PRAJÑĀ-Journal of Pure and Applied Sciences, Vol. 19: 37 - 40 (2011)
ISSN 0975 - 2595

BOOTSTRAPPING OF PHONE MODELS FOR A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION FOR GUJARATI

Himanshu N. Patel* and Paresh V. Virparia[1]

**Anand Institute of Information Science, Anand, Gujarat, India*
[1]G. H. Patel Dept. of Computer Science, Sardar Patel University, V. V. Nagar, Gujarat, India

ABSTRACT

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. 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:

In this paper we present a technique that has been used to build a large-vocabulary continuous Gujarati speech recognition system. 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.

Keywords: *bootstrapping, acoustic model, phone set mapping, LVCSR*

INTRODUCTION

An automatic speech recognition (ASR) system consists of two main components—an acoustic model and a language model. The acoustic model of an ASR system models how a given word or "phone" is pronounced.

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. By using both the acoustic model and the language model, the combined likelihood of the word is computed.

In order to train the acoustic model, a phonetically aligned speech database and acoustic models are required in order to automatically align a speech database. 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 preliminary phone models for a new language is thus a demanding job. In [1], Byrne et al. have suggested techniques to create phone models for languages which do not have a lot of training data available. They have used knowledge-based and automatic phone mapping methods to create phone models for the target language, using phone models of other languages. Previous approaches [2, 3] to generate initial phone models include bootstrapping from a multilingual phone set and the use of codebook lookup.

A codebook specifies the mapping to be used while performing the bootstrapping. The generation of this codebook requires linguistic knowledge of the languages. The technique mentioned in [2] requires a system already trained in the languages. On the other hand, the method in [3] requires labeled and segmented data in the language for which the system is to be trained. Authors in [4] describe various methods of generating the Chinese phone models by mapping them to the English phone models.

This requires the collection of specific utterances of isolated monosyllabic data that is difficult for a language such as Gujarati. Moreover, it may not be the best means for initializing the phone models that are to be used in large vocabulary continuous speech recognition tasks. Cross lingual use of recognition systems is also seen in [5], where the aim is to generate a crude alignment of words that do not belong to the language of the recognition system.

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. However, in the case of Gujarati, we have special acoustic classes, e.g., nasalized vowels and stressed plosives, which require more than one phone from the base language (English) for bootstrapping. We use this aligned data to obtain initial phone models of the target language. 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.

For training the acoustic model, baseforms for the training words are required along with the initial phone models. These baseforms are also required during recognition for each word in the vocabulary. Researchers have used a pure rule-based technique for baseform builders for phonetic languages [6]. The advantage of this technique is that once all of the rules are accounted for, the accuracy is very high; however, this requires deep linguistic knowledge that may be difficult to obtain [7]. While pronunciation rules can be extracted from existing online dictionaries, existing online dictionaries for Gujarati are not exhaustive in their word coverage or on pronunciations. Additionally, each such online dictionary for Gujarati requires a specific format in which the Gujarati characters are encoded, thus making them even more difficult to use. It is easy to capture the general linguistic nature of phonetic languages, but their idiosyncrasies and exceptions are difficult to capture by rules. On the other hand, using pure statistical techniques requires a large amount of training data that is not easily available for a new language.

Different statistical approaches have been tried for baseform builders. Decision trees [8 –11], machine learning techniques [12], delimiting, and dynamic time warping (DTW) [13] are a few of the techniques that have been studied. All of the statistical techniques require a large amount of training data for respectable accuracy. Moreover, their performance is compromised for "unknown words," typically proper nouns [9]. In order to improve the statistical techniques, other knowledge sources such as acoustics are used in conjunction with the spellings to obtain better results [14]. Pure acoustic-based baseform builders have also been built [15]. However, the techniques that use acoustics are restricted in their usage, since they require a recognition engine for the language and are better used for generating speaker-dependent pronunciations.

In this paper we present a hybrid approach that combines rule-based and statistical techniques in a novel two-step fashion. We use a rule-based technique to generate an initial set of baseforms

*Corresponding author: hnp.aiis@gmail.com, patel_himanshu6@gtu.edu.in

and then modify them using a statistical technique. We show that this approach is extremely useful for phonetic languages such as Gujarati. The phonetic nature of the language can be exploited to a greater extent by using the rule-based approach, while the statistical technique can be used to improve on this. We experimented with two different techniques as the statistical component of our hybrid system—one of them uses modification probabilities, while the other uses context-dependent decision trees.

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:

1) 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 sets, which are 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. The Gujarati phone set is presented in **Figure 1**.

For example,
BHARAT –/BH AA R AX TX/ (actual);
BHARAT –/B AA R AX TH/
(using English phone set)

In this case, the phones /BH/ and /B/ in the target language are both mapped to phone /B/ in the base language. Hence, to initially obtain the aligned data for /BH/, the data aligned with /B/ is randomly distributed between /BH/ and /B/. Phone /TX/ in the target language is mapped to phone /TH/ in the base language.

2) 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. For example, BAR – /B AA R/.

The aligned data for /B/ is randomly distributed to obtain the aligned data for /BH/ and /B/.

Gujarati phone (Y)	Gujarati alphabet	h(Y)	p(Y)	Gujarati phone (Y)	Gujarati alphabet	h(Y)	p(Y)	Gujarati phone (Y)	Gujarati alphabet	h(Y)	p(Y)	Gujarati phone (Y)	Gujarati alphabet	h(Y)	p(Y)
AA	Aા	AA	AA	DH	ધ	DH	DH	JH	જ	JH	JH	S	k	S	S
AAN	Aાં	AA	AA+N	DHH	^	DH	DH+H H	JHH	T	JH	JH+HH	SH	i	SH	SH
AE	A	AE	AE	DN	Z	DX	DX+N	K	ક	K	K	T	ધ	T	T
AEN	A	AE	AE+N	DXH	ય	DX	DX+H H + R	KH	M	KD	KD+H H	TH	'	TH	TH
AW	Aૉ	AW	AW	EY	A	EY	EY	L	g	L	L	THH	ધ	TH	TH+H H
AWN	Aૉં	AW	AW+N	EYN	A	EY	EY+N	M	d	M	M	TX	s	TH	TH
AX	A	AX	AX	F	a	F	F	N	“	N	N	UH	D	UH	UH
AXN	A	AX	AX+N	G	N	G	G	OW	Aો	OW	OW	UHN	E	UH	UH+N
B	b	B	B	GH	O	GD	GD+H H	OWN	Aોં	OW	OW+N	UW	D	UW	UW
BH	c	BD	BD+H H	HH	I	HH	HH	P	‘	P	P	UWN	J	UW	UW+N
CH	Q	CH	CH	IH	B	IH	IH	PD	‘	PD	PD	V	h	V	V
CHH	R	CH	CH+H H	IY	C	IY	IY	PH	a	P	PD+H H	Y	e	Y	Y
D	X	D	D	IYN	I	IY	IY+N	R	f	R	R	Z	S	Z	Z

Figure-1: Gujarati phonemes for characters in Gujarati. Mappings are shown using an English phone set. A mapping *h*(y) is from a Gujarati to English phone and p(y) is aligned phone set.

3) PROPOSED APPROACH

We have proposed a new technique for bootstrapping which provides more accurate initial phone models for the target language. We have modified the first approach as described above, so that the aligned speech data for two similar phones in the target language can be easily separated, for example for phones /BH/ and /B/. We propose to use both the phone sets, i.e., the phone sets of base and target languages, to avoid the confusion between the phones in the target language which are mapped to the same phone in the base language.

Figure 2 shows the technique that is used to align Gujarati speech by using an English speech recognition system. A mapping *h*() from a Gujarati phone set denoted by Γ to an English phone set denoted by π is used to generate the pronunciation of Gujarati words by the English phone set.

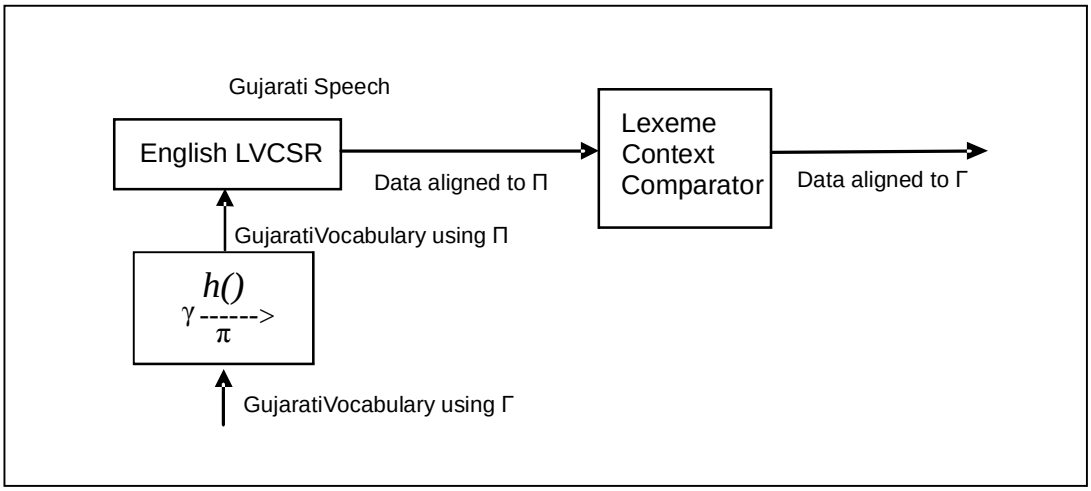

Figure-2: Alignment of the target language data. (LVCSR: large-vocabulary continuous speech recognition.)

Using linguistic knowledge, this mapping is based on the acoustic closeness of the two phones. The mapping is such that each phone $\gamma \varepsilon \Gamma$ is mapped to one and only one phone in Π. A vocabulary created by such a mapping is used to align Gujarati speech data. Since more than one element in Γ may map to a single element in Π, h() is a many-to-one mapping in general and hence cannot always be used in reverse to obtain γ from π. Therefore, in order to recreate the alignment labels with Gujarati phones, an inverse mapping $h^{-1}()$ will not be feasible. A lexeme context comparator is used to generate the correct labels from $\pi \varepsilon \Pi$. This uses the context to resolve the ambiguity which arises from the one-to-many mapping $h^{-1}()$.

To illustrate the requirement of a lexeme context comparator, we take the example of two Gujarati words, cफs and bl The base forms for these words are shown in Table 1. For both words, the alignment would be generated for the phone /B/. However, this /B/ must be replaced by /BH/ if the word is and by /B/ if the word is cफs This information is not available by using the mapping $h^{-1}()$. Therefore, a lexeme comparator is used to examine the lexemes of the words and disambiguate for such cases.

Gujarati Word	Gujarati Baseform	English Baseform
cफs	BH AA R AX TD	B AA R AX TD
bl	B AX HH UH	B AX HH UH

Table-1: Baseforms for two Gujarati words.

The algorithm can be stated in the steps mentioned below:

(i) For a feature vector labeled with a phone $\pi \varepsilon \Pi$, form a subset $\Psi \varepsilon \Gamma$ using the inverse mapping $h^{-1}()$ [since $h^{-1}()$ is a one-to-many mapping in general].

(ii) If Ψ is a singleton, change the label of the feature vector to the element $\gamma \varepsilon \Psi$.

(iii) If not, from the lexeme context of the feature vector, compare the two phonetic spellings of the two lexemes (one written with phones in Π and other with phones in Γ to which this vector belongs. Using this information, handle the disambiguate and choose the phone from Ψ that satisfies the mapping $h^{-1}()$ for the lexeme—for example, /B/ and /BH/.

This technique would generate the aligned Gujarati speech corpus without the need for a Gujarati speech recognizer. Although this alignment may not provide exact phone boundaries, it would serve the purpose of building the initial phone models. The inaccurate phone boundaries are a result of phonetic space differences in the two languages owing to the different acoustic characteristics of the languages. This depends on the two languages; if the languages are acoustically similar, we can have accurate phone boundaries using the above technique. It should be noticed that using the phone set of the target language in the lexeme context comparator not only separates the aligned data for /B/ and /BH/ but also provides the right context information for other phones in the aligned speech corpus. This context information would otherwise have been abused because of the many-to-one phone mapping from target language to base language.

4) PHONE SET MAPPING

The International Phonetic Association (IPA) [16] has defined phone sets for labeling speech databases for sounds of a large number of languages, including Gujarati. However, there are some sounds in Gujarati which are not included in the IPA phone set but are important when building phone models that are to be used for the purpose of automatic speech recognition. In continuous speech recognition tasks, the purpose of defining a phonetic space is to form well-defined, non-overlapping clusters for each phoneme in the acoustic space. This clustering makes it easier for the system to recognize the phone to which an input utterance of speech belongs. For the same number of data and phoneme models, a better phone set is one that gives a higher classification rate and is able to distinguish the words present in the vocabulary of the language. We define a Gujarati phone set which can cover all the different sounds that occur in Gujarati. This phone set takes into consideration the fact that even though Gujarati is a phonetic language, from an acoustic point of view some phones such as plosives have different acoustic properties when they occur at the end of the word. Taking these into account, we have constructed a Gujarati phone set consisting of 61 phones (including the inter-word silence *D*$ and long pause silence *X*) to represent the sounds in Gujarati. It is seen that of these 61 phones, 39 are already present in English. Figure 1 shows the corresponding characters as written in Gujarati script. In the figure, $h(\gamma)$ represents the mapping of Gujarati phones to the corresponding English phones for aligning the Gujarati data using English acoustic models, and $p(\gamma)$ represents the mapping to obtain the initial phone models for the Gujarati phones from English data. In addition to ten English vowels, Gujarati has nine nasalized vowels (AAN, AEN, AWN, AXN, EYN, IYN, OWN, UHN, UWN). Each plosive phone (B, D, K, P, T) has an additional phone (BD, DD, KD, PD, TD) to represent the acoustic dissimilarity when they occur at the end of a word. The bootstrapping approach described in the preceding subsection requires a mapping from the phones of the base language to the phones of the target language. A phone set mapping is defined using the linguistic knowledge of the two languages. We define three categories of mapping as follows:

- *Exact mapping* Some of the phones may be common to both the base and the target language. For example, many vowels

such as /AX/, /AA/, and /IY/ are common to English and Gujarati, and they have an exact mapping from one language to the other. The mappings *h()* and *p()* are the same for such phones.

- *Merging* Some of the phones in the target language may have sounds from more than one phone in the base language. For example, Gujarati has some nasalized vowels such as /AAN/ and /EYN/, which are a combination of the corresponding vowel and nasal sound /N/. For these phones, one-to-many mapping is defined from such Gujarati phones to their English counterparts. For example, the Gujarati phone /GH/ is a combination of the English phones /GD/ and /HH/ while creating the mapping *p()*. The mapping for such phones differs in the case of *h()* and *p()*.
- *Approximation* Some of the phones in the target language may not be present in the base language at all. Such phones are simply mapped to the closest phone in the base language. For example, phone /TX/ in Gujarati (cpfs - BH AA R AX TX) is mapped to phone /TH/ in English (B AA R AX TH). The mappings *h*() and p() are the same for such phones.

5) REFINING PHONE SET MAPPING

We now present a method that is used to improve the initial phone set mapping *p()*. This method is based on a measure of phonetic similarity between the phones in Γ and the phones in Π. One possible measure of similarity is the distance between the phones in the MFCC domain. Each phone of Π is modeled by a normal distribution, and the phonetic distance of a phone $\gamma \varepsilon \Gamma$ from a phone $\pi \varepsilon \Pi$ is defined as

$$D(y,\pi) = \sum_{v_i \in y} \frac{(v_i - m_\pi)^2}{\|\Gamma\|}$$

Where v_i represents a 24-dimensional MFCC vector belonging to y and m_π is the mean vector corresponding π. However, we used a distance measure based on the log likelihood of the phone models in Π for each test vector in $y \varepsilon \Gamma$. The mean of log likelihoods is taken as the measure of acoustic similarity between the phones in the two languages. This measure is calculated for each phone $y \varepsilon \Gamma$ over all of the phones in Π that are considered to be close to y. The mapping *p()* is refined if the acoustic similarity measure shows that a phone y is closer to some phone π' than it is to π, to which it was initially mapped. The log-likelihood-based distance measure produces better results. As a result of the refinement, we changed the mapping of /DDN/ from /DD + HH/ to /DD + R/ and of /DXH/ from /DD + HH/ to /DD + HH + R/.

CONCLUSION

In this paper we have presented novel technique that can be used to build a continuous large-vocabulary Gujarati speech recognition system. A new technique for fast bootstrapping the initial phone models has been presented.

REFERENCES

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

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

[3] O. Anderson, P. Dalsgaard, and W. Barry, (1994) 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, pp. 121–124

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

[5] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam, and A. Verma, (2000) Translingual Visual Speech Synthesis, *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, New York, pp. 1089 –1092.

[6] M. Choudhury, Rule-Based Grapheme to Phoneme Mapping for Gujarati Speech Synthesis, (2003) presented at the 90th Indian Science Congress of the International Speech Communication Association (ISCA), Bangalore.

[7] M. Choudhury and A. Basu, (2002) A Rule Based Schwa Deletion Algorithm for Gujarati, *Proceedings of the International Conference on Knowledge Based Computer Systems*, Mumbai, pp. 343–353

[8] E. Fosler-Lussier, (1999) Multi-Level Decision Trees for Static and Dynamic Pronunciation Models, *Proceedings of the Eurospeech Conference*, Budapest, pp. 459 – 462

[9] A. W. Black, K. Lenzo, and V. Pagel, (1998) Issues in Building General Letter to Sound Rules, *Proceedings of the 3rd European Speech Communication Association (ESCA)*, pp. 77– 80.

[10] A. K. Keinappel and R. Kneser, (2001) Designing Very Compact Decision Trees for Grapheme-to-Phoneme Transcription, *Proceedings of the Eurospeech Conference*, Scandinavia, pp. 1911–1914

[11] J. Suontausta and J. Hakkinen, (2000) Decision Tree Based Text-to-Phoneme Mapping for Speech Recognition, *Proceedings of the International Conference on Spoken Language Processing*, Beijing, pp. 199 –202

[12] F. Mana, P. Massimino, and A. Pacchiotti, (2001) Using Machine Learning Techniques for Grapheme to Phoneme Transcription, *Proceedings of the Eurospeech Conference*, Scandinavia, pp. 1915–1918.

[13] R. W. P. Luk and R. I. Damper, (1992) Inference of Letter–Phoneme Correspondences by Delimiting and Dynamic Time Warping Techniques, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Francisco, pp. 61– 62

[14] L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell, (1991) Automatic Phonetic Baseform Determination, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, pp. 173–176.

[15] B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, (1998) Acoustics-Only Based Automatic Phonetic Baseform Generation, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, pp. 309 –312.

[16] J. Wells and J. House, (1995 *The Sounds of the IPA*, Department of Phonetics and Linguistics, University College London.

[17] L. R. Bahl, P. V. deSouza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, (1991) Decision Trees for Phonological Rules in Continuous Speech, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, pp. 185–188.

[18] N. Mukherjee, N. Rajput, L. V. Subramaniam, and A. Verma, (2000) On Deriving a Phoneme Model for a New Language, *Proceedings of the International Conference on Spoken Language Processing*, Beijing, pp. 850 – 852