Perception Experiments for Effective Unit replacement for Tamil TTS

A G Ramakrishnan[#] and Laxmi Narayana M[$]

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, INDIA
ramkiag@ee.iisc.ernet.in[#], mln4u_ece@yahoo.co.in[$]

Abstract

We report our efforts in handling situations in Text to Speech Synthesis, where a particular phonemic or syllabic context is not available in the corpus. The idea is to replace such context by another one which is 'similar'. The 'similarity' of phones or syllables lies in the inability of listeners to distinguish them when placed in a particular context. Such phones were found linguistically in two south Indian languages - Tamil and Telugu, by performing listening tests and acoustically, through a phone classification experiment with Mel Frequency Cepstral Coefficients as features. Maximum likelihood classifier is used to find the most misrecognized phones. Both frame level and phone level classifications were performed to find out such phones. The classification experiments were performed on Tamil corpus of 1027 sentences and on TIMIT corpus.

1. Introduction

Text to Speech (TTS) synthesis is an automated encoding process which converts a sequence of symbols (text) conveying linguistic information, into an acoustic waveform (speech). A concatenative speech synthesis system uses the actual human speech as the source material for synthesizing speech. One of the characteristics based on which a TTS system is evaluated is its ability to produce an intelligible speech. The intelligibility of the synthetic speech depends on the selection of relevant syllables for concatenation, which match the target *context*. Even though the speech corpus covers all the phones in the language under consideration, it may not have all the phonetic contexts. Using individual mono-phones for concatenation results in discontinuities of pitch, energy and lack of coarticulation, leading to unnatural speech.

Speech synthesis based on syllables seems to be a good possibility to enhance the quality of synthesized speech compared to mono-phone or diphone-based synthesizers. This consideration is based both on the fact that more coarticulation aspects are included in syllable segments compared to diphone units and on the fact that the main prosodic parameters (pitch, duration, amplitude) are closely connected to syllables (Kopeck, Pala, 1998). So, not only the presence of a phone in the database is important, but the syllable in which the phone is present and the context in which the phone or syllable is present are important. Mono-phones are considered for concatenation only in the worst case.

2. Goal of the paper

The goal is to identify the phones whose perception is more or less similar i.e, a phone, which when replaced by another phone in that particular context, should not make much difference in perception; the listener shouldn't be able to distinguish. The knowledge of these phones can be used in synthesis. Section 3 further presents our motivations for conducting this kind of experiment. Section 4 describes the phone perception experiments carried out over telephone in languages Tamil, Telugu and English and the corresponding results. Section 5 describes the frame and phone level classification experiments performed on the Tamil and TIMIT databases. Mel Frequency Cepstral coefficients (MFCC) are used as features with Maximum Likelihood (ML) classifier for classification. Results are discussed in Section 6. Section 7 presents the conclusion.

3. Motivation

There are 12 vowels and 18 consonants in Tamil language. There are five other phones introduced for representing Sanskrit. The language has certain well defined rules which introduce seven other phones depending on the presence of consonants with respect to the vowels or the other consonants. Hence there are 42 phones in the language. If we consider phonetic contexts, any one of the 42 phones could occur between any two phones. So there are 42^3 contexts for each phone. If we take the combination of a vowel and a consonant as a syllable, then we get around 216 syllables. If we consider the phonetic contexts of each syllable, each syllable can occur between two other syllables. So for a syllable, there are 216^3 possible contexts for its occurrence. Some of them may not be valid, but the issue is, practically, for any database, it is not possible to cover all such phonetic contexts. So, while synthesis, if a 'syllable in a particular phonetic context' is not available in the inventory, another syllable by whose substitution, the listener may not be able to perceive any difference, can be considered for concatenation.

In continuous speech, a listener may not pay attention to each and every phone the speaker speaks. While speaking on telephone, sometimes, the person on the other side, who naturally never listens to each and every phone, may not exactly recognize all the words we speak. Sometimes, his prior knowledge of the words and the context makes him understand our speech. Sometimes, we might have to repeat some words or syllables, even though the phone conversation takes place in a less noisy environment. The present paper reports the perception experiments conducted to find out such

phones**.** A phone classification experiment is also performed to find out the phones by extracting MFCCs and classifying with ML classifier. The results are used in our Tamil Text to Speech synthesis. The Tamil database under consideration contains 1027 sentences, which are phonetically rich but as mentioned before, it may not contain all the contexts and this is the motivation for this experiment.

4. Perception Experiments

Listening experiments are conducted over the telephone to capture the most 'confused' phones in Tamil language. One person calls the other person and pronounces a list of 152 phones/ syllables (combination of consonant and vowel) in Tamil language and the person on the other side writes the phones *whatever he/she listens to*. Repetition of phones by the speaker is not allowed. The listener is asked to write the phone whatever he/she listens for the first time. Individual phones are chosen to find the exact confusion between phones; if words are chosen, the listener who has a prior knowledge of the word writes the word correctly even though he may have not listened properly or the word is not pronounced properly. This does not serve the purpose.

The experiment is conducted with 10 pairs of native Tamil people. On an average, 30% of the phones are wrongly identified as other phones. Another set of experiments are conducted over 2 pairs on the misrecognized phones. Not much improvement in recognition accuracy is identified. A consistency is found in the misidentification over the speaker-listener pairs. Most of the nasals are wrongly identified as other nasals. Many of the long vowels (*deergha* phones, e.g., 'A' in the English word 'call') are identified as short vowels (*hrasva* phones, e.g., 'a' in the English word 'at') and vice versa. There are two kinds of /r/ phones in Tamil language - /r/ and /R/. They are misrecognized for each another. There are three types of /l/ in Tamil - /l/, /L/ and /zh/. They are confused among themselves. Many times, the vowels like /i/, /u/ are identified as combination of a consonant and vowel - /yi/, /wu/. However, the misrecognition 'between' different groups of phones like vowels, nasals, fricatives, glides is relatively less compared to the misrecognition 'within' the groups. The reason for the misidentification 'between' some rare groups may be due to the inattentiveness of the listener and they can't be taken as similar phones which can be replaced by each another. Their frequency is also very less compared to the other prominent misidentifications. The entire set of phones and the consistently and frequently misrecognized phones are listed in Figure 1 and Table 1 respectively. Table 2 gives one example Tamil word each, for the uncommon phones in Table 1.

Similar experiment was conducted on phones of Telugu language over 8 speaker-listener pairs with 52 phones/syllables. 5 pairs over telephone and 3 pairs sitting some distance apart. In the later case, one person sits at one place and loudly utters the phone/syllable and the other one listens and writes. This did not show any difference in the recognition accuracy. It was found that the recognition depends on the clarity of pronunciation of the speaker and the attentiveness of the listener. The misidentification rate in Telugu language is less compared to Tamil language. On an average, 10 phones out of 52 were misrecognized. That comes to a 20% misidentification whereas it is around 30% in Tamil language. This is because the number of phones in Telugu language is more compared to that of Tamil; Telugu people are habituated to pronounce and to listen to more number of phones, which is not the case with Tamil people. Nevertheless, some similarity is found between the misidentified phones in the two languages. The most frequently and consistently misidentified phones in Telugu are shown in Table 1. The common misidentifications between the two languages are shown in bold in Table 3.

Similar experiment is conducted in English over 7 speaker-listener pairs with 26 letters of the alphabet. The most confused letters are shown in Table 4. The results are not so significant. The misidentification appears to be due to the presence of a common vowel in the pair (e.g, /J/-/A/, /J/=/j/+/a/).

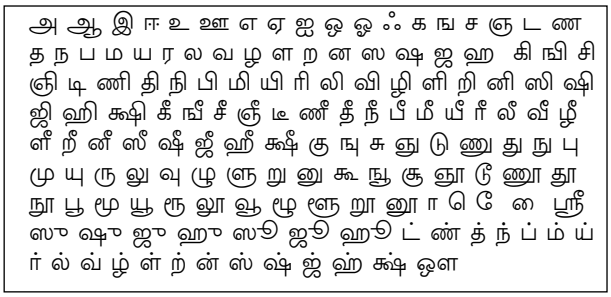

Figure 1: List of Tamil Phones used

ng - n	A – a	i - yi
ny –n	I – i	u – wu
N – n	U - u	L - l
ng - ny	**S - s**	L – zh
R - r		

Table 1: Most confused phones in Tamil. The pairs shown in bold are common misrecognitions between Telugu and Tamil. /ng/, /ny/, /N/, /n/ are the respective nasals of /k/, /ch/, /T/, /t/ groups.

/ng/ - வங்கி	/ny/ - காஞ்சி	/n/ - வானம்
/N/ - மணல்	/r/- வாரம்	/R/ - அறம்
/L/ - வள்ளி	/zh/ - அழகு	

Table 2: Example words of the phones

d – dh	**R - r**	**ng – ny**
s – S	Ksha – cha	T- p

Table 3: Phones identified consistently and frequently for one another in Telugu over 8 Speaker-listener pairs.

B - D	G – D	V – B
J - A	P - T	Q - U

Table 4: Misrecognized phones in English

5. Phone Classification Experiment

After identifying the phones, which are recognized wrongly for other phones, we took the next step of classifying the phones using Maximum Likelihood Classifier. The Tamil phones from our Tamil Corpus and the English phones from the TIMIT Corpus are classified.

5.1. Tamil Corpus

The Tamil database used for the experiment consists of 1027 sentences from a single male speaker, sampled at 16 kHz. The sentences were segmented and labeled manually using Pratt software.

TIMIT Corpus is also used to train and evaluate the phone recognizer for finding the misrecognized phones in English language and for checking the validity of the designed Maximum likelihood classifier for speaker-independence.

5.2. TIMIT Corpus

The TIMIT corpus consists of 6300 sentences spoken by 630 speakers, each speaker speaking 10 sentences sampled at 16 kHz.

Each speaker has

- 2 "SA" sentences, which are the same across all speakers
- 3 "SI" sentences, which were randomly selected by TI.
- 5 "SX" sentences, which were read from a list of 450 phonetically balanced sentences selected by MIT (Jinjin et al, 2002).

The TIMIT corpus has phone labeling, which makes it a useful database for phone classification.

S. No	Training data size	Avg. No. of FV per class	BCCA	Accuracy
1	100	6295	61%	47%
2	200	6938	65%	49%
3	400	8648	72%	53%
4	700	10603	74%	53%

Table 5: Phone level Classification results on Tamil corpus with Full Covariance matrix

Number of sentences used for testing: 100

Variable: Average number of Feature Vectors per class

S.No	Test data size	BCCA	Accuracy
1	50	73.5%	52%
2	200	72.6%	51.2%
3	400	71.73%	50.5%

Table 6: Phone level Classification results on Tamil corpus with Full Covariance matrix

Number of sentences used for Training: 700

Variable: Number of sentences used for Testing

5.3. Feature Extraction – Training

The traditional filter-bank approach (Molau et. al, 2001) is followed for extracting Mel Frequency Cepstral coefficients (MFCCs) from the speech signal. The process is very briefly presented here. The speech waveform, sampled at 16 kHz, is first divided into a number of overlapping segments (windows), each 20 ms long and shifted by 10 ms. A Hamming window is applied and the Fourier transform is computed for each frame. The power spectrum is warped according to the Mel-scale in order to adapt the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands by means of the Mel filter-bank which typically consists of overlapping triangular filters. Discrete cosine transformation (DCT) is applied to the logarithm of the filter-bank output which gives the MFCC vector of the frame. The highest cepstral coefficients are omitted. Now each 20 ms frame is represented by a 12-dimensional acoustic vector. The training data is converted to frame level data and *.feat* files which store the MFCC vectors of all the frames of the corresponding phone are created. Mean and covariance are obtained for all the .feat files.

5.4. Testing

Two types of classifications are performed frame level classification and phone level classification. In the former case, a single 10 ms frame (a 12 dimensional acoustic vector) is classified to one of the 48 (Tamil) phone classes using the Maximum Likelihood (ML) classifier (Duda et. al, 2001). In the later case, mean of all the MFCC vectors belonging to one phone is taken and classified using ML classifier. The idea behind doing this is to represent a phone with a single acoustic vector. In the case of frame level classification, a single frame does not represent a phone. Of course, this is the method mentioned in the literature to test the efficiency of a classifier, but the focus of the present experiment is to find the phones which can be used interchangeably in some contexts and there is a need to further filter the misclassified phones. So phone level classification is also done.

6. Results and Discussion

Phones are classified using full covariance matrix and diagonal covariance matrix. The classification accuracy obtained with the full covariance matrix is better compared to that obtained with the use of diagonal covariance matrix. Sentences from 'dr2' of TIMIT corpus are used to train and test the classifier. The experiments are carried out for different sizes of training and test data and the phones that are misclassified are noted down. We experimentally find that the phone level classification is better with Tamil database whereas frame level classification is better for TIMIT database. Accordingly, the results of phone level classification of Tamil corpus are presented in Tables 5 and 6 and the results of frame level classification of TIMIT corpus are presented in Tables 7, 8 and 9.

6.1. Classification Accuracy

There are 6 broad categories of phones in Tamil – Vowels (a, A, i, I, u, U, e, E, ae, o, O), Semivowels & Glides (y, r, R, l, ll, wl, Ll, L, zh, w, yl), Stops (k,T, t, p, b, kl, Tl, tl, pl, g, D, d, TR), Affricates (cl, j), Fricatives

(S, s, h), Nasals (m, n, ng, ny, N, nl, ml, Nl). BCCA (Broad Class Classification Accuracy) is the accuracy of correctly classifying a phone to its major category. For example, if a vowel is identified as vowel, a nasal is identified as a nasal and so on, the classification is considered to be accurate. The overall accuracy in the fifth column of Table 5 is the accuracy of classifying a phone to its true class. Both of them are found to increase with the training data size. When the training data size is kept constant and the test data size is varied, a slight decline in the accuracies with the increase of test data size is observed. The accuracies in case of TIMIT corpus are listed in Tables 7 to 9. Table 10 shows the Intra category classification accuracy of Tamil phone classes when the training data size is 700 and test data size is 100. For example, 86.4562% of the vowels are classified as vowels (see second entry in column 2 of Table 10).

S. No	Training data size	Avg. No. of FV/class	BCCA	Accuracy
1	200	969	75.91%	47.5%
2	400	1961	75.69%	48.45%
3	560	2762	75.34%	49.78%
4	760	3740	76.1%	50.4%

Table 7: Frame level Classification results on TIMIT database with Full Covariance matrix
Number of sentences used for Testing: 100
Variable: Average number of Feature Vectors per class

S.No	No of speakers	BCCA	Accuracy
1	4	75.1 %	47.24 %
2	10	75.65 %	49.71 %
3	15	75.67 %	50.59 %
4	20	76.1 %	50.3 %
5	26	76.48 %	50.16 %

Table 8: Frame level Classification results on TIMIT database with Full Covariance matrix
Variable: Number of speakers (2 sentences per speaker)

Sl. No	Training data size	Avg. No. of FV/class	BCCA	Accuracy
1	200	969	75.52%	45%
2	400	1961	72.11%	46.31%
3	560	2762	71.77%	45.44%
4	760	3740	71.92%	45.95%

Table 9: Frame level Classification results on TIMIT database with Diagonal Covariance matrix
Variable: Number of feature vectors per class
Number of sentences used for testing: 100

6.2. Confusion matrix

A confusion matrix of the Tamil data for the significant mismatches is shown in Table 11. The classification accuracy of the phone /a/ is relatively very high compared to that of the other phones, in both Tamil and English. Consistently, for all the cases, 25% of the 'a' phones are classified to 'A'. 72% of the /I/ phones are classified as /i/. This is not so prominent with the other vowels. So, if a *deergha* syllable ([consonant A/I] or [A/I consonant]) is not available in the corpus in a particular context, it can be replaced with the *hrasva* syllable ([consonant a/i] or [a/i consonant]). This is a major finding. The confusion between /u/ and /U/ pairs is frequent in the listening tests but not so significant in the classification test. The following results are in the case where the training data size is 700 and test data size is 400. 40.9% of /ae/s are classified to /i/ while only 29.8% of /ae/s are correctly classified to /ae/ class. 9% of /ae/s are classified to /yl/ (genitive of /y/). 38% of /yl/s are classified to /i/. 11.4% of /yl/s are classified to /ae/. There is more misclassification among the three phone classes - /i/, /yl/ and /ae/. 44.44% of /ll/s are classified to /Ll/.

Indian languages are largely phonetic in nature and English is not at all a phonetic language. The phones in the TIMIT corpus and the Tamil corpus are different. So, there are no common confused phones between the two languages. The confusion matrix of English phones is shown in Table 12.

6.3. Application to TTS

The knowledge of the phones usually misidentified is used in Speech synthesis. Blind listening tests are conducted with 4 native Tamil people. The listeners are asked to listen to a set of 11 synthesized sentences which are generated by our Tamil TTS system. The same 11 sentences are also synthesized with some phones replaced by the corresponding confused phones found, in some words. Many words had a single phone replacement and some of them also had 2 to 3 phone replacements. The original phones and the phones with which they are replaced are shown in Table 13. The listeners are asked to write the synthetic sentences of both the sets separately. The results are checked to find the validity of the phone replacement. 75% of the words for which phone replacement is done are recognized as the regular words by all the listeners. They could get the original word even though some of the phones are replaced by other phones in those words. 3 listeners did not notice a change in 50% of the remaining 25% (phone replaced) words. The most common replacements which the listeners didn't make out are shown in bold in Table 13. Some special replacements which are more language specific are shown in Table 14. The phonetic transcription of the words is shown. The IPA codes of the phonemes can be found in (Link).

The replacement of a vowel by consonant-vowel combination at the beginning of the words (e.g., i-yi) and the replacement of phones like nasals, liquids with their corresponding confused phones (e.g., /m/ - /n/ and /l/ - /zh/ respectively) at the end of words worked favourably. Interchanging of the nasals /n/ and /N/ worked at all places. Interchanging /m/ and /n/ at the beginning of the words also worked nicely. Replacement of /l/, /L/ and /zh/ among themselves worked well always. Replacement of /r/ with /R/ was good to some extent. Deergha – Hrasva replacement works well at all places since the listener who has a prior knowledge of the word gets the word right even if there is a lengthening of some vowels.

Class	Intra category classification Accuracy
Vowel	86.4562 %
Semivowels & Glides	62.8213 %
Stops	67.6737 %
Affricatives	57.6000 %
Fricatives	62.1212 %
Nasals	46.6142 %

Table 10: Intra category classification accuracy of Tamil phones

		True Class							
		A	A	i	I	ae	l	ll	yl
Assigned Class	a	3164	166	180	3	65	6	33	74
	A	1112	1461	0	0	0	4	2	0
	i	228	0	1962	110	419	2	6	407
	I	1	0	9	7	1	0	0	1
	ae	112	0	220	11	305	0	0	122
	l	0	0	0	0	0	0	0	0
	ll	0	0	0	0	0	0	12	0
	yl	61	0	130	7	92	0	0	378
	Total	5788	1633	2909	148	1023	83	369	1069

Table 11: Confusion matrix of most confused Tamil phones

		True Class										
		'jh'	's'	'z'	'm'	'n'	'ng'	'iy'	'ih'	'ey'	'y'	Total
Assigned Class	'jh'	71	15	15	0	2	0	12	3	4	4	246
	's'	8	2789	**702**	0	1	0	0	0	0	0	4312
	'sh'	156	267	92	0	0	0	1	0	0	0	2178
	'z'	0	108	277	2	5	0	2	5	1	0	508
	'm'	0	3	1	297	145	32	17	23	4	3	1116
	'n'	0	2	6	**241**	800	**132**	103	92	22	12	2639
	'ng'	0	0	0	0	4	27	4	6	6	0	104
	'iy'	3	1	3	44	136	54	1796	393	364	308	4502
	'ey'	0	0	0	2	4	8	44	100	224	4	680
	'ae'	0	0	0	11	27	19	45	60	166	7	3468
	'y'	0	0	1	0	4	0	30	7	0	47	130
	Total	264	3368	1191	901	1469	325	2590	1412	1076	431	0

Table 12: Confusion matrix of the most confused English phones

7. Conclusion

A novel way of systematic replacement of missing phones has been proposed for speech synthesis. The most confused Tamil phones which can be replaced by one another in specific contexts at the time of synthesis, if they are not available in the corpus, are found. The confused phones in Tamil are identified by conducting listening tests over telephone and also by the phone classification experiment using Maximum Likelihood classifier. The confused phones in Telugu are found by perception tests and in English by the phone classifier. The common confused phones over the two Indian languages are identified. The knowledge of the Tamil confused phones is incorporated in Tamil Text to speech synthesis and experiments show that the proposed phone replacement strategy is fairly successful. The confused phones in English are also found whose knowledge can be used for future purposes of developing bilingual TTS.

Original word	Word after phone replacement
a **g n** i	a **k N** i
E **w u** g a n ae	e **u** g a N ae
u l a **g a ng g** a L ae y u m	w u l a **g a m** g a L ae y u m
m u **kl i** y a	m u **k** y a
A y **w u** kl U D a m	A y **u** kl U D a m

Table 14: Language specific words, before and after phone replacement.

ae – e y	**n – N**	**l – zh**
m - n	**I – i**	**u – wu**
tl - Tl	**e - E**	**i – y i**
L - l	**R - r**	p - w
b - p		

Table 13: Phone replacements done during synthesis.

8. References

Duda, Hart, Stork, 2001, *Pattern Classification*, John Wiley & Sons, Second edition.

Jinjin Ye, Richard. J. Povinelli, Michael T. Zohnson, 2002, Phoneme Classification using Naive Bayes Classifier in Reconstructed Phase Space, *Proc. IEEE Signal Processing Soc. 10th Digital Signal Processing Workshop*, October, pp. 37-40.

Kopeck, I., Pala, K, 1998, Prosody Modelling for Syllable- Based Speech Synthesis, *Proceedings of the IASTED Conference on AI and Soft Computing*, pp.134-137.

Molau. S, M. Pitz, R. Schlüter, and H. Ney, 2001, Computing mel-frequency cepstral coefficients on the power spectrum, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, June, pp.73–76.

Link (http://en.wikipedia.org/wiki/Tamil_script)