Studies on Natural Variability in Human Speech and Perception for Enhancing the Quality of Synthetic Speech

A G Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore, India
ramkiag@ee.iisc.ernet.in

Laxmi Narayana M
TCS Innovation Lab – Mumbai
TCS, Yantra Park, Thane (West)
Maharastra, India
laxmi.narayana@tcs.com

Abstract

Speech production is not deterministic, in the sense that a human being does not always speak the same sentence/ phrase/ word in an identical manner whenever she/he speaks. So, why should a text to speech (TTS) synthesizer produce a similar kind of speech for a given input text, all the time? It would be great, if the synthesizer is made to speak with slightly different characteristics at different times for the same input text. In this paper, we report our studies on the natural variability in the parameters of human speech viz., pitch, amplitude and duration, when same speech is spoken at different times by a single speaker. Also we report on the perception experiments we conducted to find the most confused pairs of phones in Tamil language that can be substituted by each other in specific phonetic contexts. We discuss how the knowledge of variation in different parameters of human speech when same speech is spoken at different times and the knowledge of confused phones, may help in making a TTS system speak with slightly different characteristics at different times for the same input text. The output of such a synthesizer may be perceptually better and appear less monotonous.

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). The two major components of a TTS synthesizer are - natural language processing (NLP) module, which produces a phonetic transcription of the given text and a digital signal processing module, which transforms this phonetic transcription into speech (Thierry Dutoit, 1997). A concatenative speech synthesis system uses pre-recorded human speech as the source material for synthesizing speech. In general, any text to speech synthesizer, for a particular input text, concatenates the speech segments according to the algorithm(s) it adopts and produces a similar synthetic speech all the time. In other words, a sentence/word or phrase is synthesized by a TTS synthesizer in a similar manner all the time; we don't get different realizations from the synthesizer at different times for the same input text.

However, each time a human speaks the same sentence/phrase/word, it is unique. Whatever are the sources of expressive content in speech, they are changeable parameters. The different realizations from the speaker vary even within words, not just from one complete utterance to the next. Also, no two human beings speak alike. Probably, this is the reason why human 'speech' is considered to be a reliable biometric in personal identification systems. And perhaps, it is also why any TTS output looks monotonous, after sometime.

Study on the inherent variability of human speech has its application in many research and development related areas like speech recognition, speaker recognition and speech synthesis. The focus of the present work is directed towards making the TTS synthesizer more natural by speaking

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

slightly differently (with different characteristics) at different times for the same input text. The task before us is to introduce some novelty in the process of synthesizing speech, so that the output of the synthesizer has characteristics akin to human speech and hence, is perceived to be more natural.

In TTS, in general, the waveform is generated as follows: the corpus is searched for the availability of units that match the left and right phonetic contexts of the target. Later, among those units, the one which gives the minimal join cost[1] is selected. If the join cost is greater than a threshold, the context constraints are relaxed to only left or right phonetic context and whichever unit gives the minimal join cost is selected for concatenation. By following this or any other process of synthesizing speech, the output would be the same for a particular text and for a particular synthesizer. A sensible question may be posed in this kind of scenario. Wouldn't it be more natural for a TTS synthesizer to speak differently at different times for the same input text like a human being does? For accomplishing this task, it is required to know how a human being's speech varies at different times when she/he speaks the same sentence/ phrase/ word. This paper reports the study conducted to analyze such variability in Tamil[2] speech and the perception experiments conducted to find the most confused phones in Tamil that can be substituted by each other in specific phonetic contexts. The knowledge of human speech variability and confused phones can be used to induce naturalness in the synthetic speech by making the synthesizer speak differently at different times for the same input text.

The rest of the paper is organized as follows. Section 2 discusses the different parameters that vary in human speech for a single speaker and across different speakers, as mentioned in the literature and the parameters of human speech analyzed in the present work. Section 3 reports the experiments conducted to detect the variation in parameters of human *speech* when same sentence is spoken several times by a single speaker. Section 4 gives the details of the perception experiments conducted to find the natural variability in human *perception* and the detected confused phones in Tamil. Section 5 summarizes the inferences and conclusions and lists all the techniques adopted to induce naturalness in TTS synthesizer. Section 6 gives the conclusion and possible future directions for the work

2 Variable parameters in human speech

The uniqueness in different utterances of the same sentence by the same speaker is a result of differences in pitch, loudness and other subtle articulatory inconsistencies that constitute an individual's idiolect, or characteristic way of speaking, such as different degrees of vowel nasalization. There are still other sources of speech variability besides a speaker's idiolect. In particular, vocal organs vary in size and shape, depending upon such factors as age and sex. The fundamental frequency, perceived as pitch, varies significantly among males, females, and children; during puberty the vocal folds lengthen, resulting in a noticeable change in voice quality (David Hirtle, 2004).

Intonation naturally accompanies human speech to such a degree that it is a prime consideration (and challenge) in speech synthesis. One of the parameters that distinguishes two utterances of the same text by a single speaker is the stress associated with the different syllables. Fundamental frequency (F0) has been the most common acoustic variable studied. As mentioned in (Cairns et al, 1994), F0 rises in stressful conditions. F0 changes are smooth in normal speech, while it could be erratic in stressed speech. Intensity, duration and fundamental frequency are believed to be the primary acoustic cues that vary prominently in human speech. Therefore, they are used as the main acoustic features in the stress detection task in some studies (Wightman and Ostendorf, 1994).

For the present work, the parameters of human speech analyzed are **Pitch, Amplitude and Duration**. Hereafter, the combination of these three parameters of a phone will be called as PAD. Phones in a speech utterance are analyzed for their PAD values. "A phone has a particular PAD" means that its pitch is 'P' Hz, its amplitude is 'A' dB and it exists for 'D' seconds. P=0 for an unvoiced phone.

[1] Join cost or concatenation cost measures the mismatch in concatenating two units from the synthesis database to create a joined target; it takes care of the quality of concatenation between speech segments.

[2] Tamil is the official language of the south Indian state of Tamil Nadu, and also of Singapore, Sri Lanka and Mauritius. In addition to the above countries, it is spoken in Bahrain, Malaysia, Qatar, Thailand, United Arab Emirates and United Kingdom (Ramakrishnan & Laxmi Narayana, 2007).

3 Experiments conducted to detect the natural variability in human speech

Eight native Tamil people are asked to speak 10 Tamil sentences, at 10 different times over a period of 3 days. The time gap between two recordings of one speaker is ensured to be at least 3 hours. The sampling rate is 16 kHz. These sentences are manually segmented and labeled using PRAAT software. Out of the 10 sentences, one that has the maximum number of phones is selected for the initial analysis. The Tamil sentence and its phonetic transcription are shown in Fig 1.

இந்தத் திட்டங்களை ஏற்றுக் கொண்டால் சுதேச மன்னர்கள் நிம்மதியாகவும் பாதுகாப்புடனும் வாழலாம் என வெல்லஸ்லி பிரபு அறிவித்தார்.

> # i n d a tl i Tl a ng g a L ae # E TR u kl o N D A l # s u d E s a # m a nl a r h a L # n i ml a d i y A g a w u m # p A d u g A pl u D a n u m # w A zh a l A m # e n a # w e ll a s l i # p i r a b u # a R i w i tl A r <

Figure 1: The Tamil sentence selected for the study.

Table 1: Pitch statistics of phones in the last 3 words, from the analysis of 10 repeated utterances of the same sentence by one female speaker; Units = Hz

Phone	Mean	SD	% dev
'w'	200.5	13.3	6.6
'e'	200.8	11.5	5.7
'll'	224.2	15.8	7.1
'a'	228.7	14.6	6.4
's'	177.7	94.6	53.2
'l'	221.6	9.3	4.2
'i'	216.7	12.7	5.9
'#'	0	0	0
'p'	202.6	14.9	7.3
'i'	143.8	99.5	69.2
'r'	179.1	63.4	35.4
'a'	195.9	7.1	3.6
'b'	195.6	13.3	6.8
'u'	308.1	141.2	45.8
'#'	0	0	0
'a'	184.2	5.9	**3.2**
'R'	184.9	7.1	3.8
'i'	185.2	8.5	4.6
'w'	180.3	12.1	6.6
'i'	179.5	11.8	6.5
'tl'	127.6	88.8	69.6
'A'	177.9	10.2	5.7
'r'	60.9	79.3	130.1

Table 2: Amplitude statistics of phones in the first 3 words; Unit = dB

Phone	Mean	SD	% dev
'i'	58.3	4.1	7.0
'n'	57.9	3.9	6.7
'd'	55.3	3.6	6.6
'a'	59.7	3.1	5.3
'tl'	52.8	3.2	6.0
'i'	56.9	3.3	5.8
'Tl'	56.9	2.8	5.0
'a'	62.4	4.3	6.9
'ng'	55.9	3.5	6.3
'g'	55.3	4.2	7.6
'a'	58.7	4.8	8.2
'L'	58.2	4.3	7.3
'ae'	60.6	3.8	6.2
'#'	0	0	0
'E'	60.6	3.8	6.3
'TR'	54.3	3.1	5.6
'u'	54.3	2.6	4.7
'kl'	50.8	4.2	8.3
'o'	57.7	2.9	5.2
'N'	54.6	3.5	6.4
'D'	52.7	2.6	4.9
'A'	56.8	3.3	5.8
'l'	50.8	3.9	7.7
'#'	22.4	3.2	14.3
's'	50.4	3.6	7.2
'u'	55.9	2.4	4.2
'd'	55.8	1.6	**2.9**
'E'	61.2	2.2	3.6
's'	53.4	3.7	7.1
'a'	56.5	4.2	7.4

The 10 recordings of the sentence from a single speaker are collected for initial analysis. The PAD statistics of each phone in the sentence are collected and their mean, standard deviation (SD) and the percentage deviation about the mean are computed across the 10 utterances. The pitch and amplitude values are obtained for every 10 ms for all the 10 utterances using PRAAT software. P and A are the maximum values of pitch and amplitude of a phone and D is the duration of the phone. The PAD statistics (for the phones in 3 or more words) of the 10 utterances are given in Tables 1, 2 and 3, respectively. The pitch values for unvoiced phones in Table 1 are due to minor errors in segmentation.

The format of the tables is as follows. Columns 2 and 3 give the mean and SD of the 10 instances of the parameter analyzed and column 4 gives the

percentage deviation of the parameter, about the mean. '#' represents the pause between two words.

To detect the variability in the word speaking rate, the durations of the words are measured and their statistics are also computed. Table 4 shows the duration statistics of the 11 words over the 10 utterances of the sentence. Statistics are not collected for the entire duration of the utterances, because the SD of silence 'between' words is very high. Since the silence (pause) given by the speaker between two words is highly variable, that duration cannot be considered in calculating the speaking rate. But, a conclusion can be made that *a variable amount of silence can be given between words at different times for the same input text during synthesis*.

Table 3: Duration statistics of phones in the first 3 words; Unit = ms

Phone	Mean	SD	% dev
'i'	53.0	13.1	24.6
'n'	57.9	4.2	7.2
'd'	38.5	7.2	18.6
'a'	36.3	6.6	18.2
'tl'	109.6	4.5	**4.1**
'i'	31.7	3.8	12.1
'Tl'	112.9	6.7	5.9
'a'	50.3	2.7	5.4
'ng'	55.9	3.2	5.8
'g'	27.4	3.1	11.6
'a'	21.2	4.4	21.0
'L'	38.1	6.7	17.8
'ae'	97.6	11.6	11.9
'#'	0	0	0
'E'	85.9	9.6	11.2
'TR'	123.6	9.6	7.8
'u'	30.3	4.8	16.0
'kl'	107.4	6.1	5.7
'o'	49.6	5.8	11.8
'N'	73.0	7.8	10.7
'D'	47.6	3.6	7.6
'A'	91.7	8.5	9.3
'l'	93.3	21.1	22.6
'#'	130.2	69.8	53.6
's'	101.0	4.9	4.9
'u'	35.7	4.4	12.4
'd'	60.8	4.8	8.1
'E'	77.2	8.8	11.3
's'	95.6	6.1	6.3
'a'	34.8	2.3	6.8

Table 4: Duration statistics of 11 words in the 10 utterances of the sentence; Units = ms

Word	Mean	SD
1	730.8	24.2
2	702.9	30.6
3	405.4	9.4
4	547.9	21.8
5	803.2	16.4
6	802.9	30.6
7	481.0	30.8
8	298.4	30.2
9	446.8	35.1
10	357.5	52.9
11	621.7	123.1

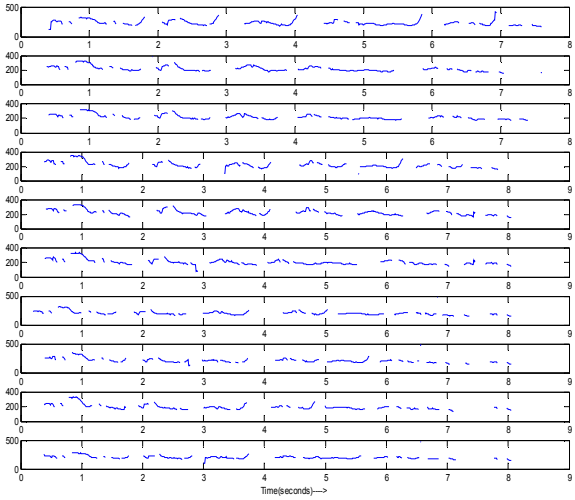

Figure 2: Pitch contours of the 10 utterances of the same sentence by a single speaker

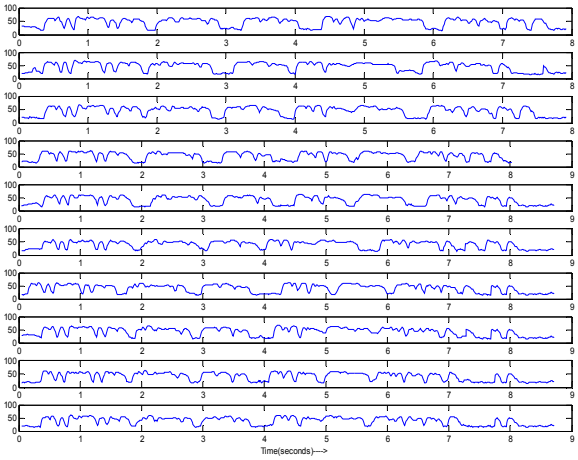

Figure 3: Intensity contours of the 10 utterances of the same sentence by a single speaker

The variations in the pitch and intensity contours of the 10 utterances are shown in Figures 2 and 3 respectively. The SD of the pitch of the last phones of words is higher than that of the remain-

ing phones. The minimum percentage of deviation in pitch about the mean is 3.2%. Thus, *pitch can be changed by a minimum of 3.2% for all the phones during synthesis*. Similarly, we observe that the amplitude of any phone can be changed up to 2.9% (see Table 2) *for all the phones during synthesis*.

In most of the cases, the SD of the duration of the last and first phones in a word are higher than that of the remaining phones. Also, the SD of a vowel in the middle of the word is significantly higher than that of the remaining phones. The SD of the phone /w/ when it occurs as the first phone of the word is more. The SD of nasals /m/ and /n/ is more when they occur as the last phone of the word. The phones in the last word of the utterance have more SD. The minimum deviation in duration about the mean is 4 % (see Table 3). So, *duration can be changed during synthesis by a minimum of 4 % for all the phones*.

Variation in phones: Our study revealed some unexpected, interesting facts. When a speaker speaks the same sentence/ phrase/ word several times, not only the characteristics of the phones in the sentence change, but we also observe *variations in the phones* uttered. Sometimes, some phones are missed or replaced by other phones. Two such words are shown in Fig 4 with the changed phones in bold. Note that these words are not spoken in isolation; they are part of a full sentence spoken at different times by a single speaker.

n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **h** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **k** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **k** a w u m	p A d u **k** A pl a D a n u m
n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m
n i ml a d i y A **g** a w u m	p A d u **k** A pl a D a n u m
n i ml a d i y A **g** a w u m	p A d u **g** A pl u D a n u m

Figure 4: Variation in pronunciation of phones when same words are uttered 10 times

According to Tamil phonology, the two words in Fig 4 should be spoken as /n i ml a d i y A g a w u m/ and /p A d u g A pl u D a n u m/. However, sometimes /g/ is spoken as /h/ and sometimes as /k/ (bold letters in Fig 4). This is natural for Tamil people; native Tamil people do not perceive such changes or even if they perceive occasionally, this variation in phones doesn't affect either the semantics of the word/ sentence or the understanding of the listener. We believe that other Indian languages may also have such phones which when spoken in specific phonetic or syllabic contexts, do not make a difference in the perception or understanding of the listeners. After observing this, we pose a question that *if a human being doesn't pronounce a sequence of phones (in a word/sentence) in exactly the similar way whenever she/he speaks and if that is perceived natural by the native listeners, why should an advanced speech synthesizer pronounce the same sequence of phones for a given text all the time?*

With an intention to take it forward, we explored more on this aspect by conducting perception experiments to find the most commonly interchangeable phones (referred to hereafter as 'confused phones') in particular phonetic contexts in Tamil. By doing so, we also see an additional benefit. If we find that two phones can be interchanged in a particular phonetic context, without any change in semantics and without a noticeable change in perception, we may substitute the unavailable phones in some contexts with their corresponding confused phones. This avoids to some extent, the limitations of not so phonetically rich TTS speech databases. And as mentioned earlier, this may also increase naturalness in the synthetic speech by making it closer to human speech.

4 Experiments to find the confused phones in Tamil

4.1. Perception Experiments

Listening experiments are conducted over the telephone to capture the most 'confused' phones in Tamil. One person calls another and pronounces a list of 152 phones/syllables (combination vowel and consonant(s)) in Tamil shown in Fig 5 and the person on the other side writes down the phones *she/he listens to*. Repetition of phones by the speaker is not allowed. Individual phones are cho-

sen to find the exact confusion between phones; if words are chosen, a listener who has a prior knowledge of the word writes the word correctly even though he might have not listened properly or the word is not pronounced properly; then the purpose of the study is not served.

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 only on the misrecognized phones. Not much improvement in recognition accuracy is observed. A consistency has been found in the misidentification over the speaker-listener pairs. Most nasals are wrongly identified as other nasals. Many 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 - /r/ and /R/. They are misrecognized for each other. 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/.

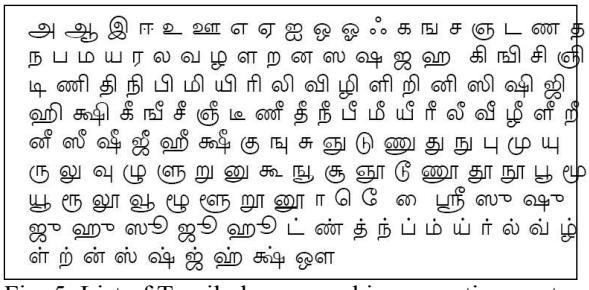

Fig. 5: List of Tamil phones used in perception expts.

Table 5: 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 – 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 6: Example words for the confused phones

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

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 occasional 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 substituted by each other. The entire set of phones and the consistently and frequently misrecognized phones are listed in Figure 5 and Table 5 respectively. Table 6 gives one example Tamil word each, for the uncommon phones in Table 5.

4.2. Phone Classification Experiments

We now proceed further to find the confusability among Tamil phones acoustically. The Tamil database used for this experiment consists of 1027 sentences from a single male speaker, sampled at 16 kHz. The sentences were segmented and labelled manually using PRAAT, by trained segmenters.

The traditional filter-bank approach (Molau et. al, 2001) is followed for extracting Mel Frequency Cepstral coefficients (MFCCs) from a speech signal. 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 and MFCCs are extracted. 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 each phone are created. Mean and covariance are obtained for all *.feat* files.

Classification is performed at two levels: frame and phone. In the former case, a single 20 ms frame (a 12-dimensional 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, the mean of all the MFCC vectors belonging to a 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. So phone level classification is also performed.

Phones are classified separately using full and diagonal covariance matrices. The classification accuracy obtained with the former one is found to be better than that obtained with the latter. 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. The results of phone level classification are presented in Tables 7 and 8.

Table 7: Phone level classification results with full covariance matrix; Number of sentences used for testing: 100; Variable: Average no. of feature vectors per class

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

Table 8: Phone level classification results with full covariance matrix; Number of sentences used for Training: 700; Variable: No.of sentences used for Testing

S.No	Test data size	BCCA	Accuracy
1	50	73.5%	52%
2	200	72.6%	51.2%
3	400	71.7%	50.5%

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) and 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 as a nasal and so on, the classification is considered to be accurate. The overall accuracy in the fifth column of Table 7 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 is observed with the increase of size of the test data.

A confusion matrix of the Tamil data for the significant mismatches is shown in Table 9. The classification accuracy of the phone /a/ is much higher than that of the other phones. 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/ pair 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/.

Table 9: Confusion matrix of most confused 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

4.3. Exploit the confused phones to increase the naturalness of synthetic speech

Recall Fig 4. According to Tamil people, /g/ replaced by /h/ is acceptable, but /g/ replaced by /k/ is not so acceptable. According to the G2P rules also, the grapheme /k/ which occurs between two vowels is changed to /g/. Here the appearance of /k/ instead of /g/ is due to the mispronunciation of the speaker and therefore not taken into account. The point here is this: *if the presence of any one of two phones in a particular context leads to the same perception of the listener, this knowledge can be used to make the synthesizer speak differently at different times for the same input text*. The knowledge of confused phones can also be used to replace the phones unavailable (in the TTS speech database) in some contexts with the corresponding confused phones.

Care should be taken while replacing the phones. The idea is not to simply replace some phones by their corresponding confused phones randomly. For the present application, intentional

replacement of phones with their corresponding confused phones is performed, with a view to make the output of synthesizer more natural. A phone is occasionally replaced even when there is an availability of phones in specific phonetic contexts. This process should not degrade the output of synthesizer, but it should make the synthetic speech closer to that of a human being's speech. The user must not be led to think that there is something wrong with the synthesizer. The idea is to mitigate the user experiencing the synthesizer as monotonous, when she/he listens to, for example, the same story several times.

5. Summary

In summary, we propose that the following steps can be taken to enhance the quality of TTS synthesizer in the sense of bringing the synthetic speech closer to human speech, by making it produce speech with different characteristics at different times for the same input text.

- The pitch can be changed by a minimum of 3.2% for any phone during synthesis.
- The amplitude can be changed by a minimum of 2.9% for any phone.
- Duration can be changed by a minimum of 4% for any phone during the synthesis.
- A variable amount of silence may be given between words at different times for a particular input text.
- Phones in specific contexts may occasionally be replaced by their corresponding confused phones.

6. Conclusion

A strategy through which synthetic speech can be generated with different characteristics at different times for the same input text is suggested. Identifying the need for the study of inherent variation in human speech to make the synthetic speech closer to human speech, the variation in parameters of speech viz., pitch, amplitude and duration is studied for different instances of a sentence by a single speaker. The percentage modification which can be made to these parameters during synthesis is determined. The data collected can be analyzed in future for acquiring more knowledge of variation in speech parameters when the same speech is spoken at different times by a single speaker and across different speakers also. Confused phones that are used interchangeably in specific phonetic contexts are found by conducting perception experiments. These phones are found to be perceived indistinguishably in specific contexts and they also cause no change in the semantics of the sentence. The knowledge of the confused phones in Tamil can be used to make the synthesizer occasionally speak differently for the same input text and also to resolve to some extent, the limitations of not so phonetically rich TTS databases. The experiments conducted are for the Tamil language; we believe that this can possibly be extended to other languages as well.

Acknowledgments

The authors would like to express their gratitude to Mrs. Shanti Devaraj for her manual segmentation of the speech corpus and Mrs. Shanti Srinivas for her help in collecting speech samples from different Tamil speakers.

References

A G Ramakrishnan, Laxmi Narayana M, *Grapheme to Phoneme Conversion for Tamil Speech Synthesis,* Proc. Workshop in Image and Signal Proc. (WISP-2007), *pp. 96-99*, IIT Guwahati, Dec 28-29 2007.

Cairns, Douglas A, Hansen, John H. L, *Nonlinear analysis and classification of speech under stressed conditions*, Journal of the Acoust Society of America, Vol 96, Issue 6, pp.3392-3400, Dec 1994.

C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," IEEE Trans. Speech and Audio Processing, 2(4), pp. 469-481, 1994.

David Hirtle, *Speech Variability: The Biggest Hurdle for Recognition*. Internal report, Faculty of Computer Science, University of New Brunswick, 2004.

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

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.

PRAAT: A tool for phonetic analysis and sound manipulations by Boersma and Weenink, 1992-2001, www.praat.org.