Multilingual Speech Recognition for Information Retrieval in Indian context

Udhyakumar.N, Swaminathan.R and **Ramakrishnan.S.K**
Dept. of Electronics and Communication Engineering
Amrita Institute of Technology and Science
Coimbatore, Tamilnadu – 641105, INDIA.
udhay_ece@rediffmail.com,{rswami_ece,skram_ece}@yahoo.com.

Abstract

This paper analyzes various issues in building a HMM based multilingual speech recognizer for Indian languages. The system is originally designed for Hindi and Tamil languages and adapted to incorporate Indian accented English. Language-specific characteristics in speech recognition framework are highlighted. The recognizer is embedded in information retrieval applications and hence several issues like handling spontaneous telephony speech in real-time, integrated language identification for interactive response and automatic grapheme to phoneme conversion to handle Out Of Vocabulary words are addressed. Experiments to study relative effectiveness of different algorithms have been performed and the results are investigated.

1 Introduction

Human preference for speech communication has led to the growth of spoken language systems for information exchange. Such systems need a robust and versatile speech recognizer at its front-end, capable of decoding the speech utterances. A recognizer developed for spoken language information retrieval in Indian languages should have the following features:

- It must be insensitive to spontaneous speech effects and telephone channel noise.
- *Language Switching* is common in India where there is a general familiarity of more than one language. This demands for a multilingual speech recognition system to decode sentences with words from several languages.
- Integrated language identification should be possible which helps later stages like speech synthesis to interact in user's native language.

This paper reports our work in building a multilingual speech recognizer for Tamil, Hindi and accented English. To handle sparseness in speech data and linguistic knowledge for Indian languages, we have addressed techniques like cross-lingual bootstrapping, automatic grapheme to phoneme conversion and adaptation of phonetic decision trees.

In the area of Indian language speech recognition, various issues in building Hindi LVCSR systems have been dealt in (Nitendra Rajput, et al. 2002). For Tamil language (Yegnanarayana.B et al. 2001) attempts to develop a speaker independent recognition system for restricted vocabulary tasks. Speech recognizer for railway enquiry task in Hindi is developed by (Samudravijaya.K 2001).

The paper is organized as follows. In sections 2, 3 and 4 we discuss the steps involved in building a multilingual recognizer for Hindi and Tamil. In section 5 automatic generation of phonetic baseforms from orthography is explained. Section 6 presents the results of adapting the system for accented English. Techniques to incorporate Language Identification are described in Section 7. Finally we conclude with a note on future work in section 8.

2 Monolingual Baseline systems

Monolingual baseline systems are designed for Tamil and Hindi using HTK as the first step towards multilingual recognition. We have used OGI Multilanguage telephone speech corpus for our experiments (Yeshwant K. Muthusamy et al.1992). The database is initially cleaned up and transcribed both at word and phone level. The phoneme sets for Hindi and Tamil are obtained from (Rajaram.S 1990) and (Nitendra Rajput, et

al. 2002). The spontaneous speech effects like filled pauses (ah, uh, hm), laughter, breathing, sighing etc. are modeled with explicit words. The background noises from radio, fan and crosstalk are pooled together and represented by a single model to ensure sufficient training. Front-end features consist of 39 dimensional Melscale cepstral coefficients. Vocal Tract Length Normalization (VTLN) is used to reduce inter and intra-speaker variability.

2.1 Train and Test sets

The OGI Multilanguage corpus consists up to nine separate responses from each caller, ranging from single words to short-topic specific descriptions to 60 seconds of unconstrained spontaneous speech. Tamil data totaled around 3 hours and Hindi data around 2.5 hours of continuous speech. The details of training and test data used for our experiments are shown in Table.1.

Lang	Data	Sent	Words	Spkrs
Tamil	Train	900	7320	250
	Test	300	1700	15
Hindi	Train	125	10500	125
	Test	200	1250	12

Table.1: Details of Training and Test Corpus.

2.2 Context Independent Training

The context independent monophones are modeled by individual HMMs. They are three state strict left-to-right models with a single Gaussian output probability density function for each state. Baum-Welch training is carried out to estimate the HMM parameters. The results of the monolingual baseline systems are shown in Table.2.

Language	Accuracy	
	Word Level	***Sentence Level***
Hindi	49.7%	46.2%
Tamil	50.3%	48.7%

Table.2: Recognition Accuracy of Monophone Models.

The difference in accuracy cannot be attributed to the language difficulties because there are significant variations in database quality, vocabulary and quantity between both the languages.

TAMIL: The recognition result for monophones shows that prominent errors are due to substitution between phones, which are acoustic variants of the same alphabet (eg.ch and s, th and dh, etc.). Hence the lexicon is updated with alternate pronunciations for these words. As a result the accuracy improved to 56%.

HINDI: Consonant clusters are the main sources of errors in Hindi. They are replaced with a single consonant followed by a short spelled '*h*' phone in the lexicon. This increased the accuracy to 52.9%.

2.3 Context Dependent Training

Each monophone encountered in the training data is cloned into a triphone with left and right contexts. All triphones that have the same central phone end up with different HMMs that have the same initial parameter values. HMMs that share the same central phone are clustered using decision trees and incrementally trained. The phonetic questions (nasals, sibilants, etc.) for tree based state tying require linguistic knowledge about the acoustic realization of the phones. Hence the decision tree built for American English is modified to model context-dependency in Hindi and Tamil. Further unsupervised adaptation using Maximum Likelihood Linear Regression (MLLR) is used to handle calls from non-native speakers. Environment adaptation is analyzed for handling background noise.

3 Multilingual Recognition System

Multilingual phoneme set is obtained from monolingual models by combining acoustically similar phones. The model combination is based on the assumption that the articulatory representations of phones are so similar across languages that they can be considered as units that are independent from the underlying language. Such combination has the following benefits (Schultz.T et al.1998):

- Model sharing across languages makes the system compact by reducing the complexity of the system.
- Data sharing results in reliable estimation of model parameters especially for less frequent phonemes.
- Multilingual models, bootstrapped as seed models for an unseen target language improve the recognition accuracy considerably.
- Global phoneme pool allows accurate modeling of OOV (Out Of Vocabulary) words.

International Phonetic Association has classified sounds based on the phonetic knowledge, which is independent of languages. Hence IPA mapping is used to form the global phoneme pool for multilingual recognizer. In this scheme, phones of Tamil and Hindi having the same IPA representation are combined and trained with data from both the languages (IPA 1999).

The phonetic inventory of the multilingual recognizer Ψ_{ML} can be expressed as a group of language independent phones Γ_{LI} unified with a set of language dependent phones Γ_{LD} that are unique to Hindi or Tamil.

$$\Psi_{ML} = \Gamma_{LI} \cup \Gamma_{LDT} \cup \Gamma_{LDH}$$

where Γ_{LDT} is the set of Tamil dependent models.

Γ_{LDH} is the set of Hindi dependent models

The share factor *SF* is calculated as

$$SF = \frac{|\Psi_T| + |\Psi_H|}{|\Psi_{ML}|} = \frac{59 + 43}{70} \cong 1.5$$

which implies a sharing rate of 75% between both the languages. The share factor is a measure of relation between the sum of language specific phones and the size of the global phoneme set. The high overlap of Hindi and Tamil phonetic space is evident from the value of SF. This property has been a motivating factor to develop a multilingual system for these languages. After merging the monophone models, context dependent triphones are created as stated earlier. Alternate data-driven techniques can also be used for acoustic model combination, but they are shown to be outperformed by IPA mapping (Schultz.T et al.1998).

4 Cross Language Adaptation

One major time and cost limitation in developing LVCSR systems in Indian languages is the need for large training data. Cross-lingual bootstrapping is addressed to overcome these drawbacks. The key idea in this approach is to initialize a recognizer in the target language by using already developed acoustic models from other language as seed models. After the initialization the resulting system is rebuilt using training data of the target language. The cross-language seed models perform better than flat starts or random models. Hence the phonetic space of Hindi and Tamil Ψ_{ML} is populated with English models Ψ_E in the following steps.

- The English phones are trained with Network speech database (NTIMIT), which is the telephone bandwidth version of widely used TIMIT database.
- To suit Indian telephony conditions, 16KHz NTIMIT speech data is down-sampled to 8KHz.
- A heuristic IPA mapping combined with data-driven approach is used to map English models with multilingual models. The mappings are shown in Table.3.
- If any phone in Ψ_E maps to two or more phones in Ψ_{ML} the vectors are randomly divided between the phones since duplication reduces classification rate.
- After bootstrapping, the models are trained with data from both the languages.

Hindi	*Tamil*	*English*	*Hindi*	*Tamil*	*English*
आ	ஆ	AA	क	-	KD
आं	-	AA+N	ख	-	KH
ऐ	ஐ	AY	ल	ள	L
ऐं	-	AY+N	म	ம	M
औ	ஔ	AW	न	ன	N
औं	-	AW+N	ङ	ங	NG
अ	அ	AX	ओ	ஓ	OW
अं	-	AX+N	ओं	-	OW+N
ब	ப	B	प	ப	P
ब	-	BD	प	-	PD
भ	-	BD+HH	फ	-	F
च	ச	CH	र	ர	R
छ	-	CH+HH	स	ச	S
ड	ட	D	क्ष	-	K+SH
ड	-	DD	ट	ட	T
द	-	DH	ट	-	TD
ध	-	DH+HH	ठ	-	TH+HH
ढ	-	DX+HH	त	த	TX
ए	ஏ	EY	थ	-	TH
एं	-	EY+N	उ	உ	UH
फ	-	F	उं	-	UH+N
ग	-	G	ऊ	ஊ	UW
घ	-	GD+HH	ऊं	-	UW+N
ह	க	HH	व	வ	V
इ	இ	IH	य	ய	Y
ई	ஈ	IY	ज़	-	Z
ईं	-	IY+N	ण	-	DX+N
ज	ஜ	JH	-	ந	N
झ		JH+HH	-	ஞ	N+Y
क	க	K	-	ண	N
-	ழ	L	-	ற	R
-	ள	L	-	ன	AE

Table.3: Mapping between Multilingual phoneme set and English phones for crosslingual bootstrapping.

The improvement in accuracy due to crosslingual bootstrapping is evident from the results shown in Table.4.

This is due to the implicit initial alignment caused by bootstrapped seed models. The results are calculated for context dependent triphones in each case. The degradation in accuracy of the multilingual system compared to monolingual counterparts is attributed to generalization and parameter reduction.

System	Accuracy
Monolingual_Hindi	95%
Monolingual_Tamil	97.6%
Multilingual	90.3%
Bootstrapped	94.5%

Table.4: Comparison of accuracy for Monolingual, Multilingual and Bootstrapped systems.

5 Grapheme to Phoneme conversion

Direct dictionary lookup to generate phonetic base forms is limited by time, effort and knowledge brought to bear on the construction process. The dictionary can never be exhaustive due to proper names and pronunciation variants. A detailed lexicon also occupies a large disk space. The solution is to derive the pronunciation of a word from its orthography. Automatic grapheme to phoneme conversion is essential in both speech synthesis and automatic speech recognition. It helps to solve the out-of-vocabulary word problem, unlike the case using a soft lookup dictionary. We have examined both rule-based and data-driven self-learning approaches for automatic letter-to-sound (LTS) conversion.

5.1 Rule-Based LTS

Inspired by the phonetic property of Indian languages grapheme to phoneme conversion is usually carried out by a set of handcrafted phonological rules. For example the set of rules that maps the alphabet to its corresponding phones is given below. The letter ப (/p/) in Tamil can be pronounced as '*p*' as in பாடம் or '*b*' as in அம்பு or '*P*' as in லாபம்.

P_Rules:

```
1. {Anything,   "pp",   Anything,    "p h"   },
2. {Nothing,    "p",    Anything,    "p*b"   },
3. {Anything,   "p",    CONSONANT,   "p"     },
4. {NASAL,      "p",    Anything,    "b"     },
5. {Anything,   "p",    Anything,    "P"     }
```

Here * indicates a pronunciation variant. It may be noted that for any word, these context sensitive rules give a phonemic transcription. These rules are ordered as *most specific first* with special conventions about lookup, context and target (Xuedong Huang et al. 2001). For example, the rule 4 means that the alphabet /p/ when preceeded by a nasal and followed by anything is pronounced as '*b*'. These rules could not comprehend all possibilities. The exceptions are stored in an exception list. The system first searches this lookup dictionary for any given word. If a match is found, it reads out the transcription from the list. Otherwise, it generates pronunciation using the rules. This approach helps to accommodate pronunciation variations specific to some words and thus avoiding the need to redraft the complete rules.

5.2 CART Based LTS

Extensive linguistic knowledge is necessary to develop LTS rules. As with any expert system, it is difficult to anticipate all possible relevant cases and sometimes hard to check for rule interface and redundancy. In view of how tedious it is to develop phonological rules manually, machine-learning algorithms are used to automate the acquisition of LTS conversion rules. We have used statistical modeling based on CART (Classification And Regression Trees) to predict phones based on letters and their context.

Indian languages usually have a one-to-one mapping between the alphabets and corresponding phones. This avoids the need for complex alignment methods. The basic CART component includes a set of *Yes-No* questions about the set membership of phones and letters that provide the orthographic context. The question that has the best entropy reduction is chosen at each node to grow the tree from the root. The performance can be improved by including composite questions, which are conjunctive and disjunctive combinations of primitive questions and their negations. The use of composite questions can achieve longer-range optimum, improves entropy reduction and avoids data fragmentation caused by greedy nature of the CART (Breiman.L et al.1984). The target class or leafs consist of individual phones. In case of alternate pronunciations the phone variants are combined into a single class.

5.3 Experiments

Both rule-based and CART based LTS systems are developed for Tamil and evaluated on a 2k word handcrafted lexicon. Transliterated version of Tamil text is given as input to the rule-based system. The performance of decision trees is comparable to the phonological rules (Table.5). We observed some interesting results when the constructed tree is visualized after pruning. The composite questions generated sensible clusters of alphabets. Nasals, Rounded vowels, Consonants are grouped together. The other phenomenon is that CART has derived some intricate rules among the words, which were considered as exceptions by the linguists who prepared the phonological rules. Statistical methods to use phonemic trigrams to rescore n-best list generated by decision tree and use of Weighted Finite State Automata for LTS rules are under study.

LTS System	Word Accuracy	Phone Accuracy
Rule-Based	95.2%	97.5%
CART	96.3%	98.7%

Table.5: LTS results using Rule-based and CART systems on Tamil Lexicon.

The results show that automatic rule generation with CART performs better than manually coded rules.

6 Adaptation for Accented English

English words are more common in Indian conversation. In OGI multi language database, 32% of Hindi and 24% of Tamil sentences have English words. Therefore it is necessary to include English in a multilingual recognizer designed for Indian languages. English being a non-native language, Indian accents suffer from disfluency, low speaking rate and repetitions. Hence accuracy of a system trained with American English degrades significantly when used to recognize Indian accented English. Various techniques like lexical modeling, automatic pronunciation modeling using FST, speaker adaptation, retraining with pooled data and model interpolation are being explored to reduce the Word Error Rates for non-native speakers.

6.1 Speech Corpus

We have used Foreign Accented English corpus from Centre for Spoken Language Understanding (CSLU) and Native American accented Network TIMIT corpus for the experiments. Table.6 gives the details of speech databases used for our study:

Database	Sent	Words	Spkrs
American Accent (N_ENG)	4500	43k	460
Tamil Accent (NN_TAE)	300	3.2k	300
Native Tamil (N_TAM)	900	7.3k	250

Table.6. Details of databases used for Tamil accented English recognition.

6.2 Previous Work

Lexical adaptation techniques introduce pronunciation variants in the lexicon for decoding accented speech (Laura.M.Tomokiyo 2000). The problem with these methods is that the context dependent phones in the adapted lexicon seldom appear in the training data and hence they are not trained properly. Statistical methods for automatic acquisition of pronunciation variants had produced successful results (Chao Huang et al.2001). These algorithms are costlier in terms of memory space and execution time, which makes them difficult to handle real-time speech. In acoustic modeling techniques the native models are modified with available accented and speakers' native language data.

6.3 Experiments

The base line system is trained on N_ENG corpus and is used to recognize NN_TAE utterances. It is a well-known fact that accented speech is influenced by native language of the speaker. Hence we tried decoding NN_TAE data using Tamil recognizer. The lexicon is generated using grapheme to phoneme rules. The accuracy dropped below the baseline, which means that there is no direct relationship between N_TAM and NN_TAE speech. The result is already confirmed by (Laura.M.Tomokiyo 2000).

Careful analysis of the accented speech shows that perception of the target phone is close to acoustically related phone in speaker's native language. As speaker gains proficiency, his pronunciation is tuned towards the target phone and hence the influence of interfering phone is less pronounced. This clearly suggests that any acoustic modeling technique should start with native language models and suitably modify them to handle accented English. Hence attempts to retrain or adapt N_ENG models using N_TAM data have degraded the accuracy. First set of experiments is carried out using N_ENG models. MLLR adaptation and retraining with NN_TAE data increased the accuracy. In the second set of experiments English models are bootstrapped using N_TAM models by heuristic IPA mapping. They are then trained by pooling N_ENG and NN_TAE data. This method showed better performance than other approaches. The comparative results are shown in figure.1.

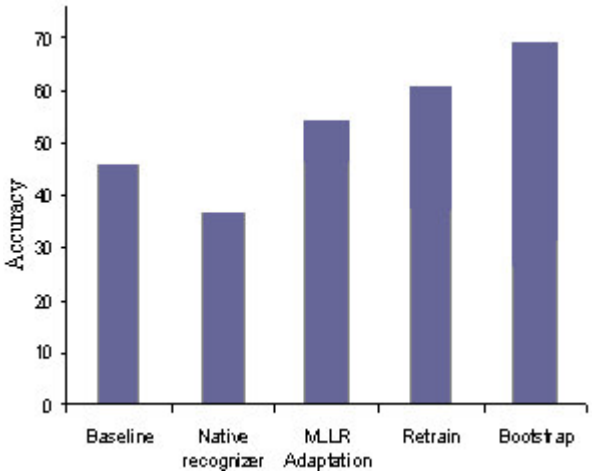

Figure.1: Comparison of Accuracies of acoustic modeling techniques on Tamil accented English.

7 Language Identification

Automatic language identification (LID) has received increased interest with the development of multilingual spoken language systems. LID can be used in Telephone companies handling foreign calls to automatically

route the call to an operator who is fluent in that language. In information retrieval systems, it can be used by speech synthesis module to respond in user's native language. It can also serve as a front-end in speech-to-speech translation. LVCSR based LID can be incorporated in both acoustic and language modeling.

In language independent approach a multilingual recognizer using language independent models is used. Tamil and Hindi bigram models are used to rescore the recognized phone string. The language providing highest log probability is hypothesized. The bigrams for both the languages are evaluated on the transcribed training data. In language dependent approach each phone is given a language tag along with its label. The models are trained solely with data from its own language. Language is identified implicitly from the recognized phone labels. This approach has the advantage of context and text independent language identification. (Lamel.L et al.1994). The results for both the approaches are given in Table.7.

LID System	2s Chunks	5s Chunks
Language Independent	87.1%	98.5%
Language Dependent	91.3%	97.8%

Table.7: Comparison of LID accuracy of language independent and Language dependent systems.

8 Conclusion and Future Work

This work presents the recent results in building a full-fledged multilingual speech recognizer for our ongoing project 'Telephone-based spoken language information retrieval system for Indian languages'. Techniques like CART based LTS, language identification using bigrams and accented English recognition by native language bootstrapping have been experimented.

Significant amount of research remains in handling spontaneous speech effects and nonverbal sounds, which are common in real world data. We have planned to explore language modeling and adaptive signal processing techniques to address these problems (Acero.A 1993). Use of model interpolation and weighted finite state transducers (Karen Livescu 1999) are presently analyzed to improve the system performance on accented English. From our experience we understood the importance of the acronym 'while there is no data like more data, there is also no data like *real* data'. Hence we have started data collection to carry out next phase of experiments.

9 Acknowledgements

The authors wish to thank Mr. C.Santosh Kumar, for his active support, great encouragement and guidance. We gratefully thank Mr.Nitendra Rajput, IBM Research Lab, India and IIT Madras speech group for their valuable help. We also thank Mr.Shunmugom, Linguistic department, Bharathiar university, Coimbatore and Mr.Omkar.N.Koul, Head of faculty of languages, Mussoori for useful linguistic discussions. We would like to acknowledge all the members of Amrita Institute for their enthusiasm in transcribing the data. We also thank Mata Amritananda Mayi for her love and blessings.

10 References

[Acero.A 1993] Acero.A 1993. Acoustic and Environmental robustness in speech recognition, Kluwer Academic Publishers.

[Breiman.L et al.1984] Breiman.L et al.1984. *Classification and regression trees*. Monterey, Calif., U.S.A Wadsworth, Inc.

[Chao Huang et al.2001] Chao Huang, Eric Chang, Tao Chen 2001. Accent Issues in Large Vocabulary Continuous Speech Recognition (LVCSR) Technical Report MSR-TR-2001-69.

[IPA 1999] IPA 1999. *Handbook of the International Phonetic Association,* Cambridge University Press.

[Karen Livescu 1999] Karen Livescu 1999. Analysis and Modeling of Non-Native Speech for Automatic Speech Recognition.

[Lamel.L et al.1994] Lamel.L.F, Gauvain.S 1994. Language Identification using phone-based acoustic likelihoods. In *Proc.ICASSP,*Adelaide, Australia.

[Laura.M.Tomokiyo 2000] Laura Mayfield Tomokiyo 2000. Handling Non-native speech in LVCSR, In *Proc.InSTIL.*.

[Nitendra Rajput, et al. 2002] Nitendra Rajput, et al 2000. A large vocabulary continuous speech recognition system for hindi In *Proc. NCC*, India.

[Rajaram.S 1990] Rajaram.S 1990. *Tamil Phonetic Reader,* Central Institute of Indian Languages.

[Samudravijaya.K 2001] Samudravijaya.K 2001. Hindi Speech Recognition, J. Acoustic Society of India, vol 29, pp 385-393.

[Schultz.T et al.1998] T. Schultz et al.1998. Multilingual and Crosslingual Speech Recognition In *Proc.* DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne, VA.

[Xuedong Huang et al. 2001] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon 2001. *Spoken Language Processing, A guide to theory, Algorithm, and System development,*Prentice Hall.

[Yegnanarayana.B et al. 2001] Yegnanarayana.B and Nayeemullah Khan.A 2001. Development of a speech recognition system for Tamil for restricted small tasks *In Proc. NCC ,* India.

[Yeshwant K. Muthusamy et al.1992] Yeshwant K. Muthusamy et al.1992. The OGI Multi-Language Telephone Speech Corpus In *Proc. ICSLP*.