NOVEL SCHEME TO HANDLE THE MISSING PHONETIC CONTEXTS IN SPEECH SYNTHESIS, BASED ON HUMAN PERCEPTION AND SPEECH

Laxmi Narayana M and A G Ramakrishnan,

Department of Electrical Engineering, Indian Institute of Science, Bangalore.

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 are performed to find out such phones. The classification experiments are performed on Tamil corpus of 1027 sentences. The natural variability in human speech is studied by analyzing utterances of same speech at different times. We observe that, not only the characteristics of phones change, when same sentence is spoken at different times, but also the phones get replaced by other phones, for the same speaker.

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 [3]. 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 also important. Mono-phones are considered for concatenation only in the worst case.

2. GOAL OF THE WORK

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 database. 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. 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 (for example), then we get around 216 syllables each of which can occur between any two syllables. So for a syllable, there are 216^3 possible contexts for its occurrence. All of them may not be valid, but the issue is, practically, for any corpus, 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 notice any difference in perception, can be used 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 or we might have to repeat some words or syllables, even though the phone conversation takes place in a less noisy environment. Further, when a speaker utters the same sentence at different times, sometimes, some phones may be missed or replaced by other phones; but still the listener can make out the sentence. The knowledge of 'what phones are replaced by what phones' can also be used to handle the missing phonetic contexts. The present paper reports the perception experiments and the phone classification experiments conducted to find out such phones**.** The results are used in our Tamil TTS System. The Tamil database under consideration contains 1027 sentences from a single male speaker, sampled at 16 kHz, which are segmented and labeled manually using Pratt software. Though the database is 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 and the person on the other side writes the phones *whatever she/he listens to* for the first time. Repetition of phones by the speaker is not allowed. 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 - /r/ and /R/. They are misrecognized for each another. The three types of /l/ - /l/, /L/ and /zh/ 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. 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.

அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஃ க ங ச ஞ ட ண த ந ப ம ய ர ல
வ ழ ள ற ன ஸ ஷ ஜ ஹ கி ஙி சி ஞி டி ணி தி நி பி மி யி ரி லி
வி ழி ளி றி னி ஸி ஷி ஜி ஹி கூ கீ ஙீ சீ ஞீ டீ ணீ தீ நீ பீ மீ யீ ரீ
லீ வீ ழீ ளீ றீ னீ ஸீ ஷீ ஜீ ஹீ கூ கு ங சு ஞு டு ணு து நு பு மு யு
ரு லு வு ழு ளு று னு கூ ங்கூ சூ ஞூ டூ ணூ தூ நூ பூ மூ யூ ரூ லூ வூ
ழூ ளூ றூ னூ ா ெ ே ை ஶ்ரீ ஸு ஷு ஜு ஹு ஸூ ஜூ ஹூ ட் ண் த் ந் ப் ம் ய் ர் ல் வ் ழ் ள் ற் ன் ஸ் ஷ் ஜ் ஹ் க்ஷ ஔள

Figure 1: List of Tamil Phones used

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

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.

Similar experiment was conducted on phones of Telugu 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 most frequently and consistently misidentified phones in Telugu are shown in Table 3. The misidentification rate in Telugu is less compared to Tamil. On an average, 10 phones out of 52 were misrecognized. That comes to a 20% misidentification whereas it is around 30% in Tamil. This is because the number of phones in Telugu is more compared to that of Tamil; Telugu people are habituated to pronounce and 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 common misidentifications are shown in bold in Tables 1 and 3.

5. NATURAL VARIATIONS IN SPEECH

Data Collection: Ten sentences from the Tamil corpus are selected and 8 native Tamil people are asked to speak the 10 sentences, at 10 different times over a period of 3 days. The time gap between two recordings of one speaker is at least 3 hours. The sampling rate is 16 kHz, which are manually segmented and labeled using Pratt software.

Variation in phones: When a speaker speaks the same sentence/phrase/word several times, we also observe variations in the phones uttered. Sometimes, some phones are missed or some phones are replaced by other phones. The words of one sentence by one speaker spoken at 10 different times are listed in Figure 2, with the changed phones shown in bold. According to the rules of the Tamil language, the two words shown in Figure 2 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`. But as it is seen from the figure, sometimes /g/ is spoken as /h/ and sometimes as /k/.

So, this observation says that, replacing the unavailable phones in some contexts with the corresponding confused phones, not only avoids discontinuities, but also induces naturalness in the synthetic speech.

```
 nimladiyAgawum # pAdugApluDanum
nimladiyAhawum # pAdugApluDanum
nimladiyAkawum # pAdugApluDanum
nimladiyAkawum # pAdukApluDanum
nimladiyAgawum # pAdugApluDanum
nimladiyAgawum # pAdugApluDanum
nimladiyAgawum # pAdugApluDanum
nimladiyAgawum # pAdugApluDanum
nimladiyAgawum # pAdukApluDanum
nimladiyAgawum # pAdugApluDanum
```

Figure 2: Variation in pronunciation of phones when same words are spoken 10 times.

6. 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 are classified.

6.1. Feature Extraction – Training & Testing

The traditional filter-bank approach [4] is followed for extracting Mel Frequency Cepstral coefficients (MFCCs) from the speech signal. The process is very briefly presented here. 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.

Two types of classifications are performed: frame level and phone level. In the former case, a single 20 ms frame (a 12 dimensional acoustic vector) is classified to one of the 48 (Tamil) phone classes using the ML classifier [1]. In the later case, mean (vector) of all the MFCC vectors belonging to one phone 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 not to design a robust classifier, but to find the confusability of phones so that the phones can be used interchangeably in some contexts. So phone level classification is also done.

7. RESULTS AND DISCUSSION

Phones are classified using full covariance matrix and diagonal covariance matrix. The classification accuracy obtained in the former case is better compared to that of the later. The experiments are carried out for different sizes of training and test data and the phones that are misclassified are noted down. The results of the phone level classification are presented in Tables 4 and 5.

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 4: 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 5: 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

7.1. 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 as a nasal and so on, the classification is considered to be accurate. The overall accuracy in the fifth column of Table 4 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.

7.2. Confusion Matrix

A confusion matrix of the Tamil data for the significant mismatches is shown in Table 6. The classification accuracy of the phone /a/ is relatively 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/.

		True Class							
Assigned Class		a	A	i	I	ae	l	ll	yl
	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 6: Confusion matrix of most confused Tamil phones

7.3. 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 7. 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 7. Some special replacements which are more language specific are shown in Table 8. The phonetic transcription of the words is shown. The IPA codes of the phonemes can be found in [2].

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 7: Phone replacements done during synthesis.

Original word	Word after phoneme 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 8: Tamil words, before and after phone replacement.

8. CONCLUSION AND FUTURE SCOPE

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 ML classifier. The confused phones in Telugu are also found by perception tests. The common confused phones over the two Indian languages are identified. The natural replacement of phones by other phones in human speech is also observed. This gives a hope that the proposed phone replacement strategy also makes the synthetic speech close to natural speech. The collected data can be analyzed for the variability in characteristics of phones and the knowledge can be incorporated in TTS to induce naturalness in the synthetic speech, in future. The knowledge of the confused phones is incorporated in Tamil Text to speech synthesis and experiments show that the proposed phone replacement strategy is fairly successful.

9. REFERENCES

[1] Duda, Hart, Stork, *Pattern Classification*, Second Edition, John Wiley & Sons, 2001.

[2] http://en.wikipedia.org/wiki/Tamil_script.

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

[4] Molau. S, M. Pitz, R. Schlüter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, Jun. 2001, pp. 73–76.