindi language and testing with Nepali and Punjabi
- ❖ Training with Nepali language and testing with Hindi and Punjabi
- ❖ Training with Hindi language and testing with Punjabi and Nepali

In the current experiment, GMM is employed as a classifier which integrates EM optimization technique. The selection of the number of Gaussians is based on the amount of training material. In this test, a range of Gaussians components from 64 to 256 per state have been found convenient. As mentioned above, for cross-lingual recognition three cases are considered. In the first case, the system is trained with Hindi (HI) and Punjabi (PU) Language and tested with Nepali (NE) language. The success rate of the cross-lingual system for varying number of Gaussians is shown in Table 4:

Language (Tr/Te)*	**SIZE OF GAUSSIANS**					**Average Success Rate**
	32	**64**	**128**	**256**	**512**	
HI/PU	28	33.5	42.3	45	45.2	**38.7**
HI/NE	34.5	39.8	41	50	50.3	**43.1**
NE/PU	22	25	32.6	36.3	36.5	30.5
NE/HI	35	36.7	43.6	48.9	48.9	**42.6**
PU/NE	29	31.5	33.7	33.7	33.7	32.2
PU/HI	31	33	37	41.6	41.6	36.8

(*Tr-Training, Te- Testing)

Table 4: Rate of Performance (%) for Cross- lingual

The cross-lingual speech recognition system gives the excellent performance when the system gets trained by the Hindi (HI) language and tested by the Nepali(NE) language followed by NE(Tr)/HI(TE), HI(Tr)/PU(Te) whereas the performance of PU(Tr)/NE(Te) was not appreciable. It has also been observed that the performance rate does not affect very much by increasing the number of Gaussians from 256 to 512. As observed, the performance rate of HI(Tr)/NE(Te) was the highest among the other language combination. This may be due to the similarity in the script of both the languages as the writing style of Hindi and Nepali is the same as both the languages use the Devanagari script for the representation and Hindi use as secondary language in Nepal. It is also evident that the phonemes क्ष(kʂ),ष(ʂ),श(ʃ),श्र(ʃrə) of the Devanagari script get confused by स (s) in Nepali. On the other hand, the performance of PU/NE, NE/PU was the poorest among others. As Punjabi is tonal language (Dua et al.,2012), it was observed that Punjabi speakers used to pronounce the phonemes with tones which change the perception of that particular phoneme. Due to difference in the speaking style of Punjabi speakers the performance of the system gets deprived. This phenomena has also been outlined by Yogesh et al.(2017).

Phone Label	**IPA**	**Hindi/Nepali**	**Punjabi**
ac	**/ə/**	**अ**	**ਅ**
a	**/ɔ/**	**औ**	**ਔ**
aq	**/ɑ/**	**आ**	**ਆ**
i	**/i/**	**ई**	**ਈ**
ic	**/ɪ/**	**इ**	**ਇ**
u	**/u/**	**ऊ**	**ਊ**
uc	**/ʊ/**	**उ**	**ਉ**
e	**/e/**	**ए**	**ਏ**
ae	**/ɛ/**	**ऐ**	**ਐ**
o	**/o/**	**ओ**	**ਓ**
k	**/k/**	**क**	**ਕ**
kh	**/kʰ/**	**ख**	**ਖ**
g	**/g/**	**ग**	**ਗ**
gq	**/ɣ/**	**ग़**	**ਗ਼**
gh	**/gʱ/**	**घ**	
ng	**/ŋ/**	**ङ**	**ਙ**
c	**/tʃ/**	**च**	**ਚ**
ch	**/tʃ ʰ /**	**छ**	**ਛ**
j	**/dʒ/**	**ज**	**ਜ**

jh	/dʒʱ /	झ	
nj	/ɲ/	ञ	ਞ
t:	/ʈ/	ट	ਟ
t:h	/ʈʰ /	ठ	ਠ
d:	/ɖ /	ड	ਡ
d:h	/ɖʱ /	ढ	
n:	/ɳ /	ण	ਣ
t	/t/	त	ਤ
th	/t̪ʰ /	थ	ਥ
d	/d̪/	द	ਦ
dh	/d̪ʱ /	ध	
n	/n/	न	ਨ
p	/p/	प	ਪ
ph	/pʰ /	फ	ਫ
b	/b/	ब	ਬ
bh	/bʱ /	भ	
m	/m/	म	ਮ
y	/j/	य	ਯ
r	/r/	र	
l	/l/	ल	ਲ
x	/x/	ख़	ਖ਼
v	/ʋ/	व	ਵ
sh	/ʃ/	श	ਸ਼
s	/s/	स	ਸ
h	/ɦ/	ह	ਹ
z	/z/	ज़	ਜ਼
l:	ɭ	ऴ	ਲ਼
r:	ɽ	ड़	
f	/f/	फ़	ਫ਼
rq	/ɾ/		ੜ
SIL			

Table 5: List of Common Phone set for Hindi, Nepali and Punjabi

6. Conclusion

In the present paper, the Monolingual and Cross-lingual speech recognition systems were developed for Hindi, Punjabi and Nepali languages. It has been observed that in the mono-lingual study the performance of Nepali language was better than the other two languages. Furthermore, we also observed in a cross-lingual study that the Nepali language for training & testing with the Hindi language have good success rate. From table 4, it may be seen that 256 number of Gaussians gives optimum performance in all combinations of cross lingual recognition. Their performance may be ranked as HI/NE, NE/HI, HI/PU, PU/HI, NE/PU and PU/NE. The experimental results can be improved by employing more language-specific features and the latest modeling techniques in both Monolingual and Cross-lingual speech recognition system. In order to design the robust speech recognition system, the large text and speech data size is required.

7. Acknowledgement

The authors would like to acknowledge the suggestions and help received from Mr. Shambhu Sharan in analysis of the speech samples collected from speakers of different linguistic backgrounds. Our special thanks to Dr. Harsh Vardhan Kamrah and Mrs. Neelima Kamrah for their support and encouragement.

8. References

B. G. Nagaraja and H. S. Jayanna, "Mono and Cross lingual speaker identification with the constraint of limited data," *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, Salem, Tamilnadu, 2012, pp. 439-443.

Dua, Mohit, et al. "Punjabi automatic speech recognition using HTK." International Journal of Computer Science Issues (IJCSI) 9.4 (2012): 359.

Hemant A. Patil, Sunayana Sitaram, and Esha Sharma, “DA-IICT Cross-lingual and Multilingual Corpora for Speaker Recognition”, Proc.IEEE, pp. 187–190, 2009.

Partha Lal and Simon King “Cross-Lingual Automatic Speech Recognition Using Tandem Features” IEEE Transactions on Audio, Speech, And Language Processing, Vol. 21, No. 12, December 2013.

R.K. Aggarwal, M. Dave, “ Integration of Multiple Acoustic and Language Models for improved Hindi Speech Recognition System”. International Journal of Speech Technol ogy, Published Online, 3 Feb 2012.

Stuker, Sebastian /Metze, Florian / Schultz, Tanja / Waibel, Alex (2003): "Integrating multilingual articulatory features into speech recognition", In EUROSPEECH-2003, 1033-1036.

T. Schultz, N. T. Vu, and T. Schlippe, “Globalphone: A multilingual text & speech database in 20 languages,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 8126–8130.

T. Schultz and A. Waibel “Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Setsin” Proc. Eurospeech, pp. 371-374, Rhodes 1997.

Yogesh Kumar et al. "An automatic speech recognition system for spontaneous Punjabi speech corpus," International Journal of Speech Technology volume 20, pages 297–303, 2017