Tamil Speech Recognition using Semi Continuous Models

Hanitha Gnanathesigar

Informatics Institute of Technology, Sri Lanka,

Abstract- In this paper novel approach for implementing Tamil Language Semi continuous speech recognition based on Hidden Markov Models is discussed. Tamil and other Indian languages share phonological features which are rich in vowel and consonant realizations. The same phone in different words has different realizations. This can be overcome by employing phone-in-context. Therefore triphone models were chosen as suitable sub-word units for acoustic training. The system is trained with speech corpus of 37 Tamil phones. Speech corpus consisted of 0.35 hours of speech. Training was done using Carnegie Mellon University (CMU)'s SphinxTrain acoustic model Trainer. Accuracy of the training is measured by decoding using PocketSphinx.

Index Terms- Speech Recognition, Tamil Phones, Acoustic Model, Hidden Markov Model, Training

I. INTRODUCTION

Speech is human's most efficient mode of communication and is an alternative to traditional methods of interaction with a computer. Beyond efficiency, humans are comfortable and familiar with speech as it is the natural mode of communication. Tamil is a Dravidian language spoken predominantly in the state of Tamilnadu in India and Sri Lanka. It is the official language of the Indian state of Tamilnadu and also has official status in Sri Lanka and Singapore [9].

II. TAMIL PHONOLOGY

Tamil phonology is characterised by the presence of retroflex consonants and multiple rhotics[7]. Tamil phonemes are categorized into vowels, consonants, and a secondary character, the āytam.

A. Vowels

With respect to orthography, vowels occur in their isolated character only in the beginning position of words. In all the other positions, such as medial and final positions, they are realized in the form of a secondary symbol.

TABLE I: Tamil Vowels [5,8]

S. No	**Vowel**	**VL**	**VH**	**VF**	**LR**
1.	அ	s	l	b	-
2.	ஆ	l	l	b	-
3.	இ	s	h	f	-
4.	ஈ	l	h	f	-
5.	உ	s	h	b	+
6.	ஊ	l	h	c	+
7.	எ	s	m	f	-
8.	ஏ	d	c	-	-
9.	ஐ	d	c	-	-
10.	ஒ	s	m	b	+
11.	ஓ	l	m	c	+
12.	ஔ	d	c	-	+

VL Vowel Length (s)hort, (l)ong, (d)ipthong, sc(h)wa, (g)eminate

VH Vowel Height (h)igh, (m)id, (l)ow, (c)losing, (o)pening

VF Vowel Frontness front, mid, back

LR Lip Rounding (+) Yes, (-) No

B. Consonants

There are 18 consonants in Tamil Language.

க ச ட த ப ற
ங ஞ ண ந ம ன
ய ர ல வ ழ ள

However depending on the context certain consonants are pronounced differently increasing the number of consonant phonemes to 25. Nasal consonants ந, ன, ண, ங and ம are pronounced variously based on the environment in which they occur. The consonants with which these nasals occur include த, ப ட and க.

க is pronounced '**g**' after nasal consonants.

Eg: அங்கே

க is pronounced '**h**' between vowels and after ர் and ய்.

Eg: பகல், ஊர்கள்

க is pronounced '**k**' in word initial position and in clusters

Eg: கரை

ச is pronounced '**s**' between vowels and optionally in word initial position

Eg: ஆசை, செவ்வாய்

ச is pronounced 'ch' in word initial position and in clusters

Eg: செவ்வாய், பச்சை

ச is pronounced 'j' after nasal consonants

Eg: பஞ்சு

ட is pronounced D after nasal consonants and between vowels

Eg: கரண்டி, ஓடம்

ட is pronounced t in word initial position and in clusters

Eg: டமாரம், பட்டு

த is pronounced dh after nasal consonants and between vowel

Eg: பந்து, அது

த is pronounced th in word initial position and in clusters

Eg: தமிழ். பத்து

ப is pronounced b after nasal consonants and between vowels

Eg: தம்பி, அபாயம்

ப is pronounced p in word initial position and in clusters

Eg: படி, அப்பா

TABLE II
Tamil Consonants [5,8]

S. No	Consonant	IPA	TC	PA	CV
1.	க	k	p	v	-
2.	க	g	p	v	+
3.	க	h	f	g	-
4.	ங	ŋ	n	v	+
5.	ச	tʃ	f	p	+
6.	ச	s	f	a	-
7.	ச	ɟ	f	p	+
8.	ஞ	ɲ	n	p	+
9.	ட	ʈ	p	r	-
10.	ட	ɖ	p	r	+
11.	ண	n	n	a	+
12.	த	t	p	a	-
13.	த	d	p	a	+
14.	ந	ɳ	n	r	+
15.	ப	P	p	b	-
16.	ப	b	p	b	+
17.	ம	m	n	b	+
18.	ய	j	m	p	+
19.	ர	R	t	u	+
20.	ல	l	m	a	+
21.	வ	v	f	l	+
22.	ழ	L	m	v	+
23.	ள	ɭ	m	r	+
24.	ற	r	t	a	+
25.	ன	N	n	u	+

TC Type of Consonant (n)asal, (p)losive, (f)ricative, appro(x)imant, (t)rill, flap or (t)ap, late(r)al fricative, lateral approxi(m)ant, (l)ateral flap
PA Place of Articulation (b)ilabial, (l)abio-dental, (d)ental, (a)lveolar, p(o)st-alveolar, (r)etroflex, (p)alatal, (v)elar, (u)vular, p(h)aryngeal, (e)piglottal, (g)lottal
CV Consonant Voicing (+) Yes, (-) No, NA Not Applicable

TABLE III
Tamil Phonemes

S. No	ARPABET	IPA	Tamil
1.	AH	ʌ	அ - அம்மா
2.	AA	aː	ஆ - ஆம்
3.	IH	ɪ	இ - இது
4.	IY	i	ஈ - ஈ
5.	UH	ʊ	உ - உலகம்
6.	UW	uː	ஊ - ஊர்
7.	EH	ɛ	எ - எண்பது
8.	EY	əɪ	ஏ - ஏற்றம்
9.	AY	aɪ	ஐ - ஐயோ
10	AO	ɔ	ஒ - ஒரு
11	OH	ɔː	ஓ - ஓடு
12	AW	aʊ	ஔ

13	K	k	க - அக்கா
14	G	g	க - அங்கே
15	HH	h	க - பகல்
16	NG	ŋ	ங - அங்கே
17	CH	tʃ	ச - பச்சை
18	S	s	ச - ஆசை
19	J	ɟ	ச - பஞ்சு
20	NC	ɲ	ஞ - பஞ்சு
21	T	ʈ	ட - பாட்டு
22	D	ɖ	ட - நாடு
23	NX	n	ண - கண்
24	TH	t̪	த - பத்து
25	DH	d̪	த - அது
26	NH	ɳ	ந - பந்து
27	P	P	ப - பத்து
28	B	b	ப - கோபம்
29	M	m	ம - மலை
30	Y	j	ய - கொய்யா
31	RR	R	ர - கரை
32	L	l	ல - பல்
33	V	v	வ - செவ்வாய்
34	Z	L	ழ - தமிழ்
35	LL	ɭ	ள - கடவுள்
36	R	r	ற - கறை
37	N	N	ன - நான்

III. CHOICE OF SUB-WORD UNIT FOR TRAINING

The number of words in Tamil is around 3 lakhs (approx.). Hence maintaining a large vocabulary is also difficult when the system needs to use Tamil[10]. For a language with large vocabulary like Tamil, training all the words adequately is problematic. Also memory requirement grows linearly with number of words. A syllable is a larger unit than a phone since it encompasses two or more phone clusters. These phone clusters account for the severe contextual effects. Tests on measuring accuracy of syllable-based Automatic Speech Recognition (ASR) reveals that the baseline results were much higher than monophone ASR and slightly worse than fine-tuned triphone ASR[2]. For both the phone and word recognition, triphone model reduced word error rate (WER) by about 50% [11]. In this scenario, when the vocabulary is high and speakers are limited, triphone based model is suitable.

IV. TRAINING

Hidden Markov Model based system, like all other speech recognition systems, functions by first learning the characteristics (or parameters) of a set of sound units, and then using what it has learned about the units to find the most probable sequence of sound units for a given speech signal. The process of learning about the sound units is called training. Acoustic models for Tamil language is created using SphinxTrain. SphinxTrain is CMU's open source acoustic model trainer[6]. It consists of a set of programs, each responsible for a well defined task and a set of scripts that organizes the order in which the programs are called.

A. Transcript File

The trainer also needs to be told which sound units you want it to learn the parameters of, and at least the sequence in which they occur in every speech signal in your training database. This information is provided to the trainer through transcript file. In this the sequence of words and non-speech sounds are written exactly as they occurred in a speech signal, followed by a tag which can be used to associate this sequence with the corresponding speech signal.

```
<s> UTAVIYI KAADCHIPADUTTUVATIL TAVARU </s> (utt13)
<s> KURAL KADDUPAATTU URUPADI PEECINAAL MEELMEESIYI KADDUPADUTTA UTAVUM </s> (utt14)
```

B. Control File

This file consists of name of each audio file used for training.

```
utt1
utt2
utt3
utt4
```

C. Dictionary Files

This file maps every word to a sequence of sound units, to derive the sequence of sound units associated with each signal. There are two dictionaries. One in which legitimate words in the language are mapped sequences of sound units and another in which non-speech sounds are mapped to corresponding non-speech or speech-like sound units. Former is the language dictionary and the latter filler dictionary.

```
AAYATTAM      AA Y AH TH TH AH M
ADIVU  AH D AY V UH
ADUTTA        AH D UH TH TH AH

<s> SIL
<sil> SIL
</s> SIL
```

D. Phone List

This tells the trainer what phones are part of the training set. It is made by listing all the above identified ARPABET phones without duplicates and arranged alphabetically.

```
AA
AH
```

AO
AY

E. Language Model

Statistical tri-gram language models were built using the Sphinx Knowledge Base Tool for a corpus of 334 sentences and 85 unique words.

F. Development of speech Corpus

Contemporary speech recognition systems derive their power from corpus based statistical modeling, both at the acoustic and language levels. Corpus is a large collection of written or spoken texts available in machine readable form accumulated in scientific way to represent a particular variety or use of a language [4]. It serves as an authentic data for linguistic and other related studies. Statistical modeling, of course, presupposes that sufficiently large corpora are available for training. For Tamil language such corpora, particularly acoustic ones, are not immediately available for processing[3]. Therefore necessary speech corpora are developed in-house. All the utterances of the transcript files are recorded and corpus is developed based on following parameters.

TABLE IV
Speech Corpus Parameters

Parameter	**Value**
File Type	mswav
File Extension	Wav
Sampling Rate	16 kHz
Depth	16 bits
Mono/Stereo	Mono
Feature File Extension	mfc
Vector Length	13

G. Training of acoustic models with sphinxTrain

It consists of following steps [6].

1. Flat-start monophone training: Generation of monophone seed models with nominal values, and re-estimation of these models using reference transcriptions. This is also called flat initialization of CI model parameters.

2. Baum-Welch training of monophones: Adjustment of the silence model and re-estimation of single-Gaussian monophones using the standard Viterbi alignment process.

3. Triphone creation: Creation of triphone transcriptions from monophone transcriptions and initial triphone training. This step creates CD untied model files and flat initialization of model files.

4. Training CD untied models: Again the Baum-Welch algorithm is iteratively used. This takes 6 – 10 iterations.

5. Building decision trees and parameter sharing: A group of similar states is called a senone. Senone is also called as a tied state. Then the senones are trained.

6. Mixture generation: Split single Gaussian distributions into mixture distributions using an iterative divide-by-two clustering algorithm and re-estimation of triphone models with mixture distributions.

H. *Decoding*

PocketSphinx, CMU's fastest speech recognition system is used to for decoding. It's a library written in pure C which is optimal for development of C applications as well as for development of language bindings. At real time speed it's the most accurate engine, and therefore it is a good choice for live applications. Also it includes support for embedded devices with fixed-point arithmetic. This is built on top of Sphinx3[1]. The results are tabulated in the following table.

TABLE V
Results – Error Rate

Type of Data	**Hours of Training**	**No. of Segments**	**Sentence Error Rate**	**Word Error Rate**
Trained Corpus	0.17 (10 min)	167	81.4% (136/167)	89.9% (283/316)
Test corpus	0.17 (10 min)	133	97.7% (130/133)	100.4% (252/251)
Trained Corpus	0.35 (21 min)	334	1.8% (6/334)	0.9% (6/632)
Test Corpus	0.35 (21 min)	7	57.1(4/7)	46.1%(11/26)

I. *Results*

Word error rate (WER) is calculated as

$$WER = \frac{S + D + I}{N}$$

where

- S is the number of substitutions,
- D is the number of the deletions,
- I is the number of the insertions,
- N is the number of words in the reference.

Word accuracy (WAcc) is calculated as

$$WAcc = \frac{N - S - D - I}{N} = 1 - WER$$

TABLE VII
Results – Word Accuracy

Type of Data	**Hours of Training**	**No. of Segments**	**Word Accuracy Rate**
Trained	0.17 (10 min)	167	10.1%

Corpus			
Test corpus	0.17 (10 min)	133	-0.4%
Trained Corpus	0.35 (21 min)	334	99.1%
Test Corpus	0.35 (21 min)	7	53.9%

V. CONCLUSION

37 phonemes are identified in Tamil Language. Of which 12 are vowels and 25 consonants. Acoustic model training for semi continuous models was performed using SphinxTrain. Results of the Decoding carried out by PocketSphinx shows that the accuracy was higher for trained corpus in compare to test corpus. Though only little amount of Training was performed it is observed that accuracy improved tremendously with increased training.

REFERENCES

[1] Carnegie Mellon University, (2010). *Pocketsphinx*. [Online] Available from: http://cmusphinx.sourceforge.net/wiki/versions [Accessed 16 January 2011].

[2] Hejtmánek, J. A. P., T., (2008). Automatic speech recognition using context-dependent syllables. *9th International PhD Workshop on Systems and Control: Young Generation Viewpoint.* Izola, Slovenia.

[3] Ganesh, K. M., Subramanian, S.(2002). Interactive Speech Translation in Tamil. College of
Technology, Peelamedu.

[4] Ganesan, M. (n.d). Tamil Corpus Generation and Text Analysis. Annamalai University

[5] IPA, (2005). "The International Phonetic Association (revised to 2005) IPA Chart." [Online]. Available: http://www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf

[6] Singh, R. (2000). SphinxTrain Documentation [Online]. Available at <http://www.speech.cs.cmu.edu/sphinxman/scriptman1.html> [Accessed 02 January 2011].

[7] Schiffman, Harold F.; Arokianathan, S. (1986). "Diglossic variation in Tamil film and fiction". In Krishnamurti, Bhadriraju; Masica, Colin P.. *South Asian languages: structure, convergence, and diglossia*. New Delhi: Motilal Banarsidass. pp. 371–382. ISBN 8120800338. at p. 371

[8] Schiffman, Harold F.; Arokianathan, S. (1999). "A reference grammar of spoken Tamil" [Online]. Available :http://books.google.com/books?id=Oqe-QsaZnnQC&lpg=PP1&pg=PP1#v=onepage&q&f=false

[9] Thangarajan, R., Nagarajan,A.M., Selvam, M., (2008). Word and triphone based approaches in
continuous speech recognition for Tamil language. WSEAS Transactions on signal processing, 4, 76-85.

[10] Thilak, R. A., Madharaci, R. (2004). Speech Recognizer for Tamil Language. Tamil Internet
2004,Singapore.

[11] Lee, K., (1990). Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition., Carnegie Mellon University.

AUTHORS

Hanitha Gnanathesigar, BSc (Hons) Software Engineering, Informatics Institute of Technology (IIT), Sri Lanka, ghanitha@gmail.com.