ISSN No. 0976-5697

Volume 2, No. 1, Jan-Feb 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Small Vocabulary Speech Recognition for Gujarati

Mr. Himanshu N Patel*
Assistant Professor,
Anand Institute of Information Science,
Anand-388001, Gujarat, India,
hnp.aiis@gmail.com

Dr. P.V. Virparia
Reader,
Department of computer Science,
Sardar Patel University, Anand-388001, Gujarat, India,
pvvirparia@gmail.com

Abstract: In this paper we describe how to recognize small vocabulary speech in Guajarati language using English Speech Recognition Engine. We present a technique for fast bootstrapping of initial phone models of a Gujarati language. The training data for the Gujarati language is aligned using an existing speech recognition engine for English language. This aligned data is used to obtain the initial acoustic models for the phones of the Gujarati language. Following this approach requires less training data. As is inherent in phonetic languages, rules generally capture the mapping of spelling to phonemes very well. This paper described technique used and result obtained.

Keywords: Gujarati Speech Recognition, Speech Recognition Engine, Acoustic model, language model.

I. INTRODUCTION

An automatic speech recognition (ASR) system consists of two main components

A. An acoustic model and
B. A language model.

The acoustic model of an ASR system models how a given word or "phone"[1] is pronounced. In most of the current ASR systems, the probability of a phone being spoken is modeled, using Baye's theorem, as follows:

$$P(M \mid O) = \frac{P(O \mid M) * P(M)}{P(O)}, \qquad (1)$$

Where **O** is the observation vector and *M* is the particular phone or word being hypothesized. Often, the probabilities *P*(*M*) are assumed to be equal for all of the phones; hence, the term *P*(**O**|*M*) is used to compute the likelihood of the hypothesized phone. The acoustic model consists of the speech signal features to be used for **O**, and a pattern matching technique to compare these features against a set of predetermined patterns of these features for a given word or phone.

The language model of an ASR system predicts the likelihood of a given word sequence appearing in a language. The most common technique used for this purpose is an *N*-gram language model. An *N*-gram model provides the probability of the *N*th word in a sequence, given a history of *N*-1 words that is, $P(Wi \mid Wi\text{-}1Wi\text{-}2..Wi\text{-}_{N+1})$. The *N*-gram model is trained over a large text corpus in the given language to compute these probabilities. For a hypothesized word, the language model score and the acoustic model score are combined to find the final likelihood of the word.

By using both the acoustic model and the language model, the combined likelihood of the word is computed as follows:

$P(W) = P(\mathbf{O} \mid Wi) * P(Wi \mid Wi\text{-}1Wi\text{-}2..Wi\text{-}N\text{-}1)$. (2)

For isolated word recognition, the above likelihood is computed for all words being considered, and the word having the highest likelihood is chosen as the recognized word. In the case of continuous speech recognition, the likelihood of a word is combined with the likelihood of other words to compute the combined likelihood of the sentence being hypothesized. To train the acoustic model, a phonetically aligned speech database is required. However, acoustic models are required in order to automatically align a speech database. Hence, it becomes a chicken and egg problem. One possible method is to manually align the speech database; however, manually aligning a large speech database is very time-consuming and error-prone. Obtaining initial phone models for a new language is thus a challenging task.

In this paper, we propose an approach for building good initial phone models through bootstrapping. We make use of the existing acoustic models of another language for bootstrapping. Following the approach proposed in [1], we define a phone mapping between the two languages to obtain an initial alignment of the target language speech data. While segmenting the aligned data for target language phones, we use a module called a lexeme context comparator, which helps in differentiating phones in the target language which were mapped to same phone in the base language. The proposed approach requires relatively lower amounts of speech data for the new language to build initial phone models.

II. BOOTSTRAPPING OF PHONE MODELS

In the bootstrapping approach, an already existing acoustic model of a speech recognition system for a different language is used to obtain initial phone models for a new language. In the literature [2, 4], there are primarily two approaches used for bootstrapping. We explain these approaches using English as the base language and Gujarati as the new or target language:

A. Bootstrapping through alignment of target language speech data

In the first approach, phonetic transcription of the target language text is written using the phone set of the base language. This is achieved by using a mapping defined between the two phone set, which is detailed in the subsection on phone set mapping. The speech data in the target language is aligned using the speech recognition system of the base language. Initial phone models for the target language can then be built from the aligned speech data.

B. Bootstrapping through alignment of base language speech

Data In the second approach, speech data of the base language itself is aligned using its speech recognition system. The aligned speech data of the base language is used as the aligned speech data for the target language using the mapping between the two phone sets.

III. TECHNIQUE

Figure 1 shows the technique that is used to align Gujarati speech by using an English speech recognition system. A mapping g() from a Gujarati phone set denoted by G to an English phone set denoted by **E** is used to generate the pronunciation of Gujarati words by the English phone set. Using linguistic knowledge, this mapping is based on the acoustic closeness of the two phones. The mapping is such that each phone g ε G is mapped to one and only one phone in E. A vocabulary created by such a mapping is used to align Gujarati speech data. Since more than one element in G may map to a single element in E, g() is a many-to-one mapping in general and hence cannot always be used in reverse to obtain g ε G from e ε E. Therefore, in order to recreate the alignment labels with Gujarati phones, an inverse mapping $g^{-1}()$ will not be feasible. A lexeme context comparator is used to generate the correct labels from e ε E. This uses the context to resolve the ambiguity which arises from the one-to-many mapping $g^{-1}()$.

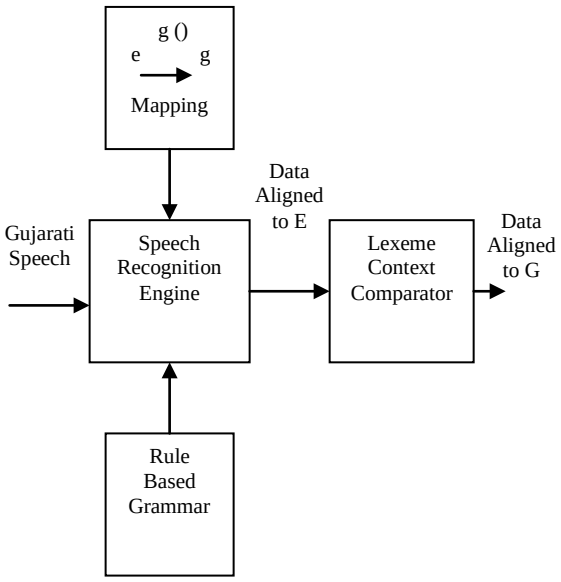

Figure-1: Alignment of the Gujarati Language Data

This technique would generate the aligned Gujarati speech corpus without the need for a Gujarati speech recognizer.

The Gujarati words of Table-1 are used for recognition.

Table-1 List of words for recognition

Sr. No	Word	Sr. No	word	Sr. No	Word
1	મહંત	11	દક્ષા	21	દેવ્યાની
2	કેશવ	12	નેનેશ	22	દીપ્તી
3	નિતિન	13	કાજલ	23	તનવી
4	હિમાંશુ	14	મીકિન	24	જીગ્નેશ
5	મીત	15	પંકતી	25	જિગીશા
6	અરવીન્દ	16	અજીત	26	અક્ષય
7	ધ્રુવી	17	મોના	27	રોનક
8	ભાવીન	18	સોનાલી	28	ભરત
9	વિશાખા	19	ભુમીકા	29	ભક્તવંતી
10	નિખીલ	20	આદર્શ	30	નરેન્દ્ર

IV. RESULTS

Total 31 speaker's speech is inputted and recognized words are displayed as Gujarati text on screen. The results are classified as recognized (successful), not recognized (unsuccessful) and misrecognition (mistake/misfire). There are 12 female and 19 male speakers.

R = Recognized Accuracy
M = Misrecognition
N = Not recognized

Table-2 All recognition Results

Gender	No of Speaker	R %	M %	N %
Female	12	85.28	12.50	2.22
Male	19	90.88	7.89	1.23
All	31	88.71	9.68	1.61

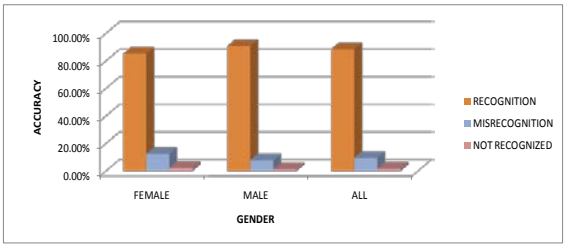

Figure-2: Gender wise accuracy

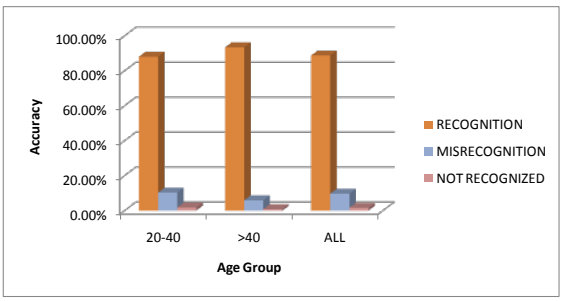

Figure-3: Age group wise accuracy

The results shows overall recognition accuracy of 88.71% with male recognition accuracy of 90.88% which is slightly higher than average and female recognition accuracy of 85.28% which is slightly less than average. The age group wise result is also nearer to average accuracy.

V. CONCLUSION

In this paper we present new techniques that have been used to build a small-vocabulary Gujarati speech recognition system. We present a technique for fast bootstrapping of initial phone models of a new language. The training data for the new language is aligned using an existing speech recognition engine for another language. This aligned data is used to obtain the initial acoustic models for the phones of the new language. Following this approach requires less training data. As is inherent in phonetic languages, rules generally capture the mapping of spelling to phonemes very well. However, deep linguistic knowledge is required to write all possible rules, and there are some ambiguities in the language that are difficult to capture with rules. On the other hand, pure statistical techniques for base form generation require large amounts of training data that are not readily available. We evaluate the performance of the proposed approaches through various recognition experiments.

VI. REFERENCE

[1] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards Language Independent Acoustic Modeling," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Istanbul, 2000, pp. 1029 –1032.

[2] J. Kohler, "Multi-Lingual Phoneme Recognition Exploiting Acoustic–Phonetic Similarities of Sounds," Proceedings of the International Conference on Spoken Language Processing, Atlanta, 1996, pp. 2195–2198.

[3] O. Anderson, P. Dalsgaard, and W. Barry, "On the Use of Data-Driven Clustering Technique for Identification of Poly- and Mono-Phonemes for Four European Languages," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Adelaide, Australia, 1994, pp. 121–124.

[4] M. C. Yuen and P. Fung, "Adapting English Phoneme Models for Chinese Speech Recognition," Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998, pp. 80 – 82.

[5] M. Kumar, n. Rajput, and a. Verma "A large vocabulary continuous speech recognition system for Hindi" IBM J. RES. & DEV. Vol. 48 no. 5/6 September November 2004