Synthesis of missing units in a Telugu text-to-speech system

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science by Research
in
Computer Science and Engineering

by

Vijayaditya Peddinti
200807011
`vijayaditya.p@research.iiit.ac.in`

Language Technologies Research Center
International Institute of Information Technology
Hyderabad - 500 032, INDIA

July, 2011

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Synthesis of missing units in a Telugu text-to-speech system" by Vijayaditya Peddinti, has been carried out under my supervision and is not submitted elsewhere for a degree.

July 7, 2011	Adviser: Dr. Kishore Prahallad

Synthesis of missing units in a Telugu text-to-speech system

by

Vijayaditya Peddinti

Submitted to the International Institute of Information Technology - Hyderabad
on July 23, 2011, in partial fulfillment of the
requirements for the degree of
Master of Science by Research in Computer Science and Engineering

Abstract

Synthesis of complex consonant clusters is a challenge in TTS systems, as it is difficult to ensure proper coarticulation. This problem surfaces in the synthesis of missing units in unit concatenation systems. Back-off techniques are used to synthesize these missing units. The current back-off techniques, which use smaller sub-word units to synthesize the missing unit, are found unsuitable for frequent usage causing severe degradation in synthesis quality. However missing units are a frequent occurrence when synthesizing unrestricted text. Hence to build open domain TTS systems, efficient back-off techniques are necessary. This problem is further aggravated when working with under-resourced languages as they lack the manually annotated speech corpora which enable the development of these high quality synthesis techniques. In this thesis we propose a two pronged strategy to improve the quality of back-off techniques. The two pronged strategy consists of refining the boundaries of units in the acoustic inventory to reduce phone insertion errors during back-off and use of a novel back-off technique to ensure coarticulation consistency.

A boundary refinement algorithm which removes necessity for manually labelled data by relying on a knowledge base of cost functions derived from landmark specific acoustic cues was proposed, to increase the quality of unit boundaries. However increase in accuracy of unit boundaries was insufficient when context-dependent units are not available in the database. Hence a back-off technique which synthesizes complex consonant clusters even when the required context-dependent units are not available, was proposed. This technique is motivated from a phenomenon observed during second language acquisition and emulates native speaker intuition in synthesis of these clusters. The technique was proven to be better than the conventional back-off techniques. Further it was proven to be robust to segmentation errors, reducing the reliance on high quality segmentation techniques which require manually labelled data.

Thesis Supervisor: Dr. Kishore S. Prahallad
Title: Assistant Professor

Acknowledgements

I would like to thank my advisor Dr. Kishore Prahallad who introduced me to the challenging field of speech synthesis, where even the minute glitch matters, where novelty requires understanding and exploring the wide variety of domains and where a contribution can have tremendous impact. The long discussions, patient hearings and insightful critique he afforded me were invaluable.

I would like to thank Prof. Yegnanarayana whose lectures on the challenges in speech synthesis humbled me and made me approach the topic with the appropriate mindset.

I would also like to thank my lab mates Gautam, Lakshmikanth, Raju, Venkatesh, Raghu, Sachin, Anand and Naresh without whose help this work would not have been possible, as they volunteered for innumerable perceptual tests without any hesitation.

Finally I would like to dedicate this thesis to all the researchers who study the rare phenomena that hold clues to the enigmatic working of human brain.

"To the abnormal and interesting"

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Text-to-Speech (TTS) systems play an important role in Human-Computer interaction. This thesis discusses the challenges in the design of TTS systems. This chapter describes the architecture of a text-to-speech synthesis system, the techniques used in these systems and the challenges involved.

1.1 Text-to-Speech

Text-to-speech synthesis is the process of converting any form of text into speech. A text to speech synthesis system can be broadly divided into front end and back end as shown in Figure 1-1. The front end deals with the conversion of natural language text to a structured linguistic representation. This front end predicts a sequence of segments called target segments from the raw text. These target segments have a variety of features determined from the text. The second part of the system which converts these target segments into a speech waveform, is referred to as the back end. The techniques discussed in this thesis pertain to the back-end or the waveform generation module of the TTS.

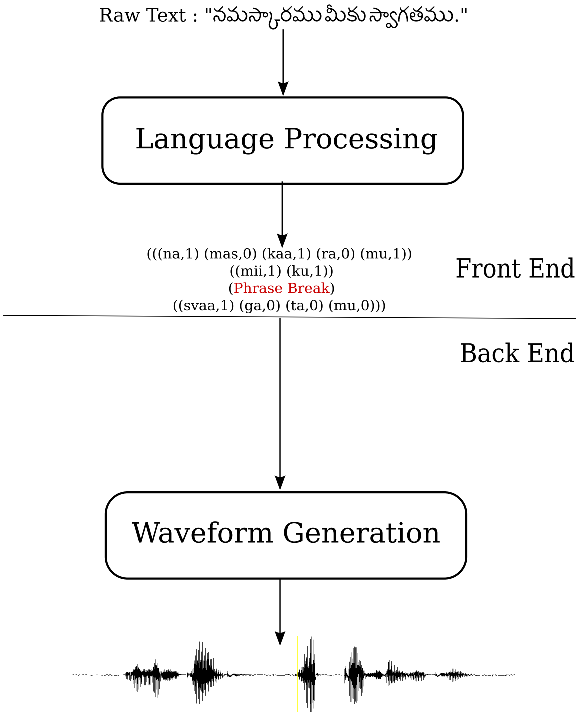
Raw Text : "నమస్కారము మీకు స్వాగతము."
Language Processing
(((na,1) (mas,0) (kaa,1) (ra,0) (mu,1))
((mii,1) (ku,1))
(Phrase Break)
((svaa,1) (ga,0) (ta,0) (mu,0)))
Front End
Back End
Waveform Generation

Figure 1-1: Parts of Text-to-Speech system

1.1.1 Waveform generation in TTS

Early efforts in TTS research focussed on parametric synthesis techniques for waveform generation, where the parameters are determined using rules designed by experts. [3], [4], [5], [6] were some prominent rule-based TTS systems. These rule-based parametric synthesis techniques, though smooth sounding, lack naturalness. Hence data-driven systems which provide comparatively more natural sounding speech have received research focus in the last decade. Further data-driven synthesis techniques reduce the effort of collecting expert knowledge and instead shift the onus to collecting data and searching for appropriate speech segments during synthesis.

Data-driven synthesis, also referred to as corpus-based synthesis, makes it possible to dramatically improve the naturalness of speech over rule based methods. Further, these techniques have gained popularity as they can be built with very limited knowledge about the language. Most of the knowledge is captured from the data [7], at least for basic systems. There are two types of prominently used data-driven synthesis techniques; the statistical parametric synthesis technique [8] and the unit selection synthesis technique [9]. Unit selection synthesis technique has been the popular technique of these two. Even with great number of disadvantages like large database size and sensitivity to the database quality, unit selection synthesis continues to be popular due to the quality of the natural sounding speech that can be produced in the best case performance [8]. This thesis deals with the unit selection synthesis technique.

Unit selection synthesis

Unit selection synthesis technique, proposed in [10] and [9], involves concatenating units of natural speech. Each type of unit has multiple instances in the inventory. The large speech databases with multiple instances of each unit allow modelling of greater variability and different speaking styles in natural speech. A number of examples of each unit in different contexts are included, and the synthesis process is formulated as a search problem. A search is performed to find the best sequence of units. The goal of unit selection speech synthesis is to select a sequence of units which requires less signal processing, or ideally no signal processing at all [11].

In this technique each instance of a unit in the database is labelled with a vector of features. These features include linguistic features, context features and acoustic features. The target segments on the other hand have a subset of these features that can be determined from the text. For synthesis of an utterance, units have to be selected from the database and then concatenated. This selection is done based on two cost metrics [10]:

Unit distortion $D_u(u_i, t_i)$ is defined as the distance between the selected unit, u_i, and a target segment, t_i, provided by the front end. It is the difference between the selected unit feature vector $\{uf_1, uf_2, uf_3, uf_4, ..., uf_n\}$ and the target segment vector $\{tf_1, tf_2, tf_3, tf_4, ..., tf_n\}$ multiplied by a weights vector W_u $\{w_1, w_2, w_3, w_4, ..., w_n\}$. This difference represents the cost of selecting the particular instance of the unit from the database to represent the target segment. Hence it is also called *Target Cost.*

Continuity distortion $D_c(u_i, u_{i-1})$ is the distance between the selected unit (u_i) and its immediately previous selected unit (u_{i-1}), defined as the difference between selected unit's feature vector and its previous unit's feature vector multiplied by a weight vector W_c. This difference represents the cost of joining two units. Hence it is also called *Join Cost.*

The **best unit sequence** is defined as the path of units from the database which minimizes

$$\sum_{i=1}^{n}(D_c(u_i, u_{i-1}) * W_c + D_u(u_i, t_i) * W_u) \tag{1.1}$$

where n is the number of segments in the target utterance, and W_c and individual W_u are weights of the target and join costs respectively.

A Viterbi search is used to find the path with the minimum cost as defined by the expression above [9]. Given a set of targets representing the utterance to be synthesized, for each target segment all units in the database with low phonetic distance to the target are identified. The unit distortion for these units is calculated with respect to the target, this list is pruned to m best ones. Next for all these candidates the continuity distortion with all candidates of the previous target is calculated. The procedure is continued through all targets finding all possible paths at each stage and the best path is selected at the end [10].

In unit selection synthesis the choice of unit type could vary. The prominently used unit types are diphones [11], phones [9], syllables [12], triphones [13] and non-uniform units [14]. The choice of unit is usually done based on various factors, language being one of the important factors. According to the concept of mental syllabary, high-frequency syllables are stored as complete gestural programs which are executed during speech production, whereas the vast majority of low-frequency and very rare syllables are assembled online using the segmental and metrical information provided by the phonological encoder [15]. Hence syllables are desirable as units in unit-concatenative synthesis, over other sub-word units. Further syllables in Indian languages have a regular structure compared to those in other languages like English [16], hence it is feasible to use syllables as units in Indian language TTS systems. In this thesis the focus is on synthesis with syllable units. However the discussion and all the techniques proposed are equally applicable to other unit types.

1.2 Issues in unit selection synthesis

The unit selection synthesis technique presents its own set of issues, which have evolved in research topics. The main issues in unit selection synthesis research are [17]:

- **Automatization of the inventory selection**

  Design of acoustic inventories, to ensure proper coverage of all units, is crucial for data driven synthesis. Acoustic inventories for TTS are usually prepared using read speech. The text used for recording is selected from large text corpora. There is a risk, however, in relying too much on frequencies of units observed in any one particular text corpus, because of the instability of unit frequency distributions across text corpora. Text corpora, no matter how large, should be viewed as samples of a larger, abstractly defined domain (e.g., the written

English language). Tying system construction too closely to a particular text corpus runs the risk of neglecting units that may prove unexpectedly frequent in new corpora. Hence coverage of all units across multiple text corpora is the goal [18]. Manual selection from these huge text corpora is a tedious task. Hence design of text selection algorithms which optimally select text, to maximize the coverage of various types of units, is crucial. Hence automatization of this optimal text selection process is a prominent research topic.

- **Synthesis of missing units**

  There are often cases where the unit frequency distribution in text corpora is extremely uneven. The number of rare units is sufficiently large and their combined probability mass makes it quite likely that even a small text sample presented for text-to-speech synthesis will contain at least one rare unit. Hence, any TTS system has to be prepared for rare units [18]. These rare units usually are missing in the acoustic inventory despite the use of optimal text selection algorithms to ensure coverage [15]. Irrespective of the quality of inventory, missing units are a consistent problem in unit selection technique [11]. Identification of techniques that tackle these missing units is another research issue.

- **Selection of parametric representations**

  Data-driven synthesis depends on large speech corpora. However storing the corpora as waveforms is costly in terms of space. Well chosen parametric representations of speech allow data-size reduction. Further, segment concatenation and prosody matching functions depend on the type of parametric representation [7]. Formants [19], linear predictive coding (LPC) coefficients [20], sinusoidal models [21], mel-frequency cepstral coefficients (mfcc) [22], weighted spectrum [23], etc., are some of the prominently used parametric representations. Each of these representations is suitable for specific purpose like preserving naturality in synthesized speech, calculation of join cost, smoothing at unit joins, prosodic modification, etc. Hence proper selection of the parametric representations to balance these competing requirements is another major issue.

- **Selection of target and join cost functions for unit selection**

  The target and join cost functions form the core of the unit selection synthesis technique and hence are critical for proper synthesis. Specification of the targets and design of the target cost function are deemed important, as the target cost is used to reduce the search space of candidate units. Hence incorrectly assigned target costs can lead to the elimination of the suitable candidates. The join cost function tries to minimize the perceptible discontinuities in the synthesized speech [24]. The calculation of this *perceived* discontinuity using objective distance measures is a difficult proposition. Hence search for distance measures which have a high correlation with the subjective human perception is a prominent research topic. A variety of studies like [25],[26] have been conducted to study the design of these cost functions.

- **Selection of concatenation algorithms**

  Though the unit selection algorithm is optimal, the final units to be concatenated are extracted from different words. Hence they present mismatches in loudness, pitch and timbre(characteristics of sound which allow the ear to distinguish sounds that have the same pitch and loudness). These lead to audible discontinuities which need to be tackled using smoothing techniques at every join of two units. The smoothing methods must produce formant trajectories and f0 trajectories that are possible to be produced by the human articulators. However excessive smoothing can lead to buzzy speech. Hence the selection of proper smoothing techniques is important for ensuring final synthesis quality. The choice of the parametric space to represent speech and the prediction of these parameters' trajectories at the joins is another research issue.

- **Choice of unit type**

  Choice of unit type is another issue in unit selection synthesis. The prominently used unit types are diphones [11], phones [9], syllables [12], triphones [13] and non-uniform units [14]. It is also known to vary according to the language and studies like [27] have been conducted to choose the best unit type for a particular language. Hence selecting the right unit type according to the language and sometimes even according to the application domain, is another research issue.

- **Automatic segmentation of speech for inventory preparation**
  Automatic segmentation is a precondition for acoustic inventory creation. The large databases used in the data-driven synthesis technique make manual segmentation a prohibitive task. Further, unlike human errors, the errors produced by automatic segmentors are known to be consistent which is preferable for use in unit concatenation and acoustic modelling. Hence data-driven synthesis research also abounds with works on automatic segmentation. However even the state of the art segmentation technique ([28]) has an accuracy of just 96% within an error tolerance limit of 20ms. In unit selection, the quality of labelling determines the quality of units, which might be affected by a range of problems including misaligned phone boundaries, mismatches between the phones that are labelled and that are pronounced, and the presence of background noise [29]. Automatic labelling methods are still error-prone, so that they are often followed by a stage of manual correction. Hence the design of high quality automatic segmentation methods is a research issue.

This thesis focuses on the issues which result from the large number of infrequent units in the text of a language. A major issue is the design and use of acoustic inventories, which results in incomplete coverage of the acoustic inventory, known as the *missing unit* problem. This thesis discusses techniques to tackle the problem of synthesizing missing units.

1.3 Issues with a large number of infrequent units

The use of larger sub-word units like syllables is preferred in concatenative synthesis technique as they capture the important transition and coarticulative variations among the phone segments and reduce the number of concatenation points. Further, even in data-driven synthesis using phone units, context information is used to categorize the data. Context includes information about the neighbouring phones. Context specific variations in sub-word units, like phones and diphones, are too prominent to ignore and the use of context-dependent units is known to increase both the intelligibility and the naturality of the synthesized speech. Even in target cost calculation, context features are given higher weightage as they capture acoustic variations that are difficult to capture using cost functions in the signal domain. The use of these larger sub-word units or context-dependent units drastically increases the number of units to be covered in the acoustic inventory.

Phonotactics of a language is the permissible combinations of phonemes that can co-occur in that language. Thus phonotactics helps reduce the number of possible larger sub-word units like syllables or context-dependent units. However it is difficult to get a complete coverage of even this reduced set in the unit inventory for the language. The problem with these larger sub-word units or context-dependent units is that the resulting symbols have distributions which are of the class LNRE (Large Number of Rare Events) i.e., the syllable frequency distributions are extremely uneven, while some syllables have a high frequency of occurrence, the vast majority of the syllables are extremely rare. This affects several modules of the TTS systems like morphological analysis, segmental duration modelling and acoustic inventory design [15]. However, in this thesis the effect of these distributions on acoustic inventories, which are critical for determining the quality of data-driven synthesis techniques, is studied and techniques are proposed to mitigate the effect of these distributions.

Mobius [15] has shown that syllable type frequency distributions in languages with complex syllable structure, such as English or German, display typical LNRE characteristics. A few hundred syllable types account for the majority of realized syllable tokens in speech production, whereas the vast majority of syllable types are very rarely used. However Indian languages, especially Telugu, which is the language of interest in this thesis, have a comparatively simple syllable structure [16]. Hence it is expected that the simple syllabic structure would lead to more manageable short-tailed syllable distributions. However, when considering unrestricted text, even languages with simple syllable structure, like Telugu, present a large number of rare syllables. Figure 1-2 represents a typical syllable frequency distribution in Telugu language, using the newspaper text. It can be seen that a vast majority of the syllables have no more than one occurrence (Region A of Figure 1-2).

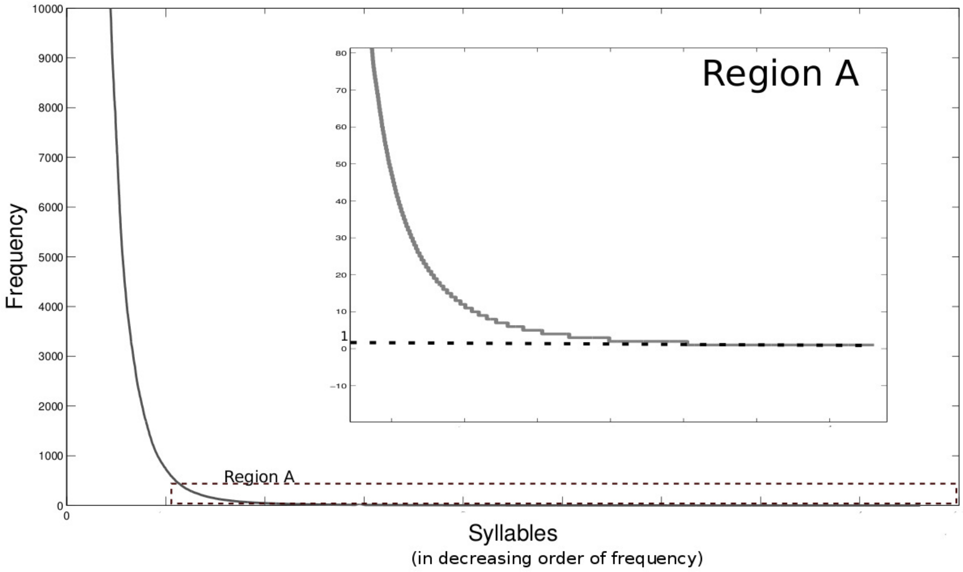

Figure 1-2: Syllable frequency distribution in Telugu newspaper text

Ensuring complete coverage of syllables, when their frequency distributions are of the LNRE class, is a difficult task. The de-facto technique in the data-driven systems to ensure coverage has been provision of more and more data. Figure 1-3 shows the database size *vs* syllable coverage, in Japanese language which also has a simple syllabic structure. The incremental increase in syllable coverage with corresponding increase in size of database is negligible after 80% of cover ratio ($\frac{no.ofcoveredsyllables}{totalnumberofsyllables} *$ 100) [1]. Similar behaviour is expected in Telugu language.

To further compound the problems above, influx of new words into a language, due to borrowings from other languages, is a fairly frequent phenomenon [30]. These borrowed words do not even necessarily conform to the phonotactics of the synthesis language. Thus, these borrowed words result in missing syllables due to the presence of new consonant clusters disallowed by the phonotactics of the synthesis language. Thus in the case of multi-lingual interactions or interactions with a large amount of foreign phrases the acoustic inventory design becomes a critical problem as the

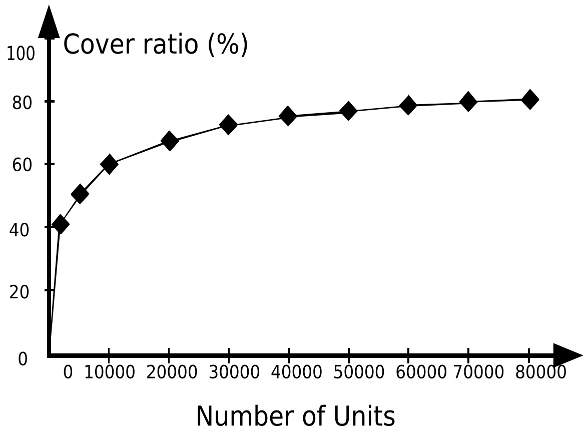

Figure 1-3: Cover ratio in Japanese text [1]

number of syllables to be covered in the database increases drastically. The solution of just providing more data to increase the coverage is not viable, as the number of languages to cover is far too many. Hence alternate techniques are to be explored for the effective use of data-driven synthesis techniques in the multi-lingual context.

Incomplete coverage in the acoustic inventory due to LNRE distributions of the syllable frequencies leads to missing units. Missing units are a common problem even in TTS systems with other unit-types like diphones [11]. In the case of diphones, though it can be ensured that all possible diphones occur in the database using carefully constructed sentences, many such diphones would not occur more than once in the database. These missing units are dealt with using back-off strategies described in Section 1.3.1.

1.3.1 Prior work on synthesis of missing units

Coverage issues affect both the data-driven synthesis techniques viz., statistical parametric synthesis and unit concatenation synthesis. However in this thesis the effects of coverage issues on the unit concatenation synthesis technique are discussed and methods are proposed to deal with these issues. In the unit concatenation synthesis technique missing syllables in the acoustic inventory lead to missing units during the synthesis stage. Hence these units are synthesized using techniques called back-off techniques. Back-off techniques are usually sub-optimal synthesis techniques and do not necessarily lead to the highest possible synthesis quality. The back-off methods either substitute the missing syllable with other syllables in the database [31] or synthesize it using smaller sub-word units (SSUs) like diphones. However even missing diphones are a common problem [11]. In these cases, diphone substitution or half-phone back-off is used to synthesize the missing diphones [32].

The design of both these back-off methods is not trivial, as each of them presents its own set of problems. Phonetic segments within syllables are found to have relatively high coarticulation, compared to those across syllable. The SSU back-off method which forms the missing syllables from smaller sub-word units has to ensure that the units selected have the right coarticulative influences, some of which span across several phones, e.g., influence of lip protrusion of the vowel is seen across many preceding consonants [33]. Further in the absence of required diphone, half-phone back-off is chosen. Half-phone back-off method splices half-phone units, at unstable regions like phone boundaries which is not desirable [25]. In addition to this, the audio databases used in TTS systems are segmented automatically into constituent phone units, from which the SSU (diphone and half-phone) boundaries are derived. The joining of units at these automatically segmented phone boundaries is a non-trivial problem and may not lead to optimal output. Usually techniques like optimal coupling [34] are used to avoid joins at phone boundaries which are unstable regions. However the back-off techniques especially at the half-phone level cannot apply optimal coupling. Hence joins at phone boundaries are performed, which result in insertion of extra phones. Figure 1-4 shows the waveform of the word [keːks][1]. The word was synthesized with a join at the phone boundary, due to the unavailability of the syllable “keːks” and also the diphone “k-s” in the acoustic inventory. It can be observed that due to improper placement of phone boundaries, the trailing phone of the unit “k” and the leading phone of the unit “s” from the original waveforms are present in the synthesized waveform. This leads to the presence of the extra perceptible phone, “u” in the synthesized waveform. Hence back-off techniques necessitate the use of high accuracy segmentation methods. The high accuracy segmentors currently available,

[1] written in IPA transcription [35]

require manually labelled data for training. However in under-resourced languages it is difficult to assume the availability of such manually labelled databases. Hence a segmentor which is capable of high accuracy segmentation and at the same time does not impose the necessity for training data is essential. Further, even though the insertion of the extraneous phones can be avoided using highly accurate segmentation algorithms, the coarticulative influences on the units result in unacceptable synthesis when the units are picked from different contexts. This is a severe problem in the back-off synthesis techniques as the proper contexts are not available during the selection of the SSU units.

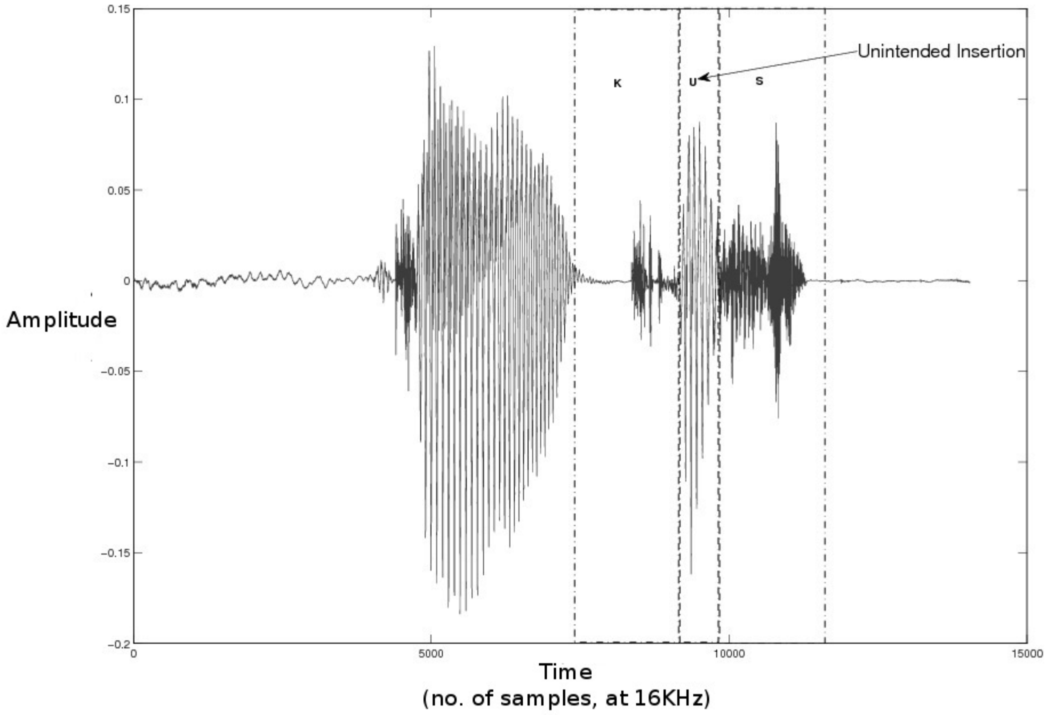

Figure 1-4: Insertion of extra phone "u" in the word [keːks] due to join at phone boundary "k-s"

Substitution methods avoid the aforementioned problems, by using other syllable or diphone units in the inventory to replace the missing units. However it takes considerable effort to design the substitution rules, requiring detailed perceptual studies when designing for new languages [31]. Even after these detailed studies, minimal pairs i.e., words which only differ in one phonological element and have distinct meanings, pose a problem when using substitution as a back-off. Hence the application of substitution rules should be done taking care that the meaning of the word does not change. Further diphone substitutions necessitate a change in the neighbouring units, thus creating a complex search problem [11].

Hence there is necessity for a back-off strategy which selects units with the right coarticulation, is robust to segmentation errors and is easy to design in terms of human effort.

1.4 Issues addressed in this thesis

It can be deduced from the previous sections that high quality speech segmentation algorithms are critical in back-off synthesis as optimal coupling cannot be applied. However the current high-accuracy segmentation techniques require manually labelled data for training. The availability of such manually labelled databases cannot be assumed in under-resourced languages. Hence a novel boundary segmentation algorithm is proposed for purpose of providing high accuracy unit boundaries without the use of manually labelled training data.

Even the state of the art segmentors have an average error of 6ms at unit boundaries. Hence the back-off synthesis strategies must be robust to these segmentation errors. In addition, coarticulative variations need to be taken care of while joining units taken from different contexts, especially during back-off synthesis. A rule-based back-off method is proposed to deal with the missing syllable and diphone units in Telugu language. It is motivated from a perceptual and speech production phenomenon, known as *vowel epenthesis.* The proposed back-off method emulates native speaker intuition in synthesis of the missing units. Hence it is found to be perceptually acceptable to the native speakers of the language. The rule-base is borrowed from the L2 (second language) acquisition research in Telugu. Further, the proposed back-off strategy is robust to the segmentation errors in automatically segmented databases, helping us ensure the quality of synthesis output.

1.5 Organization of the thesis

Chapter I provides a broad overview of the thesis. It introduces the two pronged solution proposed to tackle the problem of acoustic inventory design in under-resourced languages. Chapter II describes prior boundary refinement procedures and the unsuitability of these techniques for our problem. It describes a novel boundary refinement procedure which removes the necessity of manually labelled data. Chapter III describes the problems in the currently used back-off synthesis techniques. It describes the scenarios where the existing back-off procedures fail and proposes a novel language dependent back-off procedure for Telugu language TTS systems. Chapter IV suggests the application areas which benefit from the proposed techniques and presents the conclusion. It also discusses the future work in this direction.

Chapter 2

Refinement of segment boundaries using landmarks

Missing unit synthesis involves the assembly of smaller sub-word units(SSUs) using the segmental and metrical information. This increases the number of concatenation points, thereby increasing the probability of insertion errors (see Figure 1-4) at these joins, due to boundary errors in the acoustic inventory. Even techniques like optimal coupling, which render these joins immune to boundary errors, are effective only when the error in the boundary is of just a few pitch periods. Hence to reduce the risk of unintended insertions, boundary refinement of the units is a necessary measure. In this chapter we propose a boundary refinement technique, to enhance the accuracy of boundaries of the units in the acoustic inventory.

2.1 Phone level segmentation of speech

Use of accurately labelled phonetic segments is critical for back-off synthesis. Further, the use of accurately annotated speech data enables development of high quality data driven synthesis techniques [36]. Speech segmentation is a major task in the annotation process. The use of manually segmented data for these accurate annotations has a prohibitive cost. Automatic segmentors are used to tackle this problem to some extent. In TTS databases, transcripts are available during the segmentation procedure. Hence explicit segmentation (text-dependent) procedures, which are better performing than the implicit segmentation (text-independent) procedures are used for TTS segmentation [37].

Hidden-Markov Model (HMM) based segmentation using force alignment is prominently used for explicit-segmentation. The flat start initialization [38] accuracy of HMM based segmentation is comparatively low. Hence manually labelled data or speaker-independent (SI) acoustic models are used as bootstrap for training the acoustic models used for segmentation [39]. In addition to this, methods such as [40], [41], [28], [42] have been proposed to further refine the boundaries predicted by the HMM segmentor. These refinement techniques can be broadly classified into two types:
1) *error correcting methods*, which reduce the systematic bias in the HMM boundaries
2) *alternate boundary detection methods*, which move the HMM boundary to alternate boundary candidates in it's neighbourhood
These refinement methods depend on manually labelled data. It is difficult to obtain manually labelled data, more so in the case of under-resourced languages. Hence alternate techniques are required to guide the label movement during the refinement process. This guidance is mostly obtained from spectral discontinuities in the locality of the HMM boundary. However several spectral discontinuities are observed in close proximity in several cases. Identification of the spectral discontinuity corresponding to the current HMM boundary is a non-trivial task, especially in cases where the initial HMM boundaries are highly erroneous. In this chapter a set of boundary type specific cues and cost functions are proposed to identify the spectral discontinuity pertaining to the current HMM boundary. This method exploits the knowledge of phone-class specific sub-band energy events to identify the relevant spectral discontinuity. Kim *et al.,* [39] also suggested the use of a phone-class specific spectral discontinuity measure for detection of boundary candidates and phone-class specific time windows to identify the spectral discontinuity corresponding to the current HMM boundary. However these time windows are empirically derived, which requires manual intervention. The

proposed method on the other hand eliminates the necessity for manual intervention by using large time windows, to tackle even severe HMM segmentation errors. The process of selecting the relevant boundary is instead guided by using phone-class specific spectral cues.

The chapter is organized as follows : Section 2.2 details the existing techniques for refinement. Section 2.3.2 describes the landmark based boundary refinement algorithm. Section 2.3.3 describes the proposed spectral cue based boundary selection procedure. Section 2.4 analyses the results and discusses improvements due to the proposed method. Section 2.5 presents a discussion on the impact of speech quality with increase in the accuracy of label boundaries.

2.2 Existing techniques

A variety of high accuracy speech segmentation techniques such as [42], [28], [43], [39], [40],[41] are available, which achieve segmentation accuracies as high as 96% with a 20ms error tolerance.

Antonio *et al.*, [42] and Matousek *et al.*, [28] proposed two different phonetic feature based boundary correction techniques using classification and regression trees (CARTs). These methods exploit the boundary specific systematic bias observed in HMM segmentation. Toledano [43] uses signal features to predict the alternate boundary candidate using neural networks. Another class of boundary refinement algorithms derive the refined boundary from the fusion of several boundary candidates. The boundary candidates are usually obtained either from spectral discontinuities [39], from HMMs trained for boundary detection [40] or from other HMM segmentors with different configurations and feature sets [41]. However all these systems invariably depend on manually labelled data.

To tackle the lack of hand-labelled data in under-resourced languages, a variety of techniques have been suggested to guide the label movement in the refinement process. Niekerk *et al.* [36] proposed the use of broad phonetic class label data from other language speech corpora for bootstrapping of HMM models. Kominek *et al.*

[44] proposed the use of mel-cepstral distortion (MCD) measure, rather than manual labels for guiding the label movement. Hoffman *et al.* [45] proposed a technique which uses frames in the middle of the automatically labelled segments to guide the label refinement. However none of these methods use the knowledge of boundary specific sub-band events, which are already well researched in the literature ([46],[2]), to bring the performance of the fully automatic methods closer to the supervised techniques. The use of this knowledge base helps tackle errors, which are undetectable using conventional spectral discontinuity based techniques. It helps distinguish between closely occurring spectral discontinuities and helps tackle large misalignment errors by the HMM segmentor.

2.3 Proposed method

The proposed method uses boundary specific sub-band event information for both detection and selection of the alternate boundary candidates. The boundary candidates are detected using spectral discontinuities in sub-bands specific to the current class of boundary. From these detected boundaries the ideal candidate is selected using cost functions whose elements are spectral cues extracted from the neighbourhood of the boundary and from the center of phones on either side of the boundary. The proposed method distinguishes closely occurring discontinuities which pertain to different HMM boundaries by using different cues for each class of HMM boundary. Further care is taken to accurately place boundaries around burst segments which are of really short duration, as even minor alignment errors could lead to a mislabelling of the burst segment. This might result in missing of bursts when these units are used for synthesis. Presence of bursts is critical for the perception of the stop phone in the synthesized output.

All the refinement techniques described in Section 2.2 are combined with the isolated HMM training procedure (described in [41]) to further increase the segmentation accuracy. The refined labels are used for isolated training, where each HMM model is initialized and iteratively re-trained, exclusively using the segments of the corresponding phone. The boundaries obtained using this combined training procedure are found to be superior [39]. Hence the proposed boundary refinement technique was also combined with isolated training.

The proposed segmentation process is depicted in Figure 2-1.

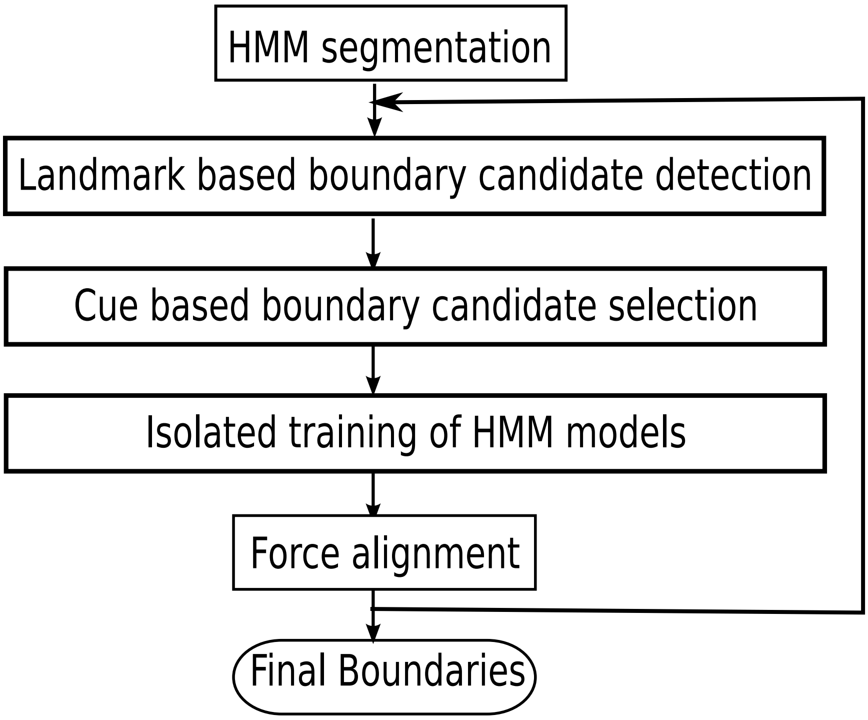
HMM segmentation
Landmark based boundary candidate detection
Cue based boundary candidate selection
Isolated training of HMM models
Force alignment
Final Boundaries

Figure 2-1: Flow chart of proposed segmentation process

2.3.1 HMM segmentation

HMM boundaries are the input to the boundary refinement procedure. Hence care is taken to attain maximum possible accuracy. Acoustic models for HMM segmentation were trained using embedded training technique with flat start initialization as the availability of manual labels or speaker-independent (SI) acoustic models is not assumed for bootstrap. Based on the analysis of segmentation performance with various HMM configurations described in [41], the HMM configuration was selected as three state left to right context-independent (CI) models, without a skip state (except for pauses), with one gaussian per state. The feature set used is 12 mel-frequency cepstral coefficients (mfcc) and normalised energy calculated with a frame length of

20ms and hop size of 5ms, along with delta and acceleration coefficients, resulting in a vector of length 39.

During the initial phases of experimentation it was identified that issues in labelling of stops and geminates were the major sources of errors in HMM segmentation. Sub-phoneme labels were used for stop closures and bursts, as they have different acoustic properties [47]. The separation of stops into closures and bursts must be accompanied by a mechanism for detection of stop stop interactions (SSI), which lead to incomplete stops (missing bursts or closures). Hence intra and inter word stop-stop interactions and other incomplete stop possibilities were considered as pronunciation variations during the alignment. Geminates, which are contiguous occurrences of the same consonant, were mapped to corresponding single consonant as they have spectro-temporal patterns similar to the single consonant except in the case of duration [41]. Forced alignment was done using these acoustic models to derive the initial segment boundaries. These boundaries are used as an input to the refinement procedure.

2.3.2 Landmark based boundary candidate detection

Landmarks are time points of lexically significant acoustic events. Consonantal landmarks represent instances of sudden signal change, e.g. consonant release or consonant closure (see Figure 2-2). Consonantal landmarks correspond to segment boundaries, hence the accurate identification of these landmarks implies increased accuracy in the detection of the phonetic segment boundaries. Hence these landmarks can be used for guiding the boundary refinement process. Chitturi *et. al,* [48] also proposed a landmark based boundary refinement procedure for HMM labels using multi-class support vector machines (SVM). However this method requires manual labels for training the landmark detection SVMs. On the other hand, the accuracies of signal processing based landmark detection methods are considerably low for use in high accuracy segmentation. However in the case of explicit segmentation, the expected landmark can be predicted from the phonetic transcript and the HMM boundary provides an approximate localization of the landmark. The proposed technique uses

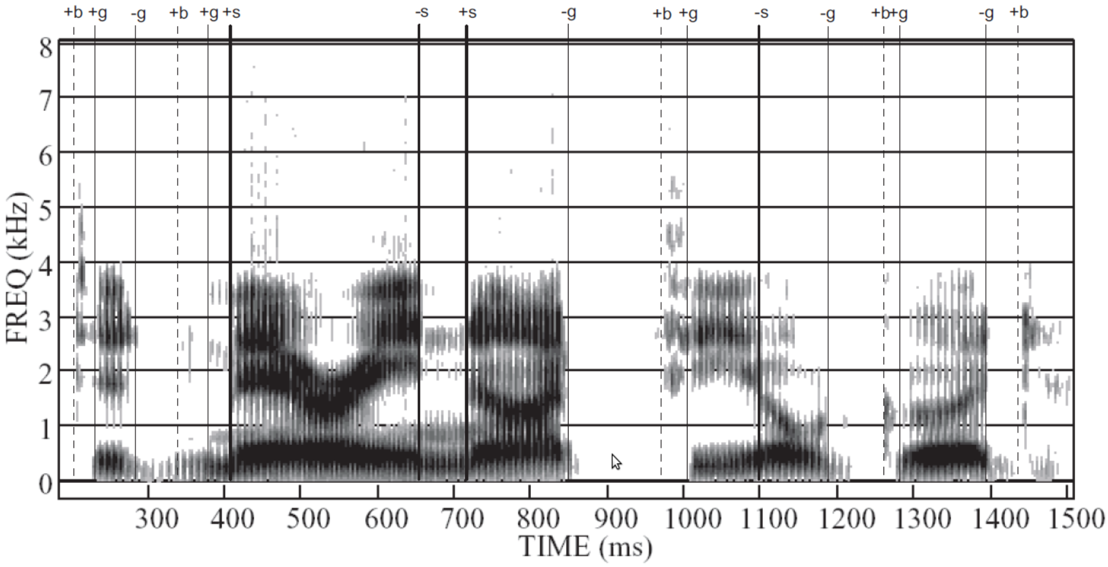

Figure 2-2: Examples of consonantal landmarks from an utterance “Did Mary not feel good?” [2]

this information for boosting the accuracy of the signal processing based landmark detection methods. The consonantal landmark expected at each HMM boundary is obtained using a mapping table described in [2] (Table 2.1). The use of this table requires the knowledge of the phone classes and their constituent phones in a language. In this chapter we do not discuss the refinement of inter-vowel boundaries and boundaries between vowels and glides.

The consonantal landmarks are classified into sonorant(s), burst(b) and glottal(g) landmarks. Each landmark is further divided into **‘+’** and **‘-’** based on the expected energy change across the landmark. An increase in energy across the landmark corresponds to a **‘+’** landmark and a decrease in energy corresponds to a **‘-’** landmark.

The position of the expected landmark is detected in the speech signal around the HMM boundary. The algorithm described here for landmark candidate detection is adapted from the algorithm described in [46]. In this method the consonantal landmarks are detected using discontinuities in corresponding sub-band energy curves. Frequency ranges of the sub-bands used for detection of each landmark type are given in Table 2.2.

Table 2.1: Mapping from pair of segment categories on either side of boundary to landmark type[1]

		Following Symbol						
		VOC	SON	FLP	IPP	HVO	FRI	SIL
Previous Symbol	VOC		-s	-s			-g	-g
	SON	+s		-s	+s		-g	-g
	FLP	+s	+s		+s		-g	-g
	IPP		-s	-s			-g	-g
	HVO						-g	-g
	FRI	+g	+g	+g	+g	+g		-b
	SIL	+g	+g	+g	+g	+g	+b	

[1] refer [2]
VOC = Vowel, Glides
SON = Nasals, Liquids
FLP = Flaps
IPP = Glottal Stops
HVO = Voiced /h/
FRI = Fricatives, affricates, stops, unvoiced /h/
SIL = Silences, closures

Table 2.2: Frequency ranges of the sub-bands used for detection of each landmark type

Band	Frequency Range (in Hz)	Landmark
1	0-400	g
2	800-1500	s & b
3	1200-2000	
4	2000-3500	
5	3500-5000	
6	5000-8000	

A search for discontinuities in corresponding sub-band energy curves, in the locality of the HMM boundary, using one-dimensional edge detection provides the required landmark candidates. Low thresholds are used in the edge detection algorithm to prevent the elimination of the desired landmark candidate from the set of identified candidates. A window w_{b_i} is selected around the i^{th} HMM boundary b_i for edge detection. This window is calculated as

$$w_{b_i} = [-1 * max(\frac{b_i - b_{i-1}}{2}, 50), max(\frac{b_{i+1} - b_i}{2}, 50)] \tag{2.1}$$

The minimum spread of the window on either side of the HMM boundary was taken as 50ms, to account for severe alignment errors of the HMM segmentor. Further, if the segment on the right side of the boundary is a stop burst, the right bound of the window is calculated till the middle of the phone following the burst. The short duration of bursts and the presence of stop closures in the left context of the bursts, results in a portion of the closure being assigned to the burst segment in HMM boundaries generated after embedded training [45]. Hence the actual stop burst does not necessarily fall at the mid of the HMM bounded segment.

The large time windows and low edge detection thresholds, used for boundary candidate detection, enable the inclusion of required landmark in the set of possible candidates. However they also result in a large number of candidates for each boundary.

2.3.3 Cue based boundary candidate selection

A cue based selection algorithm was proposed to select the boundary from the landmark candidates. Park [2] and Liu [46] proposed two different acoustic cue based techniques for selection of landmark candidates. Liu [46] uses empirically determined thresholds to eliminate extraneous landmark candidates. Park [2] uses the knowledge of manual boundaries to build probabilistic distributions of cues at segment boundaries. Both these methods necessitate manual intervention. Further these methods are designed for use in the automatic speech recognition (ASR) scenario where the exact nature of the landmark is not known prior to the selection. However as explicit segmentation provides prior knowledge of the expected landmark, we proposed a selection procedure that eliminates the necessity for manual intervention.

The proposed method is based on the hypothesis that the behaviour of acoustic cues in the immediate neighbourhood of the ideal boundary candidate, is similar to those at the center of the phones on either side of the boundary. Hence the absolute differences between the acoustic cues in the immediate neighbourhood of the ideal boundary and those at the center of the phones on the respective sides are minimal. Further as the ideal boundary has a spectral discontinuity, there is maximal difference of acoustic cues in the immediate neighbourhood. These differences are used as elements of a cost function. The boundary candidate (bc_j) which maximizes the cost function corresponding to the expected landmark is selected as the boundary.

A typical cost function is composed of three elements *viz.*,

1. $\epsilon_l = |\alpha_{IL} - \alpha_{CL}|$

2. $\epsilon_r = |\alpha_{IR} - \alpha_{CR}|$

3. $\epsilon_i = |\alpha_{IR} - \alpha_{IL}|$

where α is an acoustic cue calculated over a span of 10ms, in the immediate neighbourhood of the boundary candidate bc_j on left (IL) and right (IR) side, or over a span of 10ms at the center of the phone on the left (CL) and right (CR) side of the HMM boundary b_i.

Acoustic cue, α corresponds to one of the following

1. E_H = Average high band (1.2-8KHz) energy

2. E_G = Average glottal band (0-400Hz) energy

3. E_{5-6} = Average energy in bands 5 and 6 (see Table 2.2)

These cues are selected as they help in distinguishing different types of landmarks. A subset of these cues is used in each cost function, specific to the landmark type.

An ideal landmark candidate minimizes ϵ_l and ϵ_r while maximizing ϵ_i. Combining these elements in a simple additive unweighted cost function we have

$$C_{lm}(bc_j) = -\epsilon_l - \epsilon_r + \epsilon_i \tag{2.2}$$

where lm is the landmark type and bc_j represents the j^{th} boundary candidate.

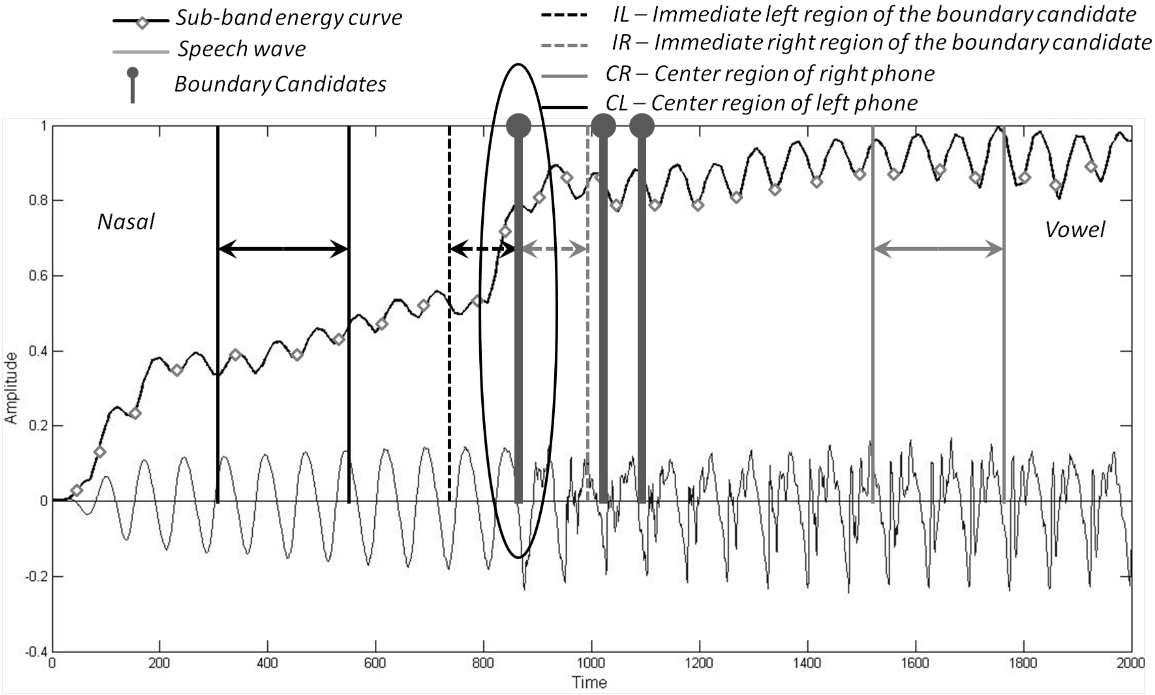

Figure 2-3: Elements of the cost function : The sub-band energy curve (S) is averaged in the identified regions (CL, CR, IL and IR). The absolute differences of these averages correspond to the elements of the cost function $C_{lm}(bc_j)$. $\epsilon_l = |S_{CL} - S_{IL}|$, $\epsilon_r = |S_{CR} - S_{IR}|$ and $\epsilon_i = |S_{IL} - S_{IR}|$. The encircled landmark candidate minimizes the cost function

The HMM boundaries, though erroneous, are used to identify the center region of the phone segments; as the HMM bounded phone segments have considerable overlap with the actual phone segments. However the average energies calculated from the center of the stop burst segment cannot be assumed to represent the burst landmark energy profile, as bursts do not necessarily fall at the center of the HMM bounded segment (see 2.3.2). Hence the cost functions corresponding to **b** landmarks are designed without considering the cues from the center regions (CL and CR). Instead, the high band energy of the closure segment before the burst is compared with the average high band energy of all silence segments in the current utterance (E_{sil}).

The **+g** and **+b** landmarks, corresponding to the scenario of a voiced phone following a stop burst, occur within a short interval. Cues in the glottal band (E_G) are insufficient to differentiate these two closely occurring landmarks, leading to boundary selection errors. Energy in the bands 5-6 (E_{5-6}) can be used to distinguish the **+g** and **+b** landmarks, as burst landmarks are accompanied by broadband noise present even in these high bands. Hence in the case where a **+g** landmark follows the **+b** landmark, a separate cost function is used to identify the ideal **+g** candidate.

Table 2.3 summarizes the details of elements in the cost function, for each landmark (LM). The cost functions are same for both the '+' and '-' type landmarks, unless explicitly specified.

Table 2.3: Elements of the cost function for each landmark type

	Elements		
LM	ϵ_l	ϵ_r	ϵ_i
s	$\lvert E_{H_{CL}} - E_{H_{IL}} \rvert$	$\lvert E_{H_{CR}} - E_{H_{IR}} \rvert$	$\lvert E_{H_{IL}} - E_{H_{IR}} \rvert$
+g	$\lvert E_{H_{CL}} - E_{H_{IL}} \rvert$	$\lvert E_{G_{CR}} - E_{G_{IR}} \rvert$	
+g after +b	$\lvert E_{H_{CL}} - E_{H_{IL}} \rvert$	$\lvert E_{G_{CR}} - E_{G_{IR}} \rvert$	$E_{5-6_{IL}} - E_{5-6_{IR}}$
-g	$\lvert E_{G_{CL}} - E_{G_{IL}} \rvert$	$\lvert E_{H_{CR}} - E_{H_{IR}} \rvert$	
+b	$\lvert E_{H_{IL}} - E_{sil} \rvert$		$\lvert E_{H_{IL}} - E_{H_{IR}} \rvert$
-b		$\lvert E_{H_{IR}} - E_{sil} \rvert$	$\lvert E_{H_{IL}} - E_{H_{IR}} \rvert$

2.4 Experiment

2.4.1 Experimental Setup

The TTS database used is a single speaker Telugu language database comprising of 4800 utterances (at 16KHz), with a total duration of 30.6 hours. The performance evaluation was done on five hours of manually labelled utterances in the database. The phone sequences for explicit segmentation were automatically predicted from the orthographic transcripts, leaving scope for undetected pronunciation variations. Hence during the calculation of accuracy within the tolerance limits (in ms), the manual and predicted phone sequences are aligned and the insertions and deletions are not considered. 98.5% of the phones were used for the calculation of the agreement percentages. The 1.5% errors in alignment were mostly due to wrongly predicted incomplete stops. The iterations of isolated training combined with landmark refinement were performed until the average shift of boundaries in successive iterations started increasing.

2.4.2 Results

Table 2.4 has a comparison of baseline HMM segmentation method (BL), the iterative isolated training method using embedded HMM labels (IT) (described in [37]), proposed landmark based refinement algorithm without iterative isolated training (LM) and proposed landmark based refinement with iterative isolated training (LM+IT). The accuracies of these segmentation techniques are measured as percentage of boundaries within 5ms, 10ms and 20ms deviation from the manual boundary.

Isolated training using baseline labels, disregarding the knowledge of the spectral cues, results in the HMM models capturing some errors in baseline labels introduced due to embedded training. This can be clearly observed in the performance of the isolated training procedure, in the placement of boundaries corresponding to **g** landmarks. However it can be seen that the LM refinement method increases the accuracy

Table 2.4: Performance of the refinement methods

	Deviation from manual boundary	BL	IT	LM	LM+IT
Total	5ms	28.9	25.9	34.8	37.0
	10ms	50.7	52.1	59.5	65.0
	20ms	73.3	81.9	82.2	88.6
s	5ms	38.5	40.4	37.0	44.2
	10ms	69.4	68.5	64.8	74.0
	20ms	94.0	89.8	89.4	93.4
g	5ms	35.3	19.9	44.5	45.1
	10ms	59.4	45.6	73.4	75.3
	20ms	76.2	78.4	88.7	93.4
b	5ms	9.6	25.6	18.9	22.8
	10ms	20.2	51.7	35.7	47.5
	20ms	48.9	85.3	66.8	82.4

BL = Baseline
LM = Landmark based refinement
IT = Iterative isolated HMM training technique

of boundaries to 88.7% ($< 20ms$). On the use of these refined boundaries in the iterative isolated training procedure the accuracy is boosted to 93.4%.

The overall accuracy (inclusive of even inter-vowel and boundaries between vowels and glides) of the baseline labels is 79.0% ($< 20ms$). On the inclusion of errors, in the placement of sub-phoneme boundaries before the stop bursts, the baseline accuracy falls to 73.3%. Using the proposed method the accuracy was increased to 88.6% (including the sub-phoneme stop burst boundaries). [45] and [39] report increases in accuracy to 88.4% and 94.8% respectively. However these refinement methods operate on labels whose baseline accuracies are 80.2% and 87.3% respectively. The current method on the other hand is able to operate on labels with initial accuracies as low as 73.3% due to the use of large time windows, during the detection and selection of boundary candidates. Application of large windows is made possible due to use of boundary specific cost functions. These cost functions help select the desired boundary even from a number of boundary candidates within the large time window.

The baseline accuracies in the placement of boundaries corresponding to **+g** landmark, following a **+b** landmark (**+g after +b**), are as high as 97.2% ($< 20ms$). This is a misleading statistic as 59.0% of these **+g** boundaries encroach on the preceding burst segment. Stop burst segments are typically of small durations ($\sim 25ms$) and a boundary error of even a few milliseconds could lead to mislabelling of the burst segment. Use of these mislabelled segments in unit selection synthesis, results in missing bursts in the synthesized output. As stop bursts are crucial for perception of stops, these mislabelling errors are detrimental to synthesis quality and have to be avoided during refinement. The proposed landmark based refinement algorithm reduces the error in the placement of the **+g** boundaries, while ensuring that the burst segment is not mislabelled due to encroachment. The proposed landmark based spectral correction technique ensures that these encroachments are reduced to 18.9% while maintaining the accuracy at 97.0% for this class of boundaries. The use of phone-class specific cue (E_{5-6}) ensures the proper assignment of these boundaries.

2.5 Impact on speech synthesis

Context-dependent units are usually used in unit concatenative synthesis. When context dependent units are available for concatenation, optimal coupling technique is used to perform the join. This technique makes the joins at phone boundaries robust to the segmentation errors within a few pitch periods [44]. Hence improved automatic segmentation which brings the boundaries within a few milliseconds of the manual boundaries would be beneficial in the cases where the context-dependent SSUs (smaller sub-word units) are available, to synthesize the missing syllable.

However in cases where the context-dependent units are not available, optimal coupling cannot be applied. Hence the joins are performed at comparatively inaccurate phone boundaries. This poses problems in two aspects:

1)During Viterbi search join cost plays an important role in the selection of units. Errors in boundary placement shift the boundary away from the transition, to more stable regions near the center of phones on either side. Join costs at these comparatively stable regions were found to be lesser than the join costs calculated at transition regions, when considering out-of-context units. Hence the unit selection process is biased to select the erroneous units within the database, despite the availability of accurately segmented units. Even the state of the art segmentation techniques have average error of 6ms at each boundary, hence the selection of two units with such erroneous boundaries would result in an unintended insertion of average duration 12ms at the join. This could lead to the perception of an additional phone and hence result in lexical ambiguity.

2) Joins at phone boundaries need to be done with special care as the transitions at the phone boundaries are crucial for perception of the segments on either sides of boundary. Further it is difficult to ensure perceptual continuity at these transitions, as these transitions need to be similar to those produced by human articulators.

Hence in the cases where these context-dependent units are not available better back-off techniques robust to even minor segmentation errors are necessary.

2.6 Summary

In this chapter we proposed a boundary refinement algorithm which does not necessitate the use of manually labelled data. We proposed the use of knowledge base, in the form of landmark specific cost functions, to guide the boundary movement during the refinement process. The use of the knowledge base helped to distinguish between closely occurring spectral discontinuities and to correct large alignment errors of the HMM segmentor. The overall segmentation accuracy was increased to 89% (within 20ms from manual boundary) from the baseline accuracy of 73%. Thus the proposed refinement procedure is suitable for segmentation of TTS databases in under-resourced languages where the initial HMM labels have low accuracies, due to lack of properly annotated speech corpora for boot-strapping the acoustic models.

The only resource necessary for applying the proposed method to a speech database in a new language is the knowledge of phone-classes and their constituent phones in the language, required to identify the expected landmarks.

However in the case of SSU back-off using out-of-context units even minor errors were found to be detrimental to the speech quality. Hence a more robust back-off technique is necessary for synthesis in scenarios where context-dependent units are not available.

Chapter 3

Synthesis of consonant clusters using epenthesis

SSU back-off technique synthesizes the missing syllables from smaller sub-word units (SSUs) like diphones, phones, etc. However the use of this technique is accompanied by two challenges:
1)Maintaining coarticulative consistencies
2)Avoiding insertion errors at phone boundary joins

The boundary refinement algorithm proposed in the Chapter II helps reduce the insertion errors, by refining the segment boundaries. In this chapter we propose a back-off technique to ensure the coarticulative consistency. This technique is motivated from second language acquisition phenomena seen in native speakers. It is shown that this technique not only helps maintain coarticulative consistencies but is also robust to the segmentation errors in the database which are inevitable even in the state of the art segmentors. Further, we identify exceptions where the proposed back-off solution is not applicable and suggest alternate strategies. Finally we provide a generic synthesis strategy for all missing units with various types of consonant clusters.

3.1 Need for back-off methods in TTS system

Coarticulative and contextual variations in speech sounds pose a major challenge for speech synthesis systems. Unit concatenative speech synthesis systems try to capture these variations by using larger sub-word units, *viz.,* syllables, diphones, etc., or by using context-dependent units. Coarticulation tends to be stronger within syllables than across syllables. Further syllabary model of speech production proposes that speakers have access to precompiled motor programs for articulatory gestures during the production of frequent syllables [49]. Hence TTS systems in languages like Japanese and Indian languages use syllable level units.

In the data-driven approach the quality of speech synthesis depends on the syllable coverage of the speech database. Hence the speech database is recorded using an optimally selected text to maximize syllable coverage [18]. However missing syllable units during synthesis are a frequent occurrence, especially for unrestricted text [15]. These missing units usually have unseen and complex consonant clusters (see section 3.3). The synthesis of units with these unseen consonant clusters is not a trivial problem.

Back-off methods are used to tackle these missing units. The back-off methods either substitute the missing syllable with other syllables in the database [31] or synthesize it using smaller sub-word units (SSUs) like diphones. However, even missing diphones are a common problem [11]. In these cases, diphone substitution or half-phone back-off is used to synthesize the missing diphones [32]. The design of both these back-off methods is not trivial, as they present a new set of problems. Phonetic segments within syllables are found to have relatively high coarticulation, compared to those across syllable. The SSU back-off method which forms the missing syllables from smaller sub-word units has to ensure that the units selected have the right coarticulative influences. This is difficult to ensure as coarticulative influences span across several phones, e.g., influence of lip protrusion of the vowel is seen across many preceding consonants [33]. In cases where the required context-dependent units are not available ensuring proper coarticulation is a major issue. Further, in cases where the required context-dependent units are not available units have to be joined at phone boundaries. Phone transitions are important for perceiving the identity of phones on either sides of the phone boundary. Hence it is not preferable to perform joins at these boundaries, especially within syllables where the phonetic segments have high coarticulation.

Unit substitution methods avoid the aforementioned problems, by using other syllable or diphone units in the inventory to replace the missing units. However it takes considerable effort to design the substitution rules. Efforts to build these rule bases require detailed perceptual studies when designing for new languages [31]. Even after these detailed studies, minimal pairs i.e., words which only differ in one phonological element and have distinct meanings, pose a problem when using substitution as a back-off. Hence the application of substitution rules should be done taking care that the meaning of the word does not change. Further diphone substitutions necessitate a change in the neighbouring units, thus creating a complex search problem [11].

In this chapter, we present a novel rule-based back-off method to tackle the problems of coarticulative inconsistencies and sensitivity to segmentation errors in acoustic inventories. The proposed technique is motivated from a perceptual and speech production phenomenon, known as vowel epenthesis, and emulates native speaker intuition in synthesis of the missing units. Hence it is found to be perceptually acceptable to the native speakers of the language. The rule-base is motivated from the L2 (second language) acquisition research.

This chapter is organized as follows. Section 3.2 describes the baseline TTS system and its shortcomings. Section 3.3 details the vowel epenthesis phenomenon in Telugu. Section 3.4 describes the issues in utilizing epenthesis as a potent back-off strategy in Telugu TTS systems, and proposes resolutions for these issues. Section 3.5 identifies the consonant clusters where epenthesis is rejected, identifies the probable reasons for this rejection and also identifies alternate strategies for synthesis of these specific consonant clusters. Conclusions are presented in Section 3.6.

3.2 Baseline TTS system

The baseline TTS system is a unit concatenation system, built using the Festival framework [11] for Telugu language. Unit type was chosen as syllable, as it was found to be the most suitable unit for Indian languages [16]. This system uses the SSU back-off technique using phone units to synthesize the missing syllables. The instances of each phone were automatically clustered using the CART based technique described in [25]. This is similar to using context-dependent phone units with context oriented clustering to cover the missing contexts. Optimal coupling was used to perform the unit joins wherever possible.

3.2.1 Database

Telugu has a phone set of 35 consonants and 15 vowels, with more than 10,000 possible syllables. The database used in our experiments in this chapter covers only 3,000 of these syllables, using 12 hours of speech recordings. Further simple syllables constitute a major portion of the database and it does not cover all the consonant cluster combinations. Thus providing us ample opportunity to test the back-off strategies, for synthesis of unseen consonantal clusters. Labels used in the voice building process were generated using automatic segmentors with average accuracy of 80% within 20ms deviation from the manual boundary.

3.2.2 Performance of Baseline TTS system

A MOS test was conducted to grade the speech synthesis quality of the baseline TTS system. 30 utterances belonging to three different domains were synthesized. These sentences contained 10 words, on average. The sentences belonging to these three domains were grouped in to sets and the order of these sets were randomized in order to avoid any learning effects. Naturally spoken sentences were also introduced into these sets, to normalize the scores given by the scorers [50]. There were 23 subjects for MOS tests. The domain-wise MOS scores are given in Table 3.1, along with the average number of back-off calls per sentence. The reduction in the quality of synthesis with increase in back-offs is evident.

Table 3.1: Domain-wise back-off rate and MOS

Domain	Average back-offs / sentence	MOS
Sports	5.6	2.46
Science	1.2	2.83
Story	1.1	2.89

Phone-based back-off strategies could not synthesize speech of acceptable quality with increase in frequency of back-off per sentence. Especially for text belonging to domains like sports, international news, etc., as such text contain large number of foreign words. These foreign words, as will be explained in Section 3.3, usually lead to missing units during synthesis. Increasing the coverage of the database to tackle these units is not a viable option.

Synthesis of these unseen consonant clusters, especially when the required context-dependent units are not available in the database, is a non-trivial problem. Perceptual experiments support the hypothesis that humans focus on the regions of abrupt change where a large amount of information is concentrated [2]. The regions of these abrupt changes include phone transitions. Jenkins [51] performed an experiment that supports the hypothesis that information near consonant-vowel boundaries plays an important role in classification of vowels. Further perceptual continuity plays a role in the extraction of segmental identity from acoustic patterns [52]. Hence the joining of out-of-context units, where it is difficult to ensure continuity at the transition regions, is not an optimal solution to the problem of synthesizing missing units. Hence the requirement of an efficient back-off technique arises.

The proposed synthesis technique is suitable for maintaining coarticulative consistencies in the synthesized speech even when the phonetic segments in the consonant cluster are not available in the required contexts, in the acoustic inventory. Further, the proposed back-off strategy is robust to the segmentation errors in automatically segmented databases, helping us ensure the quality of synthesis output.

3.3 Lessons from L2 acquisition research

Influx of new words into a language, due to borrowings from other languages is a fairly frequent phenomenon [30]. Phonotactics of a language is the permissible combinations of phonemes that can co-occur in that language. These borrowed words do not necessarily conform to the phonotactics of the synthesis language. Hence these words lead to new phone clusters.

During L2 (second language) acquisition, native speakers of Telugu break such new consonant clusters, through vowel insertion, to conform to the phonotactics of Telugu [53]. This phenomenon is known as vowel epenthesis. e.g., The English word "*bulb*" which is pronounced by Telugu speakers trained in English as [balb][1] [35], is pronounced as [bal**u**b**u**] by native Telugu speakers untrained in English. As the consonant cluster "lb" is new to native Telugu speakers, they perform an insertion of the vowel "u" to break it. Another "u" is also inserted after the word final stop consonant b, as words in Telugu do not end with stop consonants. These inserted vowels are called *epenthetic vowels*.

Epenthesis is widely observed phenomenon during L2 acquisition in several languages like Arabic [54], Japanese [55], Korean [56], etc. Rules to identify the identity and place of the epenthetic vowels (or phones) are also proposed in L2 acquisition research of these languages. Hence the proposed technique could be easily adapted to these languages.

Epenthesis which is used by native speakers to deal with complex and unknown phone clusters could also be emulated by the TTS systems to produce speech acceptable to the native speakers. Hence epenthesis could be explored as a potent back-off strategy for complex consonant cluster synthesis.

[1] written in IPA phonetic transcription

3.4 Consonant cluster synthesis using vowel epenthesis

Phonotactics helps us reduce the number of syllables or diphone units that need to be covered in the unit inventory of a language. However it is difficult to get a complete coverage of even this reduced set of syllables or diphones. Usually about 80% syllable coverage is achieved, even in languages with simple syllabic structure and further increase in coverage is not proportional to the increase in the size of the inventory when considering unrestricted text which has borrowed words. In the case of diphones, though it can be ensured that all phonotactically possible diphones occur in the database using carefully constructed sentences, many such diphones would not occur more than once in the database [15]. Hence missing units are a common problem in unit selection synthesizers.

The TTS system can emulate the epenthesis phenomenon to break the consonant clusters. The newly formed syllables after epenthesis are simple and conform to the phonotactics of Telugu, thus they can be found in the unit inventory with a greater probability. It is safe to assume that it would be acceptable for native Telugu speakers to hear a TTS system, which performs epenthesis (insertion) of the vowel emulating the native speakers. Thus the insertion of vowel “u” when faced with the missing unisyllable unit [balb] in the database, to produce the syllables [ba], [l**u**] and [b**u**] could be acceptable. This provides a new back-off strategy for the missing unit problem. However the resolution of the following issues is necessary to successfully incorporate vowel epenthesis as a potent back-off strategy for Telugu TTS systems :

1. How to determine the identity of the epenthetic vowel?
2. How to use epenthesis to simplify frequent clusters?
3. Is multiple epenthesis within a complex consonant cluster perceptually acceptable?
4. Is the performance of vowel epenthesis based back-off strategy better than other back-off strategies?

The following sub-sections detail these issues and the proposed resolutions for these issues.

3.4.1 Identity of the epenthetic vowel

If a consonant cluster in a syllable violates the phonotactic constraints of Telugu, it is broken using epenthesis. The identity of this epenthetic vowel is determined by the vowel harmony rule [53]. According to the vowel harmony rule, if the epenthetic vowel is inserted in a word medial consonant cluster, it's identity is dependent on the identity of the vowel following the cluster. If the epenthetic vowel is inserted in a word final consonant cluster, it's identity is determined based on the word final consonant. These rules are tabulated in Tables 3.2 and 3.3.

Table 3.2: Identity of word medial epenthetic vowel in Telugu

Following vowel	Epenthetic vowel
a,a:	a
i,i:,e,e:	i
u,u:,o,o:	u

Table 3.3: Identity of word final epenthetic vowel in Telugu

Word final consonant	Epenthetic vowel
non-palatal consonant	u
palatal consonant	i

In the case of words with inflexional suffixes, the epenthetic vowels are determined treating the root word and suffix as isolated words. Similarly in the case of compound words the epenthetic vowels are determined separately for each word. The knowledge of syllable boundaries is a prerequisite for epenthesis. The syllabification algorithm adopted in this chapter is explained in [53].

Epenthesis can be used to break consonant clusters within and across syllable boundaries. When occuring across syllable boundaries, the epenthesis rule is combined with another rule called the coda erasure rule [53]. For example the word

[porlutu] in Telugu is composed of the syllables [por],[lu] and [tu]. Epenthesis can be used to break the consonant cluster [rl] across the syllable boundary, to simplify the syllable [por]. The syllables post epenthesis are [po], [ru], [lu] and [tu].
[porlutu] ⟼ [por**u**lutu]
Hence in this chapter consonant cluster represents a group of consonants in a word without an intervening vowel, irrespective of the syllable boundary.

3.4.2 A reduced vowel for Epenthesis

The TTS system can emulate vowel epenthesis to tackle missing syllables. For example syllable [kloː] when missing, can be replaced with syllables [ku] and [loː] after epenthesis of the vowel [u] in the consonant cluster [kl]. However preliminary experiments showed that tackling all missing syllables using this strategy was not suitable. Some consonant clusters though frequent in current usage (e.g., [ʈɽ]), are not valid according to the phonotactics of Telugu. Breaking such clusters with vowel insertion, to tackle missing syllables, was found to be unacceptable, in preliminary experiments. In addition to this, the TTS system has to cater to Telugu speakers with various degrees of fluency in other languages. Informal experiments also showed that Telugu speakers fluent in other languages, like English, from which words are borrowed, do not prefer vowel insertions in the consonant clusters of those words. In order to deal with these issues, we propose the insertion of a reduced[2] form of a vowel, spanning just a few pitch cycles, in place of the complete vowel. We claim that the proposed method can be used to simplify all kinds of consonant clusters in Telugu. The experiments below are designed to test this hypothesis.

The first experiment was designed to understand if the vowel insertions, of a few pitch cycles, were perceptually acceptable in various kinds of consonant clusters. A set of 12 words containing biconsonantal clusters, both frequent and infrequent, were selected. This set also includes borrowed words. The bi-consonantal clusters contain transitions between consonants of various classes (glides, fricatives, obstruents, lateral

[2] Please note that the word reduction is not used here in the phonetics sense, and simply means reducing the number of pitch cycles in the vowel.

approximants, nasals, affricates). These words were synthesized by performing a vowel insertion in the consonant cluster, according to the previously mentioned ruleset (Tables 3.2 and 3.3). After synthesis of these words, the number of pitch cycles in the epenthetic vowels were reduced. Three samples containing zero, two and four pitch cycles of the epenthetic vowel were generated manually from each synthesized word. These three cases represent various degrees of vowel reduction[2].

Subjects were asked to rate the samples generated, for naturality and intelligibility, using the mean opinion score (MOS) scale. The results of the subjective study are summarized in Table 3.4. The MOS of the frequent category supports the hypothesis that insertion of even four cycles of the epenthetic vowels, do not lead to unacceptable synthesis due to perception of the extra vowel in well recognized and highly frequent consonant clusters. As it can be seen from Table 3.4, zero cycle insertions are the most preferred category, with high MOS scores. Zero cycle insertions are equivalent to picking a smaller sub-word unit from the immediate context of the epenthetic vowel, which ensures proper coarticulation within the resultant syllable. For example, the syllable [kloo] when missing is effectively being replaced with the smaller sub-word unit [k] having right context of [u] and syllable [loo], when the epenthetic vowel [u] is reduced to zero pitch cycles.

Usually TTS systems use syllable units with boundaries derived from automatic segmentors. These units have considerable segmentation errors. Due to these erroneous boundaries, it might not be possible to ensure reduction to zero pitch cycles. This leads to unintended insertion of a few pitch cycles of the vowel. The state of the art segmentors produce boundaries with an average error of 6ms. Assuming that the units on either side of join have 6ms of error the total insertion at the join is 12ms, which translates to 4 pitch cycles for average pitch ranges. Hence in the rest of the chapter we analyse the performance of the proposed method in the worst case scenario, by using four pitch cycle insertions of the epenthetic vowel. This is ensured by manually reducing the epenthetic vowel to four pitch cycles.

AB tests were conducted to see if subjects noticed considerable difference among the samples with zero, two and four pitch cycles of the epenthetic vowel, in the

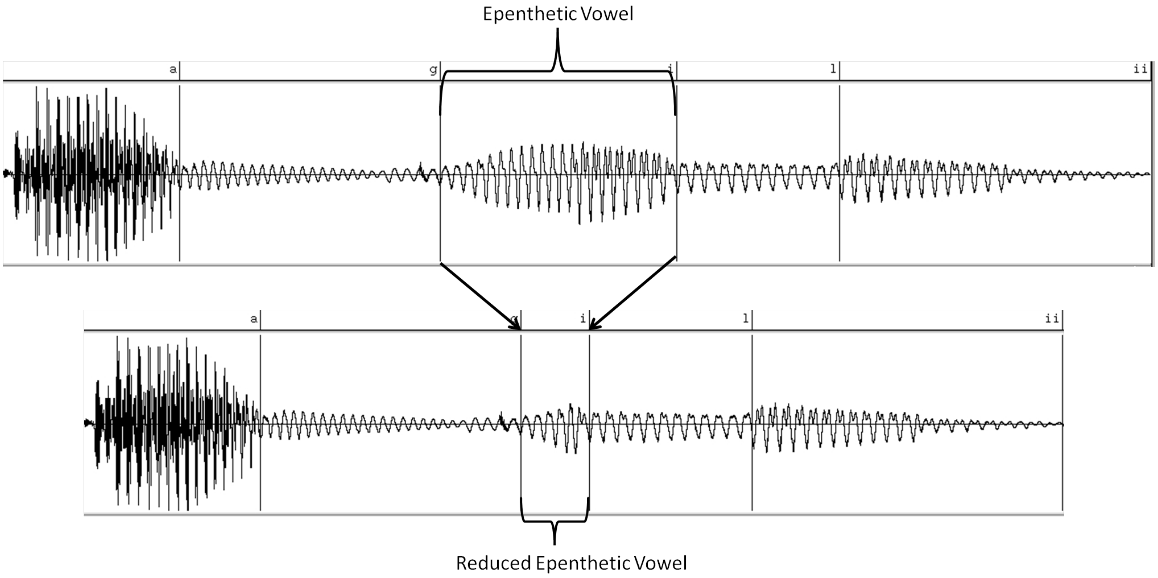

Figure 3-1: Reduction of the epenthetic vowel: The word [agliː] was synthesized as [ag**i**liː] after performing epenthesis. The epenthetic vowel [i] was reduced by removing pitch cycles from the synthesized waveform

words with bi-consonantal clusters. The results are tabulated in Table 3.5. It can be observed that the extent of insertion is perceptually indiscernible in most cases.

The above results show that vowel insertions of few cycles are acceptable in both frequent and infrequent bi-consonantal clusters. Thus the proposed method enables the use of epenthesis as a back-off strategy for missing syllables with both types of clusters.

Table 3.4: MOS scores[†] for various degrees of reduction in the epenthetic vowel

Word Type	MOS		
	4c*	2c	0c
Infrequent	3.80	4.08	4.29
Frequent	4.09	4.30	4.56

∗ c = pitch cycles

† No. of trials in each MOS test = 8 subjects X 12 words

Table 3.5: AB tests† for various degrees of reduction in the epenthetic vowel

	4c* vs 2c	4c vs 0c	2c vs 0c
Prefer 4c	5%	10 %	-
Prefer 2c	17%	-	10%
Prefer 0c	-	30%	20%
No Preference	78%	60%	70%

∗ c = pitch cycles

† No. of trials in each AB test = 5 subjects X 12 words

3.4.3 Epenthesis in tri and quadra consonantal clusters

In the previous experiment perceptual acceptability of the vowel insertion in bi-consonantal clusters was tested. However the main application of the backoff method is in simplifying syllables with complex clusters containing 3 or 4 consonants. Hence an experiment was designed to study multiple vowel insertions in words with tri and quadra consonantal clusters i.e., a cluster $C_1C_2C_3V$ is simplified with multiple vowel insertions as $C_1V_{ins}C_2V_{ins}C_3V$ where V_{ins} represents the reduced epenthetic vowel and C represents consonants. Each of the consonant-consonant junctures has four pitch cycles of vowel insertion (to represent worst case, as discussed previously). Ten words with complex clusters were synthesized using the proposed method, after being fitted into a sentence. The clusters in these words were broken by multiple epenthesis of reduced vowels. Subjects were asked to rate the samples generated, for naturality and intelligibility, using the MOS scale. The MOS score averaged over ten subjects was 4.1. Thus proving the applicability of the method for simplifying even complex clusters.

The results of the above subjective tests, support the hypothesis that vowel insertion can be used to deal with missing syllables, by creating simplified syllables and subsequently reducing the pitch cycles in the epenthetic vowels.

3.4.4 Comparison with conventional back-off methods

The proposed back-off method was compared with the conventional back-off strategy of synthesis using smaller sub-word units (SSUs). An AB test was conducted to identify the preferred method. The test consisted of ten subjects judging ten sentences with words containing complex clusters, synthesized by the candidate methods. Results summarized in Table 3.6 show a preference for the proposed method.

Table 3.6: Results of AB test † comparing proposed back-off method with SSU back-off technique

Prefer proposed method	59%
Prefer conventional back-off	17%
No Preference	24%

† No. of trials = 10 subjects X 10 sentences

The proposed method is also superior to the unit substitution method in two aspects. Firstly, when epenthesis is used for back-off the new phone sequence with the reduced vowel(s) may not result in a minimal pair with the original phone sequence of the word. Thus reducing the overhead of tracking minimal pairs. Secondly epenthesis is already well researched in several languages and this reduces the effort needed in creating the rule base for the proposed back-off method.

3.5 Analysis of rejection of epenthesis in specific consonant clusters

In all the perceptual experiments, described above, subjects consistently gave low MOS scores or chose conventional back-off techniques when presented samples with epenthesis in some consonantal clusters. This section has an analysis of these exceptions.

In experiment one of section 3.4.2, perceptual tests were conducted using samples synthesized with reduced vowel epenthesis in various bi-consonantal clusters. In the perceptual test it was identified that epenthesis in consonantal clusters of specific phone classes was given substantially low MOS scores by the subjects. These constitute :

1. *stop stop clusters*

   e.g., [aːptuɖu] ⟼ [aːp**u**tuɖu] (rejected)

   [ɟaʈkaː] ⟼ [ɟaʈ**a**kaː] (rejected)

   Subjects rejected samples with even minor degree of epenthesis in stop-stop clusters. Clusters with two stop consonants are separated by the syllable boundary. In case syllables on either side of this boundary are missing in the acoustic inventory, the application of coda erasure rule (refer 3.3) to break this cluster was rejected by the subjects. This rejection was observed for both voiced and unvoiced stops

2. *fricative stop clusters*

   e.g., [sʈaːp] ⟼ [s**a**ʈaːp] (rejected)

   [pepsi] ⟼ [pep**i**si] (rejected)

   [aʂʈaː] ⟼ [aʂ**a**ʈaː](rejected)

   [aːʃcarjamu] ⟼ [aːʃ**a**carjamu] (rejected)

   Reduced vowel epenthesis in fricative stop clusters was also rejected. Rao [53] has suggested that epenthesis is not suitable in clusters where the consonants are assimilated (i.e., change in identity of the consonants due to other consonants in the cluster). This could explain the rejection in some fricative stop clusters generated due to assimilation (e.g., when [ct] ⟼ [st] in some words). However it was observed that *reduced vowel epenthesis* was rejected in all fricative stop clusters irrespective of the assimilation status of the consonants in the cluster.

3. *nasal-homorganic obstruent sequences*

   [aɳʈa] ⟼ [aɳ**a**ʈa](rejected)

   [baɳɖi] ⟼ [baɳiɖi] (rejected)

   Sequences with nasal followed by a stop with the same place of articulation (POA) are known as homorganic nasal obstruent sequences. *Reduced vowel epenthesis* was rejected even in this cluster.

In addition to these epenthesis was also rejected in geminates. However geminates do not present issues in ensuring coarticulation as they are contiguous occurrences of the same phone. Hence they can be synthesized by extending the duration of the normal phone, in case a geminate is missing in the acoustic inventory. Thus they are not included in this analysis.

Olive *et al.,* [57] identify pairs of consonants with minimal coarticulation. These include the fricative stop clusters, fricative fricative clusters, stop stop clusters and nasal stop clusters. They proposed that selection of units from the immediate context of silence segments is sufficient to synthesize the clusters with minimal coarticulation and context-dependent units are not necessary.

Clusters with minimal coarticulation cover all the cluster types where epenthesis was rejected by subjects. Epenthesis is a mechanism used by native speakers to deal with consonant clusters which are difficult to coarticulate. Hence the presence of minimal coarticulation in production of above clusters, would not necessitate epenthesis. This could be a reason for the rejection of *reduced vowel epenthesis* in these clusters. Since native speakers do not use epenthesis to pronounce these clusters, they might perceive pronunciation of these clusters with epenthetic vowel, though of a short duration, as incorrect. Further as these clusters have minimal coarticulation, they can be synthesized with joins at phone boundaries and are less likely to have perceived discontinuities, at these joins [58]. Hence the *reduced vowel epenthesis* technique which was introduced in this chapter as a technique to ensure proper coarticulation, to avoid perceptual discontinuities and to avoid joins at phone boundaries, need not be applied in these clusters in the first place.

3.5.1 Synthesis techniques for rejected clusters

Stop-stop clusters show intergestural overlaps, known as *Stop-Stop Interactions* (SSIs), which lead to incomplete stops. These interactions can be observed even across word boundaries.An incomplete stop consonant is characterized either by an indistinguishable closure or a missing burst. A stop-stop interaction (SSI) is comprised of two incomplete stops and the acoustic events of an SSI typically resemble that of a complete stop i.e., one closure followed by a single burst. Though these sounds share similar spectro-temporal patterns with a single complete stop, they have some variations. The mean duration of closure in an SSI is consistently higher than that of a single complete stop consonant. Ghosh *et al.*, [59] have shown that closure duration is an important feature to distinguish complete stops from SSI-stops. Hence in the synthesis of stop-stop clusters care has to be taken to simulate the SSIs, by increasing the duration of the closure region. Increase in the duration of the closure, by insertion of a silence segment is sufficient for simulating the SSI in the synthesis of these clusters (for both voiced and unvoiced stops) [57].

Fricative stop clusters could be synthesized by performing joins at phone boundaries, as they have minimal coarticulation. However fricatives are known to be affected by anticipatory coarticulation due to following vowels [60]. Some characteristics, like influence of lip protrusion of the vowel, are seen even though there are intervening consonants [33]. Hence care has to be taken when combining fricative and stop units taken out-of-context, to ensure coarticulative consistency. The technique of joining segments from the immediate context of silence (described above) is preferable in fricative stop clusters as the presence of phrasal (or sentence) boundary at the silence inhibits any cross-word coarticulation effects on the fricative.

Nasal-homorganic obstruent sequences have low or zero amplitude at the transitional phase from the nasal to the stop closure. Hence the previously defined technique of selecting units from the context of silence segments is found to be sufficient for synthesis of even this cluster [57].

Hence technique of inserting silence or selection of units from the context of silence, is sufficient for synthesizing the consonant clusters, where epenthesis is rejected. Previously this technique was used to enable pruning of diphone databases in [57]. However the use of this technique was extended to complex consonant cluster synthesis in this chapter.

3.5.2 Evaluation

An AB test was conducted to compare the performance of the *reduced vowel epenthesis* and *silence insertion* techniques against the conventional SSU back-off technique. A variety of bi, tri and quadra consonatal clusters were synthesized using these techniques. However for the synthesis of most tri and quadra consonantal clusters the application of both these techniques was found necessary to synthesize the cluster. *Reduced vowel epenthesis* was used for synthesizing all the consonant-consonant transitions in a consonant cluster except those across stops, across fricatives, or across fricatives and stops; and from nasal to homorganic stop. For these specific transitions the *silence insertion* technique was used. Once again it was manually ensured that at least 4 pitch cycles(∼20ms) of each epenthetic vowel were present, to measure the performance of *reduced vowel epenthesis* technique in the worst case scenario (see Section 3.4.2).

A total of 22 words containing complex clusters were synthesized, after embedding them into three word sentences. Small sentences were chosen to enable the subject to focus on the consonant cluster, not being affected other artefacts of the TTS system. Subjects were asked to consider naturality and intelligibility of the consonant cluster when choosing the better sample. They were also given the option to state *no preference*. The results are summarized in Table 3.7.

The overall performance of the proposed techniques was comparatively higher than the conventional SSU back-off technique, both when applied in isolation or in combination. In some consonantal clusters subjects did not mind the phone insertion errors, caused by conventional back-off techniques. These errors were imperceptible in continuous speech due to comparatively shorter duration. This behaviour was especially evident in the samples synthesized for comparison with silence insertion technique. The 31.25% preference for conventional back-off techniques, in this scenario, was mainly due to two samples where the phone insertion errors though present were not rejected.

Finally, to synthesize quadra and tri consonantal clusters both the reduced vowel

Table 3.7: Results of AB tests comparing proposed back-off techniques, *reduced vowel epenthesis* and *silence insertion*, with SSU back-off technique

reduced vowel epenthesis vs conventional back-off [1]	
Prefer proposed method (epenthesis)	60.9%
Prefer conventional back-off	20.5%
No Preference	18.6%
silence insertion vs conventional back-off [2]	
Prefer proposed method (silence insertion)	51.25%
Prefer conventional back-off	17.5%
No Preference	31.25%
reduced vowel epenthesis + silence insertion vs conventional back-off [3]	
Prefer proposed methods	71.4%
Prefer conventional back-off	12.9%
No Preference	15.7%
Overall Scores [4]	
Prefer proposed method (epenthesis & other technique)	60.9%
Prefer conventional back-off	20.5%
No Preference	18.6%

1 No. of trials = 10 subjects X 7 sentences
2 No. of trials = 10 subjects X 8 sentences
3 No. of trials = 10 subjects X 7 sentences
3 No. of trials = 10 subjects X 22 sentences

epenthesis technique and silence insertion technique were necessary, as these clusters contain both types of consonant transitions. The combination of these techniques was preferred in 71.4% of the cases showing the suitability of the above technique to ensure quality of synthesis compared to the SSU back-off technique. In every single speech sample synthesized using this combination, the proposed method was rated higher than the conventional technique. Hence combining *silence insertion* technique with the *reduced vowel epenthesis* technique we can achieve synthesis quality above the conventional SSU back-off technique, especially in the case of synthesizing complex consonant clusters of higher order, which is the goal of this thesis.

The proposed techniques could also be used for synthesizing missing units of other types. The epenthesis phenomenon can also be exploited in missing diphone synthesis, using half-phones. Half-phone joins are done at phone boundaries, which are derived automatically. Optimal coupling necessary to make these joins robust to segmentation errors [25], cannot be done during half-phone back-off. Thus half-phone back-off involves the splicing of half-phone units at phone boundaries, with a small portion of the neighbouring phone due to segmentation errors, which in turn leads to traces of undesirable phone(s) insertions at the joins. It can be ensured that these insertions are the epenthetic vowels (or silence segments), by picking the half-phone units with the immediate context of the epenthetic vowel (or silence segments) from the inventory, wherever possible. The splicing of such half-phones at the phone boundaries produces a diphone, which would have at most few pitch cycles of the epenthetic vowel (or a small duration of silence) at the join. These minor insertions have been shown to be perceptually acceptable to the native speakers.

3.6 Summary

In this chapter the issues in synthesizing missing units were identified. Missing units contain complex consonant clusters which are difficult to synthesize. A rule-based back-off strategy motivated by the vowel epenthesis phenomenon was proposed to synthesize these clusters. Epenthesis of reduced vowels was proposed, to enable the application of this back-off strategy for larger set of missing syllables, corresponding to both frequent and infrequent consonantal clusters. The rule-set used in the proposed method is adapted from L2 acquisition research in Telugu, thus making the rule design a minimal effort process. AB tests conducted to compare the performance of the epenthesis based method with conventional back-off methods, showed a preference (59%) for the proposed method. The proposed method was found suitable to synthesize complex tri and quadra consonantal clusters while ensuring coarticulative consistencies. The proposed back-off method is also robust to segmentation errors of automatically segmented databases. Further few consonantal clusters where epenthesis could not (and need not) be applied were identified. We show that these specific clusters have minimal coarticulation and epenthesis need not be applied in these cases in the first place. An alternative technique of silence insertion was identified for the synthesis of these clusters. We showed that the combined application of these two techniques during back-off was preferred in 71.4% of cases when compared with conventional SSU back-off technique. The final technique which combines the application of the reduced vowel epenthesis technique and silence insertion technique, for the synthesis of complex consonantal clusters, is summarized in Figure 3-2.

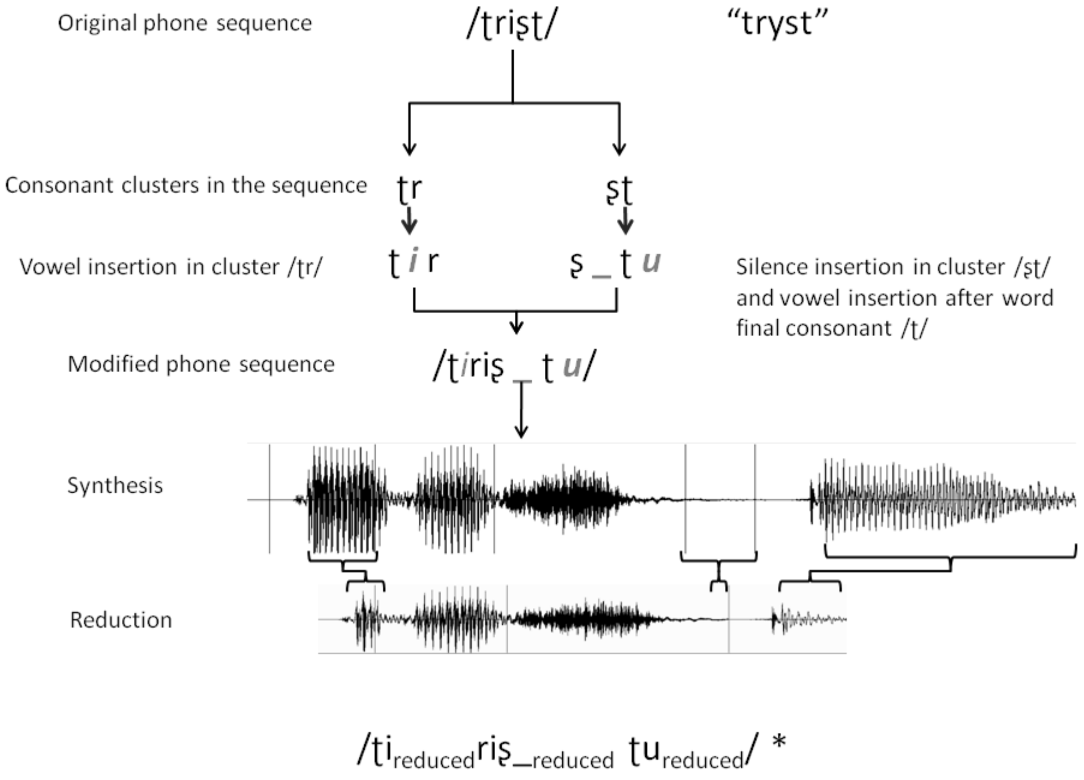
Original phone sequence
/ʈriʂʈ/
"tryst"
Consonant clusters in the sequence
ʈr
ʂʈ
Vowel insertion in cluster /ʈr/
ʈ i r
ʂ _ ʈ u
Silence insertion in cluster /ʂʈ/ and vowel insertion after word final consonant /ʈ/
Modified phone sequence
/ʈiriʂ _ ʈ u/
Synthesis
Reduction
/ʈi_reduced riʂ_reduced ʈu_reduced/ *

Figure 3-2: Synthesis of the english word "tryst": This word pronounced as [ʈriʂʈ], by native Telugu speakers trained in English, was synthesized as [ʈiriʂ_ʈu] after performing epenthesis and silence insertion. The epenthetic vowels [i], [u] and the inserted silence(_) were reduced in the synthesized waveform

The proposed method is not limited to Telugu language, as epenthesis phenomenon is widely observed in several other languages like Japanese, Swahili, Arabic, Korean, etc. The proposed technique can be easily adapted to these languages.

Chapter 4

Conclusion

In this thesis, a solution was proposed for resolving the data insufficiency issues in preparation and use of acoustic inventories of TTS databases. The data insufficiency issues were caused due to the LNRE (Large Number of Rare Entities) nature of syllable frequency distributions. The data insufficiency issues in acoustic inventory translate to missing units, in unit concatenation synthesis systems. The missing unit problem is exacerbated in the scenario of foreign word synthesis. Further, manually annotated speech corpora, which enable development of high quality acoustic inventories are not readily available in Telugu language, which is the language of interest in this thesis. Hence the objective of this thesis is to tackle the missing unit problem to enable high quality speech synthesis, using techniques which do not require manually labelled data and are robust to the errors in acoustic inventories.

The proposed two pronged solution, uses the techniques of boundary refinement and robust back-off synthesis to tackle the problem of missing units. The missing units are synthesized from smaller sub-word units. Combining smaller sub-word units requires tackling of two challenges, which are joins at phone boundaries and coarticulative inconsistencies. Joins at phone boundaries require units with refined boundaries, hence a boundary refinement algorithm was proposed. The proposed boundary refinement algorithm uses knowledge of phone class specific landmark events, rather than using manual boundaries, to guide the boundary movement. The boundary refinement method increased the accuracy of unit segmentation from 73% to 89% within a deviation of 20ms from the manual boundary. These refined units were suitable for SSU back-off when combined with the optimal coupling technique. However in many instances optimal coupling technique is not applicable due to lack of context dependent units in acoustic inventory. In these cases it was observed that even minor errors in unit segmentation lead to perceptible insertions when combining error ridden units in SSU back-off. Hence a back-off technique robust to these errors is required. Further it is difficult to ensure coarticulative consistencies when combining units which are taken out-of-context.

A novel back-off technique, inspired from second language acquisition phenomenon known as vowel epenthesis, was proposed to tackle the issues of coarticulative inconsistencies and unintended insertions, in SSU back-off. The coarticulative consistency is ensured by joining units from the immediate context of the same vowel. In this case boundary refinement errors lead to insertion of a portion of this vowel in the synthesized speech. The proposed technique tackles the errors in acoustic inventories, by ensuring that phone insertions due to segmentation errors are acceptable to the native speakers of the language. The identity of the vowel is chosen based on a rule set adapted from second language acquisition research in Telugu.

Using the two pronged strategy described above, the quality of synthesis of consonant clusters in missing units can be ensured. Missing units are a major problem in unrestricted text synthesis due to the inclusion of large number of foreign words or foreign phrases. Hence these techniques enable the use of the TTS system in the open domain scenario.

Further in the scenario with global exchange of information in multiple languages, it is not sufficient to create voices which can deal with just one language. Polyglot synthesis is thus an essential quality in next generation TTS systems. Polyglot synthesis opens up a variety of problems not commonly seen as severe in normal systems. Primarily the change in phonotactics across languages makes it difficult to collect phone clusters present in all these languages, in natural settings. Thus creating severe data insufficiency problems for data-driven synthesis techniques. The conventional techniques of back-off strategies in data-driven synthesis systems are not designed to tackle the frequent usage required by polyglot systems. The alternate techniques of using speaker and language adaptation techniques on multi-lingual and multi-speaker corpora, to tackle the problem of data insufficiency, do not provide a quality of synthesis on par with that from single speaker databases. Hence the task of designing efficient back-off techniques in the conditions of data scarcity is crucial for quality polyglot synthesis systems.

Building synthesizers using speech data from pre-recorded databases like audio books is another area of application. In many cases voices are built using pre-recorded speech databases which are not optimized for speech synthesis in terms of coverage. The techniques described above can be used here to ensure the voice quality even in this scenario.

4.1 Future Work

4.1.1 Use in acoustic modelling

In acoustic modelling for both speech synthesis and recognition, phone models are created after tagging them with context information (e.g., triphones), to efficiently model the coarticulative variations. However context dependent modelling presents the problems of data scarcity for several triphones. In the cases where insufficient number of samples are available for modelling, parameter tying technique is used. Proximity of the various clusters is determined in the feature space, and the combined samples of nearby clusters are used to determine a common set of parameters for the corresponding models. However the feature spaces currently used do not reflect the intricacies of the human auditory perceptual space.

The lack of sufficient data, in a database with optimal coverage, implies that the particular triphone is a rare triphone in the language. These triphones are rare or missing, mostly because the phonotactics of the language do not allow such a combination in the language. The process of parameter tying does not necessarily reflect the way in which humans process or produce such rare phone clusters. Humans simplify such clusters to fit them into the phonotactics of their language. Hence synthesis systems which emulate such behaviour in the acoustic modelling are expected to produce speech more acceptable to native speakers of the language.

Phone epenthesis is one such technique used by humans to tackle phonotactically illegal phone clusters. It is widely observed in many languages like Japanese, Telugu, Korean, etc. Experiments on using this phenomenon as a back-off strategy in unit concatenation synthesis systems in Chapter III, have proven the superiority of this method over the conventional back-off techniques of unit substitution or synthesis from smaller sub-word units. It was found that synthesis systems which tackle the problem of missing units in a manner emulating native speakers were preferred by the native speakers of the language. This is a promising result and similar behaviour is expected on application of this technique to HMM based speech synthesis systems too.

4.1.2 Pattern Mining in audio databases

In Chapter III we had seen how the knowledge of language specific phone cluster changes, like epenthesis phenomenon, can be used improving synthesizers. The use of these natural phenomenon is just limited to already recognized phenomena. Even in Telugu language, the rules described for identification of the vowel identity in the section 3.4.1 are just limited to the standard variety. The rules for determining the vowel identity are known to vary according to the region. Hence procedures to automatically identify and exploit these phenomena are necessary, to design synthesizers which cater to regional dialects. This is a problem of pattern mining from audio databases.

This pattern mining procedure requires a preprocessing step of efficient identification of pronunciation variations at microphonal level. Techniques for separating normal coarticulative variations or variations with respect to rate of speech, etc., from these variations of interest, are necessary. Representations suitable for the efficient detection and identification of these variations are to be explored. The detected pronunciation changes can then be used as an input to data mining techniques for association rule mining. Current research on pronunciation variation identification provides a good starting point to initiate the investigation.

List of Publications

1. Vijayaditya Peddinti, Kishore S. Prahallad, “ Significance of vowel epenthesis in Telugu Text-To-Speech synthesis”, in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May, 2011

2. Vijayaditya Peddinti, Kishore S. Prahallad, “ Exploiting phone-class specific landmarks for refinement of segment boundaries in TTS databases ”, accepted in *INTERSPEECH,2011*

3. Hema A. Murthy, ..., Vijayaditya Peddinti, Kishore Prahallad, “Building Unit Selection Speech Synthesis in Indian Languages: An Initiative by an Indian Consortium”, in *Proceedings of Oriental COCOSDA*, Kathmandu, Nepal, December 2010

4. Veera Raghavendra Elluru , Vijayaditya P. and Kishore S. Prahallad, “Speech Synthesis using Artificial Neural Networks”, in *Proceedings of National Conference on Communications* , Chennai, India, January 2010

5. Vijayaditya Peddinti, Kishore S. Prahallad, “ Synthesis of complex consonant clusters in Text-to-Speech systems ”, in preparation for submission to a journal

Bibliography

[1] K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, "A Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese," in *Eurospeech*, 1999, pp. 839–842.

[2] C. Park, "Consonant landmark detection for speech recognition," Ph.D. dissertation, MIT, Cambridge, Mass, Jun. 2008.

[3] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: the MITalk system.* New York, NY, USA: Cambridge University Press, 1987.

[4] D. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[5] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of the Acoustical Society of America*, vol. 82, pp. 737–793, 1987.

[6] R. Carlson and B. Granstrom, "A text-to-speech system based entirely on rules," in *Proceedings of ICASSP*, vol. 1, Apr. 1976, pp. 686 – 688.

[7] T. Dutoit, *An Introduction to Text-to-Speech Synthesis.* Norwell, MA, USA: Kluwer Academic Publishers, 2001.

[8] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[9] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, vol. 1, May 1996, pp. 373 –376.

[10] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proceedings of Eurospeech*, 1995, pp. 581–584.

[11] R. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[12] S. Kishore, R. Kumar, and R. Sangal, "A data driven synthesis approach for indian languages using syllable as basic unit," in *Proceedings of Intl. Conf. on NLP (ICON)*, 2002, pp. 311–316.

[13] F. J. Huang, E. Cosatto, and H. P. Graf, "Triphone based unit selection for concatenative visual speech synthesis," in *Proceedings of ICASSP*, vol. 2, May 2002, pp. II–2037 –II–2040.

[14] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proceedings of ICASSP*, Apr. 1988, pp. 679 –682 vol.1.

[15] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 57–71, 2003.

[16] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 1317–1320.

[17] J. P. H. Van Santen, *Progress in speech synthesis.* Springer Verlag, 1997.

[18] J. P. H. Van Santen and A. L. Buchsbaum, "Methods for optimal text selection," in *Proceedings of Eurospeech '97*, 1997, pp. 2–5.

[19] G. Fant, “Analysis and synthesis of speech processes,” *Manual of phonetics*, pp. 173–277, 1968.

[20] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[21] E. George and M. Smith, “Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model,” *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 5, pp. 389–406, 2002.

[22] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, “Speech reconstruction from mel frequency cepstral coefficients and pitch frequency,” in *Proceedings of ICASSP*, vol. 3. IEEE, 2002, pp. 1299–1302.

[23] H. Kawahara, “STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[24] J. Vepa, S. King, and P. Taylor, “New objective distance measures for spectral discontinuities in concatenative speech synthesis,” in *Proceedings of IEEE Workshop on Speech Synthesis*, 2003, pp. 223–226.

[25] A. W. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Proceedings of Fifth European Conference on Speech Communication and Technology*, 1997, pp. 601–604.

[26] J. Vepa and S. King, “Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1763 –1771, Sep. 2006.

[27] S. P. Kishore and A. W. Black, “Unit size in unit selection speech synthesis,” in *Proceedings of Eighth European Conference on Speech Communication and Technology*, 2003, pp. 1317–1320.

[28] J. Matoušek and R. J., "Automatic Pitch-Synchronous Phonetic Segmentation," in *Proceedings of INTERSPEECH.* Brisbane, Australia: ISCA, 2008, pp. 1626–1629.

[29] S. Pammi, M. Charfuelan, and M. Schroder, "Quality control of automatic labelling using hmm-based synthesis," in *Proceedings of ICASSP*, Apr. 2009, pp. 4277–4280.

[30] T. Bynon, T. Bowden, and B. Bunbury, *Historical linguistics.* Cambridge, NY: Cambridge University Press, 1977.

[31] E. Veera Raghavendra, B. Yegnanarayana, and K. Prahallad, "Speech synthesis using approximate matching of syllables," in *Proceedings of IEEE SLT Workshop*, 2008, pp. 37–40.

[32] J. Louw and M. Davel, "Halfphones: A Backoff Mechanism for Diphone Unit Selection Synthesis," in *Proceedings of 17th Annual Symposium of the Patt. Recog. Assoc. of South Africa*, vol. 29, 2006.

[33] R. Kent and F. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics*, vol. 5, no. 2, pp. 115—-133, 1977.

[34] A. Conkie and S. Isard, "Optimal coupling of diphones," *Progress in speech synthesis*, pp. 293–304, 1997.

[35] P. Lagedfoged, *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet.* Cambrdige University Press, 2000.

[36] D. R. van Niekerk and E.Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proceedings of the Interspeech*, Sep. 2009, pp. 880–883.

[37] I. Mporas, T. Ganchev, and N. Fakotakis, "A hybrid architecture for automatic segmentation of speech waveforms," in *Proceedings of ICASSP*, Mar. 2008, pp. 4457–4460.

[38] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[39] Y. Kim and A. Conkie, "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in *Proceedings of ICSLP*, Denver, Colorado, USA, 2002, pp. 145–148.

[40] S. Jarifi, D. Pastor, and O. Rosec, "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," *Speech Commununication*, vol. 50, no. 1, pp. 67–80, 2008.

[41] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech & Language*, vol. 24, no. 2, pp. 273–288, 2010.

[42] J. A. Antonio and A. Bonafonte, "Towards Phone Segmentation For Concatenative Speech Synthesis," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004, pp. 139–144.

[43] D. T. Toledano, "Neural network boundary refining for automatic speech segmentation," in *Proceedings of ICASSP*, Washington, DC, USA, 2000, pp. 3438–3441.

[44] A. W. Black and J. Kominek, "Optimising segment label boundaries for statistical speech synthesis," in *Proceedings of ICASSP*. IEEE, Apr. 2009, pp. 3785–3788.

[45] S. Hoffmann and B. Pfister, "Fully Automatic Segmentation for Prosodic Speech Corpora," in *Proceedings of INTERSPEECH*, 2010, pp. 1389–1392.

[46] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," Ph.D. dissertation, MIT, Cambridge, Mass, 1995.

[47] J. N. Marcus and V. W. Zue, "A variable duration acoustic segment HMM for hard-to-recognize words and phrases," in *Proceedings of ICASSP*, Washington, DC, USA, 1991, pp. 281–284.

[48] R. Chitturi and M. Hasegawa-Johnson, "Novel Entropy based moving average refiners for HMM Landmarks," in *Proceedings of Interspeech*, 2006.

[49] J. Cholin, W. Levelt, and N. Schiller, "Effects of syllable frequency in speech production," *Cognition*, vol. 99, no. 2, pp. 205–235, 2006.

[50] I. L. T. consortium, "MOS evaluation report," http://lantana.tenet.res.in/TTSconsortiumWiki/lib/exe/fetch.php?media=mos_evaluation_report.pdf, IIIT-Hyderabad, IIT-Madras, Tech. Rep., 2011.

[51] J. J. Jenkins, W. Strange, and T. R. Edman, "Identification of vowels in vowelless syllables," *Perception and Psychophysics*, vol. 34, no. 5, pp. 441–450, 1983.

[52] R. E. Remez, "Toward a principle of perceptual continuity in speech," *The Journal of the Acoustical Society of America*, vol. 67, no. S1, pp. S65–S66, 1980.

[53] G. U. M. Rao, "A nonlinear analysis of syllable structure and vowel harmony in Telugu," *PILC Journal of Dravidic Studies*, no. 6, pp. 55 – 84, 1996.

[54] P. Jordens and J. Lalleman, *Investigating second language acquisition.* Mouton de Gruyter, 1996.

[55] S. Peperkamp, I. Vendelin, and K. Nakamura, "On the perceptual origin of loanword adaptations: experimental evidence from Japanese," *Phonology*, vol. 25, no. 01, pp. 129–164, 2008.

[56] G. Iverson and A. Lee, "Perception of contrast in Korean loanword adaptation," *Korean Linguistics*, vol. 13, pp. 49–87, 2006.

[57] J. Olive, J. van Santen, M. Bernd, and C. Shih, *Multilingual text-to-speech synthesis: the Bell Labs approach.* KLUWER academic publishers, 1998, ch. Synthesis, pp. 191–205.

[58] I. Bulyko and M. Ostendorf, "Unit selection for speech synthesis using splicing costs with weighted finite state transducers," in *Proceedings of Eurospeech*, vol. 2, 2001, pp. 987–990.

[59] P. Kumar Ghosh and S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, 2009.

[60] S. Soli, "Second Formants in Fricatives: Acoustic Consequences of Fricative-Vowel Coarticulation," *Journal of the Acoustical Society of America*, vol. 70, no. 4, pp. 976–984, 1981.