MODELING PRONUNCIATION VARIATION FOR SPEECH RECOGNITION

A THESIS

submitted by

GOPALA KRISHNA ANUMANCHIPALLI

for the award of the degree

of

Master of Science (by Research)

in

Computer Science & Engineering

LANGUAGE TECHNOLOGIES RESEARCH CENTER

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD - 500 032, INDIA

February 2008

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

This is to certify that the work contained in this thesis titled **Modeling Pronunciation Variation for Speech Recognition** submitted by **Gopala Krishna Anumanchipalli** for the award of the degree of Master of Science (by Research) in Computer Science & Engineering is a bonafide record of research work carried out by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Date

Mr. Kishore Prahallad

Dr. Mosur Ravishankar

ACKNOWLEDGMENTS

I would like to thank Mr. Kishore Prahallad for his support and guidance throughout my research. He has helped me in more ways than he knows and I am immensely indebted to him. He has been a great friend and a mentor.

I am thankful to Dr. Mosur Ravishankar for supervising my work at CMU, which forms a major chunk of this thesis. I have greatly benifited from the long discussions with him. I thank him for being so approachable and for patiently answering my often naive enquiries.

I am grateful to Prof. Raj Reddy, for his care and advice during my stay at CMU. He has been a great source of inspiration since early years of my undergraduation and working with him was a dream come true. It was his readiness to spare time to see a demo or giving feedback that kept me working.

I am thankful to Prof. Rajeev Sangal for his support and encouragement. I also gratefully acknowledge the LTRC for financial support and for computational resources.

Of the many people at CMU I would like to thank are Dr. James Baker, Dr. Alan Black and Dr. Alex Rudnicky for ideas, resources and comments on my work. Thanks also to Ms.Vivian Lee for keeping me from starving.

I would like thank members of the speech lab and my undergraduate friends for their company and for keeping me sane during the last year. Special thanks to Nataraj for helping me debug my own code.

My biggest thanks go to my parents and sister who have believed in me and have always been cheerfully supportive of my decisions, despite circumstances.

ABSTRACT

Keywords: *automatic speech recognition; pronunciation modeling; baseform inference; automatic dictionary generation.*

Automatic Speech Recognition (ASR) is a sequential pattern recognition problem. It aims to correctly hypothesize a spoken utterance into a string of words. The conventional statistical framework employed to accomplish the speech-to-text conversion comprises three major components- acoustic models, language model and the pronunciation dictionary.

The pronunciation dictionary(or the lexicon) is a mapping table, a representation of the system's vocabulary in terms of its acoustic modeling units. In general, while acoustic and language models are outputs of statistical optimization procedures, lexicons are usually taken off the shelf. During the training, the system is provided with speech data, the corresponding transcription and a pronunciation dictionary. At the decoding run-time, the acoustic models and language models trained on the task are used while one of the standard dictionaries (CMUdict, Pronlex etc.) is used as the lexicon. Standard lexicons are manually built by linguists to provide the most generic pronunciations of the words. However, variations occur in pronunciation due to a number of factors including gender, accent, dialect, mode of speaking etc. While a generic pronunciation dictionary remains to be the safest bet, it may not be optimally suited for a test condition. Thus, there is a need to adapt a lexicon in order to best match the test conditions. This thesis shows that the lexicon can also be improved for a task by adding the variants to pronunciations, inferred using the resources already provided for acoustic model training.

The focus of pronunciation modeling research has to been to incorporate all kinds of genuine pronunciation variation into the speech recognition models. Several approaches operating at various components of the recognizers have been proposed to account for these kinds of variation. Existing methods fall into two broad categories. Modeling at the acoustic

level, (including front-end signal processing and model parameter adaptation given some adaptation data) and modeling at the lexical level, modifying the entries in the dictionary. This thesis falls into the latter category.

The primary contribution of this thesis is a generic framework for generation of pronunciation variants of words seen in the training data. The framework is realized by construction of a pronunciation grammar network along which to search for the variants used in the acoustics. The network itself is constructed via a statistical model (here decision trees) built to predict likely candidate phones of the word. The framework provides for adjusting the reliability of each source of information about the pronunciation: the acoustic evidence, the orthography and the phone transition patterns in the language. The three sources are effectively combined to score each candidate in the search space. The hidden variant of the pronunciation is inferred from the network via a Viterbi trace-back. The thesis presents a thorough analysis of the nature of the inferred variants and proposes criteria to select the best set of variants that when augmented to the lexicon will enhance the system's tolerance to such variations, thereby increasing its performance.

The thesis validates the proposed framework and techniques on three different tasks. It uses an isolated word task, the OGInames corpus for establishing the importance of orthography as the most reliable source of information about the pronunciation. The experiments involving evaluation of inferred variants are done on single speaker and multi-speaker continuous speech tasks, ARCTIC and TIMIT databases respectively. Among the two, the largest improvement was seen on the single speaker ARCTIC database, in which the automatically learned pronunciations corrected 14% of the errors made by the baseline system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	- Artificial Intelligence
AM	- Acoustic Model
ANN	- Artificial Neural Network
AP	- Acoustic-Phonetic
ASR	- Automatic Speech Recognition
CART	- Clustering And Regression Trees
CHMM	- Continuous Hidden Markov Model
CIHMM	- Context Independent Hidden Markov Models
CDHMM	- Context Dependent Hidden Markov Models
CD	- Consonant Deletion
CI	- Consonant Insertion
CS	- Consonant Substitution
CSR	- Continuous Speech Recognition
CV	- Consonant-Vowel
CWR	- Connected Word Recognition
DP	- Dynamic Programming
EM	- Expectation Maximization
FST	- Finite State Transducer
GMM	- Gaussian-Mixture Model
HMM	- Hidden Markov Model
SCHMM	- Semi-Continuous Hidden Markov Models
IWR	- Isolated Word Recognition
LER	- Lattice Error Rate

LM - Language Model

LVCSR - Large Vocabulary Continuous Speech Recognition

MAP - Maximum A posteriori Probability

MFCC - Mel-Frequency Cepstral Coefficients

ML - Maximum Likelihood

OOV - Out Of Vocabulary

PDF - Probability Density Function

PER - Phone Error Rate

PLP - Perceptual Linear Prediction

SVM - Support Vector Machine

TIMIT - Texas Instruments and Massachusetts Institute of Technology

VD - Vowel Deletion

VI - Vowel Insertion

VS - Vowel Substitution

WER - Word Error Rate

WFST - Weighted Finite State Transducer

CHAPTER 1

Automatic Speech Recognition

The problem of speech recognition is defined as the conversion of spoken utterances into textual sentences by a machine. An utterance that is given as input to an Automatic Speech Recognition (ASR) system is digitized and processed using signal processing algorithms to extract representational vectors $X = \mathbf{x_1}, \mathbf{x_2}...\mathbf{x_t}$, where t depends on the length of the utterance. If the hypotheses space of word sequences is ζ, the problem of speech recognition can be formally stated as

$$W^* = \underset{W \epsilon \zeta}{\operatorname{argmax}} P(W|X) \tag{1.1}$$

In other words, an ASR system tries to find a string of words W^* that has the highest probability for the given acoustic waveform. The direct computation of posterior probability $P(W/X)$ is difficult and hence Bayes' rule is applied to split $P(W/X)$ into realizable subcomponents:

$$W^* = \underset{W \epsilon \zeta}{\operatorname{argmax}} P(W|X) = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

$$= \underset{W}{\operatorname{argmax}} P(X|W)P(W) \tag{1.3}$$

The prior $P(X)$ is constant over all candidates of the hypotheses space and hence may be ignored in the denominator of Equation 1.2. From Equation 1.3, it can be observed that an ASR system needs to model two probability distributions: (1) the probability of the acoustics matching a particular hypothesis $P(X|W)$, and (2) the prior probability of the candidate hypotheses $P(W)$. The estimation of likelihood $P(X|W)$ involves modeling the relationship between the acoustic sequence and all possible word strings which is computationally expensive.

Since hypothesis W is a sequence of words $W = w_1, w_2...w_N$,

$$\max_W P(X/W)P(W) = \max_W P(X/w_1, w_2...w_N) * P(w_1, w_2...w_N) \tag{1.4}$$

Words in utterances are represented by sub-word units called *phones*. If w_i is a word model then $P_i = \mathbf{p_1}, \mathbf{p_2}...\mathbf{p_n}$ is corresponding sequence of phones. Each phone may further be realized as a sequence Q of a defined number of states, $Q = \mathbf{q_1}, \mathbf{q_2}...\mathbf{q_s}$. Thus models of utterances are deconstructed into a phone state sequence Q_i. Three different terms that comprise the entire probability distribution are:

$P_A(X|Q_i)$: The probability of acoustics given the phone state sequence (known as the *acoustic model*)
$P_P(Q_i|W)$: The probability of a state sequence given the words (the *pronunciation model*)
$P_L(W)$: The prior probability of word sequences (the *language model*)

These three models, P_A, P_P, P_L are related to Equation 1.4 as follows :

$$\max_W P(X/W)P(W) = \max_{Q_i}\{P(X/Q_i)P(Q_i/W)P(W)\} \tag{1.5}$$

$$= \max_{Q_i}\{P_A P_P P_L\} \tag{1.6}$$

Equation 1.5 follows from probability theory and the assumption that acoustic likelihood is independent of word models given the state sequence. To decode the best state sequence according to Eqn. 1.6, a *Viterbi approximation* is often employed. The Viterbi algorithm tries to find the state sequence which has the highest posterior probability $P_A P_P P_L$ on the observations. Hence, it is also referred to as the maximum *a posteriori* (MAP) decoding.

The typical ASR system has different components that estimate each part of the model. To begin with, acoustic features ($X = \mathbf{x_1}, \mathbf{x_2}...\mathbf{x_t}$) of the acoustic signal are produced by signal processing routines. MFCCs (Mel Frequency Cepstral Coefficients), a typical choice of representation for speech recognition are used in this work.

1.0.1 Acoustic Model

The factor P_A of Eqn. 1.6 is a model for the acoustics. It can be calculated in several ways. In Hidden Markov Models (HMM) systems, the distribution at any state q is modeled as a Gaussian Mixture Model (GMM). The distribution over each state can be determined for individual acoustics and phones $(P(x_t|q_j))$, the context independent (CI) models; these estimates are multiplied together to give an overall estimate of the probability $P(X|Q)$. Formally, the output distribution GMM at each state q is modeled as:

$$P(x_t = x|q_j = q) = \sum_{i=1}^{K} c_{i,q} N(x, \mu_{i,q}, \Sigma_{i,q}) \tag{1.7}$$

where K is the number of Gaussian components, $c_{i,q}$, $\mu_{i,q}$ and $\Sigma_{i,q}$ are the mixture weight, mean and covariance matrix of the ith component of the observation distribution of state q, respectively, and each Gaussian is modeled as

$$N(x, \mu_{i,q}, \Sigma_{i,q}) = \frac{1}{(2\pi^{D/2})|\Sigma_{i,q}|^{1/2}} exp\{\frac{-1}{2}(x - \mu_{i,q})^T \Sigma_{i,q}^{-1}(x - \mu_{i,q})\} \tag{1.8}$$

The state level emission densities *N* and the state transition probabilities are estimated during the acoustic model training. The data usually available during training is the speech data and its corresponding transcription. The state level information is assumed to be hidden. The Forward-Backward algorithm, also called the Baum-Welch algorithm [1] is employed to compute the model parameters given the training data under the Maximum Likelihood (ML) criterion. Baum-Welch is an instance of the iterative Expectation Maximization (EM) algorithm. [2] may be referred for a formal treatment of the EM algorithm.

1.0.2 Language Model

The language model (LM) provides an estimate of P_L of Eqn.1.6. It is typically an *n*-gram grammar for large-vocabulary decoders. In general, the probability of a word sequence W can be decomposed as follows:

$$P_L(w_1...w_t) = P(w_t|w_{t-1}, w_{t-2}..., w_1)P(w_{t-1}|w_{t-2}, ..., w_1)...P(w_1) \quad (1.9)$$

$$= \prod_{i=1}^{t} P(w_i|w_{i-1}, ..., w_1) \quad (1.10)$$

An n-gram grammar makes the assumption that word histories more than $n-1$ words before the current word do not affect the probability:

$$P_L(h_1...h_t) \approx \prod_{i=1}^{t} P(h_i|h_{i-1}, .., h_{i-(n-1)}) \quad (1.11)$$

Language models play an important role in constraining the search space of the decoder. It also helps in disambiguating between acoustically confusing word sequences (e.g. 'I scream' vs 'Ice cream'). Smoothed word trigrams with back-off are the most common LMs nowadays. Recent attempts claim significant improvements by employing higher order models [3].

1.0.3 Pronunciation Dictionary

For the most part in this thesis, the auditory front-end, the acoustic model and the language models are assumed as given. This thesis is concerned with the pronunciation model, P_P of Eqn. 1.6. The pronunciation model serves an important role: it acts as an interface between acoustic and language models, creating mappings between the two. The pronunciation dictionary determines how the acoustic modeling units are concatenated. The HMM phone models give the distribution and durational constraints for the individual phones.

In most systems, the dictionary is a mere look-up table, providing phonemic representations[1] of each word. Words may have more than one representation, in which case the table is called as a *multiple pronunciation dictionary*. A multiple pronunciation dictionary provides a model of baseform sequences, $P_B(B|W)$, as part of the overall pronunciation model.

$$P_P(Q|W) = P_D(Q|B)P_B(B|W) \quad (1.12)$$

[1] Phonemic representations are in the dictionary are also called *baseforms*

Where $P_D(Q|B)$ is the prior probability on the baseforms. Most dictionaries assume no prior bias on a baseform, although it may be helpful for some applications. In general, the baseform pronunciations of a word $P_D(Q|B)$ are assumed to be independent of the word context.

1.1 DECODING: RECOGNIZING THE SPEECH

The language model can be represented as a Markov chain, and since the acoustic model itself is HMM-based, the joint model can be realized as a single large HMM. MAP decoding according to Eqn. 1.6 is employed to search this huge network for the most likely path given the acoustic observations. Usually a form of Viterbi decoding [4] is used to obtain the MAP hypothesis.

Most formulations of the Viterbi algorithm entail traversal through a HMM graph. The huge pronunciation network is usually constructed and dynamically replicated at run-time according to the language model. While traversing along the graph, each state j at the time instant t is associated with a likelihood of the best path that ends in the current state j. The likelihood is computed having seen all observations until the time instant t (the forward algorithm). Back pointers are also stored to give the most likely state sequence(s) for each state j. Beam search [5] is often applied to prune candidate paths at each state for further processing. Once all frames of speech are processed ($t = T$), trace back information of the maximum scoring state is used to recover the most likely word sequence. The trace back information may also be used to produce a compact representation (known as a *lattice*) of the candidate hypothesis. An *n*-best list of n most likely hypotheses may be generated. The lattice/list may be rescored using higher knowledge sources to get the best word sequence(s). Several decoding strategies can be found in the literature [6] [7] [8].

1.1.1 Evaluation Criteria

ASR systems are usually evaluated under the Word Error Rate (WER) criterion. The WER metric is defined to be the ratio of the number of recognition errors to the number of words

in the reference (truth). The number of recognition errors is calculated as the minimum number of insertion, substitution or deletion operations required to obtain the same string as the reference from the recognizer output (hypothesis). This WER metric is an instance of the Levenshtein distance measure [9] computed using dynamic programming techniques. Human transcriptions are usually taken to be the reference.

$$WER = \frac{Substitutions + Insertions + Deletions}{\#\ words\ \ in\ \ reference} \tag{1.13}$$

Based on the associated task, Phone Error Rate (PER) and Lattice Error Rate (LER) may also be relevant performance measures.

1.1.2 Errors made by speech recognizers

Though ASR research has come a long way, today's systems are far from being perfect. Speech recognizers are brittle and make errors due to various causes. [10] attempts a detailed characterization of errors made by speech recognizers. Accordingly, most errors made by ASRs fall into one of the following categories:

1. *OOV errors*[2]: Current state of the art speech recognizers are closed vocabularies. So, they are incapable of recognizing words outside the system's vocabulary. Besides mis-recognition, the presence of an OOV in an input utterance causes errors to its neighboring words. In a large vocabulary system, each OOV is known to cause about 1.2 errors in the decoder's output [11].
2. *Search errors*: This class of errors is due to pruning of the candidate hypotheses by beam search (Sec. 1.1). It may be possible that the correct hypothesis is pruned because of a low score (this can be caused by multiple reasons).
3. *Homophone Substitution*: These errors are caused if more than one lexical entry has the same pronunciation (phone sequence). While decoding, they may be confused with one another causing errors. In general, the language model disambiguates in the event of such a confusion.

[2]Out of Vocabulary words are referred as OOV

4. *Language model bias*: Because of an undue bias (effected by high language weight) towards the language model, the decoder may be forced to reject the true hypothesis in favor of a spurious candidate. These errors may occur along with analogous acoustic model bias.

5. *Multiple acoustic problems*: This is a broad category of errors comprising those to due to bad pronunciation entries; disfluency, mis-pronunciation by the speaker himself or confused acoustic models (possibly due to noise, data-model mismatch etc.)

1.2 ISSUES ADDRESSED IN THIS THESIS

The research in this thesis attempts to improve the pronunciation dictionary. In part, the thesis addresses the last category of errors in the preceding section. As would be elaborated in the next chapter, earlier attempts to this problem have been sub-optimally implemented and only a few have shown improvements on complex speech tasks. This thesis addresses the problem in the context of continuous speech in a fully data-driven fashion. The first stage is inference of pronunciation from available data and information. This is followed by selection of the right variants for improving the dictionary.

Existing attempts follow a sequential step-by step procedure to inferring pronunciation. The time complexity in these approaches is a function of the number of variants and the length of the acoustics, iterated more than once. The current work employs a framework which infers pronunciation in a time synchronous manner. Thus, significantly reducing the time complexity, the proposed technique only requires a linear time in the length of the utterance. In brief, the technical contributions of this thesis are as follows:

- Establishing the significance of orthography as a reliable information source than acoustic evidence in determination of surface form pronunciations.
- An integrated framework for accurate time synchronous pronunciation inference from acoustics via the use of a novel combination function to combine and tune the information from available knowledge sources.

- Modifications to the Levenshtein distance to measure distance between phonetic strings to select suitable pronunciation variants for augmenting to the ASR lexicon.
- Significant reduction of continuous speech recognizer WERs on ARCTIC and TIMIT databases, the largest tasks so far studied for this problem.

1.3 MOTIVATION

There are several reasons that underline the need for automatic data-driven dictionary improvement/generation techniques:

- *Significance:* In a cheating experiment, [12] shows that word error rate on switchboard corpus dramatically decreased from 40% to 8% if the dictionary pronunciation matched the actual pronunciation. This proves that dictionary improvement is a promising direction for significant error rate reductions.
- *Adequacy:* [13] analyzed a corpus of conversational speech and identified that the baseform pronunciations are quite inadequate. For instance, the word 'that' appears 328 times in the corpus used and has 117 different realizations (variations in pronunciation). Also, the most frequent variant only covers 11% of all instances. Hence, there is a need for addition of better representative baseforms of the word pronunciation. Also, variants should be economically added so that improvement is not offset by the added confusibility due to the new lexical entries.
- *Consistency:* In [14], the switchboard corpus was phonetically annotated and human labelers disagree on more than 20% of the surface forms. This alludes to the fact that manually built dictionaries have a drawback of being inconsistent. This calls for principled ways for automation to impart more consistency to the dictionary building process.

Adaptation techniques for acoustic and language models are thoroughly researched and are put to practice in deployed real-time systems. As for the pronunciation model, one of the standard available dictionaries is plugged into the system, regardless of the task. This is

because, most earlier techniques have shown only marginal improvements on lexica. Also, most analysis was limited to isolated speech tasks. [15] argues that improvements in pronunciation modeling research have been elusive because most kinds of variations that were studied are already captured by context-dependent acoustic modeling. This thesis focuses solely on understanding the within-word variation in the lexical baseforms relevant to the task using context independent models.

Another motivation for inference and enrollment techniques is their ability to handle OOV words. This is in the context of applications which allow corrective feedback from the user. In such scenarios, data-derived pronunciations of a new word can be augmented into ASR lexicon. This enables the system to hypothesize the word in a subsequent encounter. The larger implication is the viability of systems supporting open/dynamic vocabularies, a sought after feature in ASR systems.

1.4 ORGANIZATION OF THE THESIS

The remaining chapters are organized as follows:

- Chapter 2 presents a survey on techniques for capturing variation via pronunciation modeling. It also discusses the limitations of existing attempts and an overview of the proposed approach in this thesis.
- Chapter 3 proposes the technique for inferring surface forms[3] using lexical and acoustic information. The chapter presents the decision tree based technique for capturing grapheme-to-phoneme relations. This is followed by empirical observations on the nature of the inferred surface forms with varying weights (confidence) on the different information sources, within an isolated word task.
- Chapter 4 presents an integrated framework with adjustable weights to combine the different information sources. Criteria for improved inference and rejection of spurious candidate variants are also proposed. The performance of the variants is pre-

[3]This thesis uses the terms surface form and pronunciation variant interchangeably

sented on two continuous speech tasks.

- Chapter 5 presents the concluding remarks and outlines the contributions of the thesis. The chapter also presents the limitations and assumptions made in the thesis.

CHAPTER 2

Approaches to Pronunciation Modeling

For humans, knowledge about pronunciation is intuitive and effortless. To this day, the faculties that inherently give us this intuition are not completely understood. For inquiry into pronunciation, linguistics dedicates two related sub-fields: *Phonetics* and *Phonology*. *Phonetics* deals with the range of vocal sounds that are produced during spoken language generation while *Phonology* deals with capturing the variation of pronunciation within these sounds.

In automatic speech recognition systems, information about pronunciation is captured via the pronunciation dictionary. While alternatives exist, the commonly used representation of pronunciation in ASR systems is a first order Markov chain of the phonemes. Fig. 2.1 shows example pronunciations of words *one*, *two* and *three*.

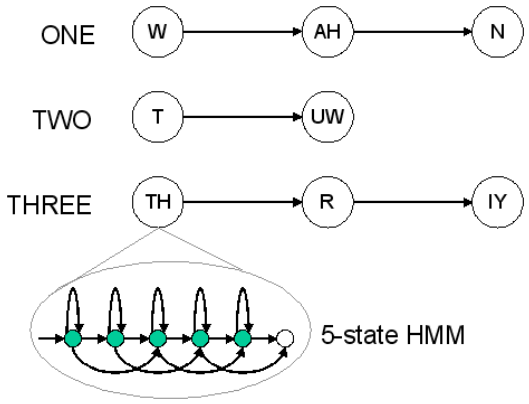

Fig. 2.1: Pronunciations as first order Markov chains

Figure 2.1 also shows the realization of the phoneme /th/ as a HMM at the state level, its acoustic model. As mentioned in chapter 1, the dictionary is an interface between the acoustic models and the language model. It gives the composition of each word in terms of the acoustic modeling units.

2.1 WHAT KIND OF VARIATION IS DEALT IN THE LEXICON?

As an interface between the acoustic and language models, a model for pronunciation must deal with variation from both sides: variation in pronunciations caused by factors such as predictable word sequences or increased speaking rate etc. The pronunciation dictionary together with the acoustic models enables the system to handle pronunciation variation. Linguistic variation includes a number of factors known to dynamically affect pronunciation variation, including the surrounding phones, the prosodic/accent context, the identity and probability of neighboring words and the presence of disfluencies and silence near the target word. The phone sequences in the lexicon (canonical baseforms) are hence built to be reflective of most realizations of the word.

In controlled conditions as in isolated speech, variations range from gross variations like those of the word 'tomato' in American and British English ([t ow m ey t ow] in *AmE* & [t ow m aa t ow] in *BrE*) to more subtle changes like that in the word 'all' ([aa l] & [ao l]). Canonical baseforms are expected to cover most variations, which are then complemented by the context dependent acoustic models to capture the co-articulation effects. Among other factors, the validity of this assumption weakens with increasing variation within the task (isolated speech, carefully read speech, conversational speech, sloppy speech and so on). An obvious solution to handle this is to have the lexicon cover all possible variations of the word pronunciations. This technique, however is helpful only to an extent. Recalling the Viterbi algorithm (Sec. 1.1), since the decoder finds the best phone string rather than the best word string, it biases against words with multiple pronunciations. Furthermore, [11] shows that as the vocabulary size grows, acoustic confusibility among the lexical entries increases and it becomes a non-negligible source of recognition errors. So, variants have to be added cautiously, taking into consideration the improvement due to added variants and also the offset due to the increased confusibility.

Pronunciation dictionaries are generally hand-crafted by linguists to reflect the most agreed pronunciations of each word. In theory, sound units are divided into two basic types: phones and phonemes. *Phones* are the fundamental sound categories that describe the range

of acoustic features found in languages. *Phonemes* on the other hand are abstract, language-specific entities that may represent one or more phones.

Phonology is dedicated to capturing the variations in the *surface forms* of pronunciation. It investigates which phone a particular phoneme would assume in a given *context*[1]. The various phones that are candidate realizations of a phoneme are referred to as *allophones*. For instance, [ah], [ae], [ay] are allophones of the phoneme /ah/ in different contexts. In ASR literature, several techniques have been proposed for capturing this variation into the lexicon. Section 2.2 presents an overview of related attempts. Although the section is a bird's eye view of earlier attempts, more specific comparisons will be made in the following chapters as applicable.

2.2 MODELING VARIATION: OVERVIEW OF APPROACHES

An important distinction that is often drawn in modeling pronunciation variation is that between within-word and cross-word variation. The underlying phonetic mechanisms are different in the two and hence the need to address them separately. Approaches to handle cross-word variation have widely employed the use of multi-words [17] [18] [19] [20] [21], wherein frequent word clusters are concatenated as one lexical entry. This technique can account only a small portion of cross-word variation, like the variation between words that occur in very frequent sequences. Due to this limitation, other techniques involving rewrite rules based on word context etc have also been proposed like those described in [22] [23] [24] [25].

Within word variation is the kind of variation that can be modeled at the level of the lexicon by adding pronunciation variants [26] [27] [22] [28] [24] [17] [29] [18]. On similar lines, this thesis delves into modeling within-word variations. Earlier approaches to this problem have all employed and differed within two broad phases:

1. Finding the information on variation of pronunciation
2. Integrating this information into ASR

[1] [16] elaborates the definition of context as applied to pronunciation modeling

2.2.1 Information about pronunciation variation

An important step in modeling variation is investigating the sources of information on pronunciation variation. This can be obtained by knowledge-based or data-driven techniques:

2.2.1.1 Knowledge-based approaches to study variation

In knowledge-based approaches, information on pronunciation is mainly derived from sources that are already available. Existing sources can be pronunciation dictionaries or rules on pronunciation variation from linguistic studies [26] [22] [23] [17] [18] [19]. In general, these rules are optional phonological rules concerning insertions, deletions and substitution of phones. A drawback of knowledge-based methods is that these sources usually only provide qualitative information about pronunciations (like the possible allophones a phoneme can assume). It doesn't provide any quantitative information, which is essential for system building. Another drawback is that existing knowledge is available from analysis done on laboratory speech, and may not hold for all testing conditions.

2.2.1.2 Data-driven approaches to study variation

The idea behind the use of data-driven techniques is to simulate the test conditions. Also, they carry limited or no biases to linguistic theories, relying on the techniques to automatically discover information and rules from the data. The various realizations of word pronunciations are obtained directly from the speech signals [25] [19] [16] [29] [13] [30] [31] [20] [32]. The acoustic signals are analyzed in order to observe the different ways in which the words are realized. A common stage in the analysis of the signal is the transcription of speech which is done either manually [16] [13] [32] or (semi-)automatically [26] [27] [30] [24] [29] [20] [32]. The latter is usually done using a phoneme recognizer or by means of forced Viterbi alignment. The transcriptions are either used directly for new word pronunciations or for formalizations derived from them [27] [24] [29] [18] [20]. This is done by comparing the transcriptions against canonical transcriptions (using existing baseforms). The comparison is done by a dynamic programming based alignment. The resulting

alignments are then used to:

- derive rewrite rules to characterize the variations undergone by the baseforms [27] [24] [18] [20],
- train an artificial neural network to model the phone change process from the canonical form to the transcription [29],
- train decision trees to learn the phone change as per the context in the canonical form [16] [32] or to
- calculate a phone confusion matrix [33]

2.2.2 Using the information on pronunciation

In Section 2.2.1, an overview of various sources of information used by existing approaches is given. This section presents an overview of how this information has been used for improving ASR. Adaptation of the lexicon is done by adding pronunciation variants to it. So the first stage in the process is generation of candidate variants. This is either done manually [22] [32] or by automatic procedures such as:

- using learnt rules to generate the possible candidates of variants [26] [27] [18] [19] [20],
- artificial neural networks [29],
- grapheme to phoneme converters,
- using a phoneme recognizer [19] [20] [21],
- maximum likelihood optimization [30] and
- decision trees [16] [32]

The variants that are likely to capture relevant variation are selected from those generated above. The assumption made here is that since multiple pronunciations are present, the recognizer can select from among the different baseforms to match the acoustics. This reduces the errors made due to mismatched pronunciations. However, addition of variants adds to the confusibility within the lexicon, which may lead to degradation of the recognizer performance. So, variants should be cautiously added to balance the improvement due their

addition and the offset introduced by the added confusibility. For this purpose, different criteria for variant selection are used such as:

- frequency of the variant's occurrence [18] [20] [32],
- a maximum likelihood criterion [30],
- confidence measures [21] and
- degree of confusibility between the variants [21].

2.3 LIMITATIONS OF EXISTING APPROACHES

Data-driven methods score over knowledge-based methods for the current problem, since they make use of available evidence. Among the data-driven methods cited above, the emphasis was put on selecting the right kind of variants from automatically/manually generated transcriptions. There was not much work in improving the transcription accuracy itself. So, the errors of transcription persist through the later steps in the process.

Another limitation is the choice of the candidate variants that different techniques considered. Since this choice puts a hard bias on the hypotheses evaluated, it is an important early stage decision. Two closely related works employ techniques as follows-

[30] employs a tree-trellis algorithm based on the A^*-algorithm [34] for finding optimal path through an elaborate phoneme tree. This effectively has the same exhaustive hypotheses space and is as error-prone as raw phonetic decoding.

[35] employs a rule based generation of candidates allowing each phoneme of the canonical baseform to be optional per candidate. Although, it is stated to be an adhoc setting, the rule is too simplistic and overly biased to the canonical baseform to perform any commendable inference. Hence, this is not generalizable.

At an implementation level, barring a few exceptions [36] [30] [37], a distinction earlier attempts did not acknowledge is between inference and recognition. While recognition tries to overlook variations to identify the intended hypothesis, inference aims to capture the actual phones in the acoustics. So, techniques that are adapted from general recognition

frameworks (e.g. CDHMM modeling) are incompatible for inference. This may be a possible justification for marginal improvements in previous attempts. Another disadvantage of tailoring existing recognition tools for this problem is the limited control it leaves on the process, besides the irrelevant time and complexity overhead.

2.4 APPROACH ADOPTED IN THIS THESIS

This thesis proposes improvements at different stages of the process. At the first stage, techniques are investigated to improve the accuracy of variant inference, rather than relying on the canonical baseforms or directly the inferred surface forms. In the second stage, it attempts to build better pronunciation dictionaries that improve the ASR accuracy.

As would be shown in Chapter 3, acoustic evidence is sub-optimal for generation of candidates for further selection; this is invariably done in all earlier data-driven methods. Rather than considering the exhaustive candidate space of the raw phonetic decoder [30] or overly pruning the candidates [35], this thesis chooses a balance between the two. This thesis uses the least error-prone resource, orthography as the primary information source for candidate generation and uses acoustics only as an evidence to improve the inference accuracy.

This thesis employs an exclusive framework for inference of the surface form from acoustics. The integration of knowledge from other information sources is done at the time of alignment, making the entire process, time and computation efficient. The thesis also discusses possible techniques to improve agreement among inferred variants, thereby weakening the possible offset due to lexical confusibility. Like few other approaches [35] [18], findings from error analysis are incorporated in the selection algorithm for further efficiency. ARCTIC and TIMIT are used as the test sets to validate the proposals made.

2.5 PERFORMANCE COMPARISON

Although a direct performance comparison of approaches cannot be made due to the varying assumptions, baselines, vocabulary sizes and test sets in each, this section tries to contrast the specifics of a few attempts.

[33] uses 200 OGI names as a test set for isolated name recognition task and reports a 19.4% relative WER improvement in a controlled dialog task. they use hybrid techniques combining both knowledge-based and data-driven strategies. In contrast, the orthography-driven inference techniques proposed in Ch. 3 show a 21.4% improvement in the phone error rate on the same task.

[30] employs a maximum likelihood approach for variant selection on the 900 word task Resource Management database. They note a 18.4% relative improvement over manual dictionaries, the best performance caused by allowing 1.3 alternatives for each lexical entry. Another closely related work, [35] employs a rule-based data-driven technique for selection of pronunciation variants. It shows an 8% relative improvement on a 1288 word Dutch spontaneous speech allowing 2 variants per word. In this thesis, the inference and selection are validated for continuous speech on the 2366 vocabulary single speaker ARCTIC data. A 14.6% relative WER improvement at 1.4 variants per word is shown. The same techniques have shown to give an 8% improvement using 1.6 variants per word on the multi-speaker TIMIT database.

CHAPTER 3

Capturing Variants using Lexical and Acoustic Information

3.1 INTRODUCTION

This chapter focuses on improving the quality of the inferred pronunciations from acoustic samples. Hence, an investigation of various information sources about pronunciation is necessary. As seen in the Chap. 2, earlier methods (Sec. 2.2.1) have broadly used either existing phonological knowledge in the language or have used acoustics to infer pronunciations. This chapter presents techniques to exploit the available knowledge sources further to infer better pronunciation baseforms in a data-driven fashion. As mentioned earlier, the inference techniques described here apply at the word level.

3.1.1 Sources of information about pronunciation

It is important to identify the various sources of information relevant to pronunciation before proceeding to techniques to infer it automatically. This section presents the various sources of information. It should be noted that these factors and the extent of their roles largely depend on the language in consideration.

- The knowledge of the language: the phonology, stress patterns etc.
- The spelling of the word (for languages that have a written form).
- Spoken example(s): the direct source of pronunciation.
- Context: at all levels (the effect of neighboring words; discourse context; speaker emotion etc)

Depending on the scenario, only a few of these sources are available at disposal for inference. Table 3.1 gives a few example scenarios and the available information sources from which to learn about pronunciation.

Table 3.1: Pronunciation knowledge sources available in different scenarios

Example Scenario	Available information
Human hearing	innate psycho-acoustic faculties
ASR training	spelling; acoustics; context
Dictation systems	spelling (upon corrective feedback); acoustics; context
Dialog systems	acoustics; context
TTS systems	spelling; context

Accordingly as per table 3.1, the available knowledge sources vary with the application. In ASR systems, pronunciation baseforms in dictionary are usually either 1) manually built or 2) derived from acoustics or 3) generated from the spelling (letter-to-sound rules) or 4) a combination of these. Acoustics driven methods implement a Viterbi decoding on the acoustics using sub-phone (arc) acoustic units and a phone transition model to derive one or more pronunciations for each word [38] [37]. Orthography-based methods widely use finite state transducers (FST) or decision trees to determine the pronunciation [39] [40]. However, the quality of orthography based pronunciations depends on the grapheme-phoneme correspondence of the language. Hence, they cannot be directly used as baseforms in the ASR lexicon.

Of late, data-driven techniques combining both linguistic and acoustic information have gained focus owing to the better performance and wide range of application scenarios providing such a setting. To name a few are automatic lexicon generation [36] and systems supporting dynamic vocabularies [41] [33]. [33] and [42] for example, use syntactic and semantic information to incorporate dynamic classes allowing OOV detection and enrollment. Also, [36] applies a letter-to-sound constrainer within the decoder to take advantage of the spelling of the OOV word.

However, earlier attempts have only partially exploited the availability of rich information sources. This chapter exploits the linguistic information further by efficiently constructing the *n*-best list of pronunciation alternatives and scoring them using decision trees. The hypotheses are further rescored with costs in acoustic alignment and phone transition, usually modeled using a phone *n*-gram model. Thus, this thesis uses all the information sources presented for ASR training.

3.2 CANDIDATE GENERATION AND RESCORING

This chapter uses tree based letter to sound models to characterize allophonic variations based on phonemic context. Conventional approaches use the acoustics to generate an *n*-best list of possible phone/sub-phone strings. The *n*-best alternatives are re-ranked using additional knowledge sources, like a language model, to improve the intelligibility of the best alternative. Typically the first best alternative is the output of the decoder. Consider the following schemes-

[30] uses a tree-trellis method for variant generation. It uses a maximum likelihood criterion to generate the baseforms that maximize the joint probability of being realized as the spoken examples instances of the word presented for training. It searches an exhaustive space of phones which is redundant besides being error-prone.

[35] uses rule based generation of candidates from the canonical baseform where each phoneme can be optional. The candidates are later Viterbi-aligned and frequency-based rules for variant selection are investigated. The search space here is too small (in fact equal to the length of the baseform, one candidate with each phoneme missing).

These cases are on two extremes: one considering an unnecessarily exhaustive space and the other considering a highly constrained set. In this chapter, a set of potential candidates is generated from the available information. This chapter shows that as generally employed in earlier works, acoustic evidence and the canonical baseform are not very informative for generation of candidate pronunciations. The novelty of the method proposed here lies in inverting the common relationship. This chapter uses the spelling information to generate an

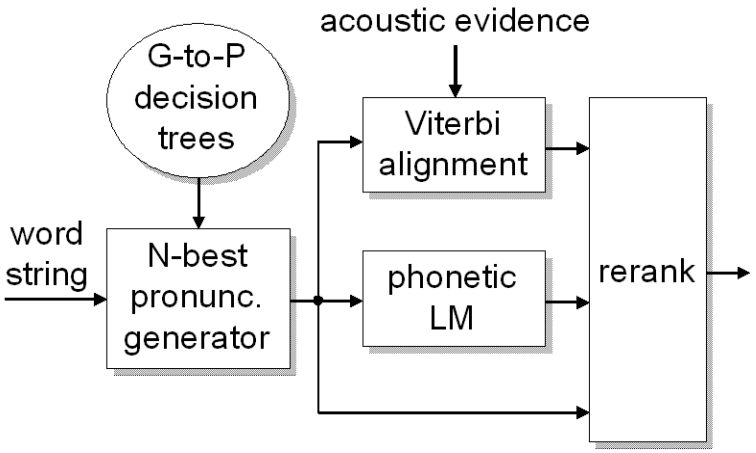

Fig. 3.1: Schematic diagram for baseform inference

n-best list of pronunciation hypotheses, which can be subsequently rescored using available acoustic evidence and phone transition costs. The *n*-best list referred here is analogous to the same in continuous speech recognition (Sec. 1.1). This thesis presents the superiority of spelling over a single spoken exemplar in surface form inference. Fig. 3.1 describes the process of baseform inference employed in this chapter.

The bias towards using orthography for generating the *n*-best list is justified by the fact that, on an average, spelling can give more information about the pronunciation than a single acoustic exemplar, as borne out by the results in (Sec. 3.4.1, below). The following subsections present the three information sources: The spelling, the spoken evidence and known phone transition patterns in the language. Sec. 3.2.1 describes the decision-tree based approach for the generation of *n*-best pronunciation hypotheses followed by section 3.2.5 for subsequent rescoring.

3.2.1 Learning grapheme-to-phoneme rules

Both rule-based techniques (FSTs, mapping tables etc) and statistical methods (decision trees, discrete HMMs, SVMs etc) are widely used for capturing the grapheme-to-phoneme (or letter-to-sound) rules. In this work, decision trees are used. Decision trees offer flexibility in choice of the modeling context and hence are chosen as the statistical paradigm

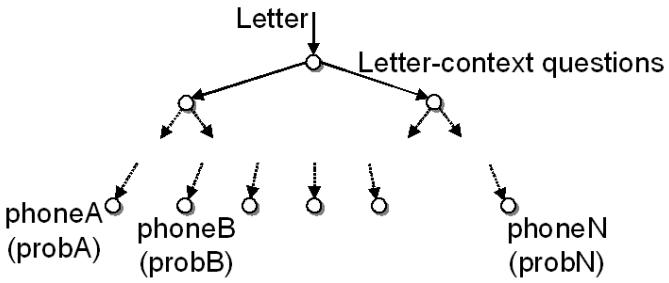

Fig. 3.2: Letter to sound capture by decision trees

for capturing letter-to-phone rules from a large training lexicon. Separate decision trees are trained for each letter of the alphabet. The leaves of the tree are discrete probability distributions of the phones and the internal nodes are questions about the neighboring context (e.g., next letter='a'? etc.). Fig. 3.2 shows the decision tree format for capturing letter to sound rules as used in this chapter.

For the experiments reported here, training and testing set features for each letter are extracted from CMUDICT [43] of 130k words. The trees are built using the CART based letter-to-sound module within the FESTVOX [40] framework. Various context lengths of 1, 2, 3 and 4 letters on either side of the target letter are tried. The performance of the resulting trees in predicting the phone produced by a letter in an untrained word is studied. 1-letter context and 4-letter context trees are discarded for being overly general and over-training the decision trees. Fig. 3.3 shows the relative performance of the 2-letter and 3-letter context trees on a held-out set, consisting of 10% of the lexicon. As would be expected, 3-letter context trees outperform 2-letter context trees. Also, it is interesting to note that irrespective of the context length, relative performance within the letters remains the same in both cases. Furthermore, letters that produce vowel sounds (a, e, i, o & u) perform significantly worse than the other consonant letters, which also agrees with intuition.

3.2.2 Orthography based n-best pronunciation generation

Given the grapheme-to-phoneme decision trees, multiple (*n*-best) hypotheses of pronunciations for a word are generated as follows: From the spelling of the given word, features are

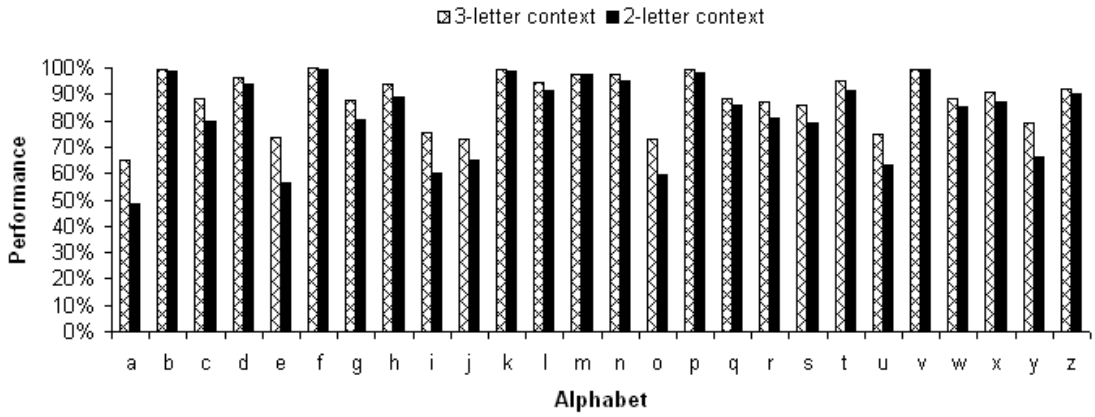

Fig. 3.3: Performance of 2-letter and 3-letter context trees on the held-out data

drawn for each letter using the same context length (2 or 3 letters) as that of a chosen set of trees. When queried with these features, the corresponding G-P trees return a list of phones, with their probabilities, for each letter in the word. A variant of best first search algorithm can be used to traverse through all of the phones predicted for each letter, thus generating several pronunciation alternatives. Each pronunciation also receives a score which is the product of probabilities of the constituent phones, as given by the decision trees. The product is referred to as the *n*-best likelihood. This is a model based on orthography used in the n-best list rescoring process described further below.

3.2.3 Learning from Acoustic Evidence

While spelling captures the gross aspects of pronunciation, the acoustic instance is the direct evidence of the pronunciation and hence an important source of information. In this work, acoustic alignment is used as an important factor for determining the surface form. Each hypothesis in the *n*-best list of pronunciations is aligned (using Viterbi alignment) against the single speech sample of the word, producing an acoustic likelihood score for the hypothesis. The acoustic likelihood is used in re-ranking the *n*-best list, as will be described in section 3.4.1. For the alignment in this chapter, acoustic models consisting of three state context independent phone models with left-to-right topology, and 8 Gaussian mixture com-

ponents per state are used. The models are trained on the officially designated training set of the TIMIT data [44]. The *sphinx3_align* tool from the Sphinx suite [45] is used for Viterbi alignment.

3.2.4 Phone transition model

Phone transition patterns in the language give important information as to which phone sequences are acceptable in the language. The function of the phone transition model is similar to that of a language model (Sec. 1.0.2) in continuous speech recognition. It provides a prior probability to each hypothesis in the *n*-best list. In this chapter, a phone bigram model trained on CMUDICT is used. Word beginning and ending markers are also considered while computing the probabilities. The model was then smoothed with a uniform distribution, to avoid over-fitting to the training data. The smoothing was done as follows: If N is the number of phones, and $P(p_2|p_1)$ the unsmoothed probability of transitioning from phone p_1 to phone p_2, the smoothed transition probability is given by:

$$P_{Interpolated}(p_2|p_1) = \omega * P(p_2|p_1) + (1-\omega)/N \tag{3.1}$$

The scaling factor $0 < \omega < 1$ can be chosen according to the reliability and comprehensiveness of the dictionary. The cleaner and larger the dictionary, the higher ω can be. An optimal value for ω can be determined empirically using the deleted interpolation technique. $\omega = 0.5$ in the experiments reported here.

3.2.5 N-best rescoring criteria

The n-best list of pronunciations generated according to sec. 3.2.2 is rescored by combining the three scores: *n*-best likelihood (sec. 3.2.2), acoustic likelihood 3.2.3, and phone transition costs (sec. 3.2.4). Since the three have widely differing ranges, a following combination function is proposed. For each alternative, ϕ in the n-best list, the function $\xi(\phi)$ is computed where:

$$\xi(\phi) = (Acoustic\,likelihood) * (nbest\,likelihood)^{\eta} * (Phone\,transition\,penalty)^{\gamma} \quad (3.2)$$

The exponentiation weights η and γ are determined empirically (similar to 'language weight' in most speech recognition systems). The highest ranking pronunciation, according to ξ is chosen as pronunciation for the word.

3.3 EVALUATION

In the experiments reported here, Phone Error Rates (PER) of the inferred baseforms are used as the performance measure. The baseline for our comparison is the PER of the top hypothesis in the original *n*-best list (before rescoring). For the test data, we chose them to be exclusively proper names, which are a good representative of OOV words in many applications. Furthermore, the peculiarities of the spoken form of proper names as opposed to their written form, makes them an appropriate tough test for the current problem. 173 randomly selected first and last names from the OGI names corpus [46] are used, these are the subset that is publicly available for use (Appendix A). This test set was excluded from the training data for acoustic, G-P trees, and phone transition probability models. 3-letter trees are chosen as the decision trees for the *n*-best list generation step.

3.4 EXPERIMENTAL RESULTS

In Table 3.2 the baseline PERs of the top hypothesis of the original *n*-best list, re-ranked by each of the three scores individually (i.e., not in combination with any of the others). The table shows the average error rates obtained on the test data. The table suggests that orthography determines the pronunciation more reliably than a single instance of the speech. This may change when more than just a single instance is provided. (Furthermore, relying solely on phone transition probability to rank the *n*-best list is clearly useless, and is only included here for the sake of completeness).

Table 3.2: Baseline phone error rates of the factors contributing to rescoring criterion.

Baseline	PER (%)
Orthography based n-best	22.9
Acoustic alignment	37.8
Phone transition	68.6

The orthography-based performance of 22.9% PER is the baseline for comparison in the following sections, which deal with combining the three sources of information effectively in re-ranking the *n*-best list of pronunciations.

3.4.1 Use of Spelling and Acoustic Evidence

The effectiveness of combining acoustic likelihood with *n*-best likelihood in *n*-best selection, ignoring phone transition costs is examined here. To study this combination, a wide range of values for η are tried, measuring the PER from the best re-ranked *n*-best hypothesis in each case. Fig. 3.4 shows the performance with varying η The dotted line represents the baseline performance of 22.9% PER using *n*-best likelihood alone. It can be observed that as η increases the PER drops rapidly from the acoustic-likelihood baseline of 37.8% (η=0), and reaches a minimum of approximately 19.5%. The combined information from orthography and acoustics is able to provide a 3.4% absolute improvement (14.8% relative improvement) over the *n*-best likelihood baseline performance of 22.9% PER.

3.4.2 Use of Spelling, Acoustic Evidence and Phone Transition penalty

The performance can be further improved by bringing in phonotactic constraints via the phone transition penalty. To study the effect of this factor, the *n*-best likelihood weight η is kept constant around the middle of the steady-state region in Fig. 3.4 (η=28 is chosen). The phone transition penalty weight γ is varied in computing $\xi(\phi)$ and the error rates from the re-ranked *n*-best list are recorded. Fig. 3.5 summarizes the behavior. As shown, a further

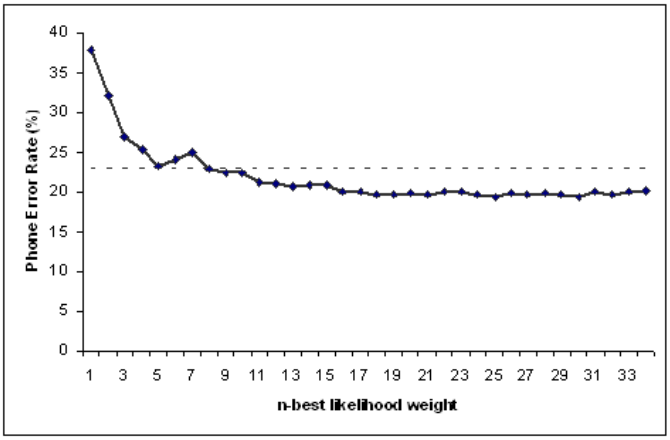

Fig. 3.4: Effect of increasing the Orthographic exponentiation weight η on performance

reduction in PER can achieved, reaching a minimum of around 18%, which is a 21.4% relative improvement over the orthography baseline of 22.9% PER.

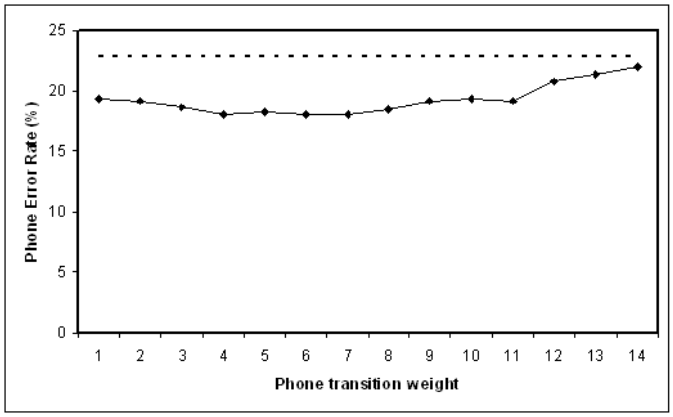

Fig. 3.5: Effect of increasing the Phone transition exponentiation weight γ on performance

3.5 SUMMARY

This chapter introduced a technique for pronunciation of words, employing an orthography-driven *n*-best list generation, and rescoring using acoustic and other evidence. The orthographic information is shown to be more accurate than a single spoken exemplar. Accord-

ingly the *n*-best list generation is based on the the richest information available. All other information is used to re-rank the list. A comprehensive evaluation and analysis of the approach shows that the *n*-best list likelihoods and phone transition priors can be used to reduce phone error rates of the inferred pronunciation surface forms significantly.

On the test set employed here, the PER is reduced from the orthographic baseline of 22.9% to about 18%, a 21.4% relative reduction. Obviously, the true error rate is highly task and application dependent. Chapter 4 validates the extent to which this improvement in PER translates into WER improvements on continuous speech.

CHAPTER 4

Improving Pronunciation Dictionaries for Continuous Speech Recognition

4.1 INTRODUCTION

In Chapter 3, a framework is proposed for baseform inference exploiting orthographic and acoustic information. The performance was measured in terms of Phone Error Rates (PER) on spoken proper names, an isolated speech task. Often times, a PER decrease may not be translated into Word Error Rate (WER) improvement. The focus in this chapter is on application of the inference techniques for improving WERs on continuous speech. Single speaker ARCTIC database and multi-speaker TIMIT database are used as the two continuous speech tasks. Also, Chap. 3 uses a conventional step by step procedure:*n*-best list generation, Viterbi alignment followed by rescoring the list using various scores. A possible set of *n*-best variants were generated from the spelling using CART performing grapheme-to-phoneme conversion. Information sources such as the acoustic evidence and phone language model were used to rescore the *n*-best list to give the highest score to the true variant of pronunciation. The scores from various information sources were effectively combined using a combination function.

In this chapter, pronunciation inference is applied to continuous speech on TIMIT [44] corpus. The approach used in the earlier chapter (Fig.3.1) involved Viterbi alignment of each hypothesis from *n*-best pronunciation generator and has a high time complexity involved in generating the variants. In this chapter, the Viterbi and combination of information sources are integrated into a single phase as shown in Fig.4.1. A method of selection of variants based on frequency and a precedence based distance from its baseform is also incorporated.

The precedence-based distance measure proposed constraints on the addition of variants into the ASR lexicon. The new scheme is highly time effective for continuous speech tasks which otherwise need an order times more time employing the scheme in Fig. 3.1.

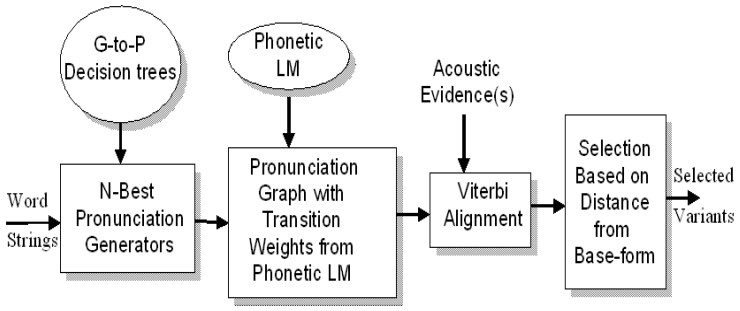

Fig. 4.1: Time efficient approach for baseform inference

Significant WER reductions have been shown in the recognizer's performance by augmenting the inferred variants to lexicon. Error analysis is also done on the tasks to understand the phonological processes in continuous speech.

4.2 VARIATION IN CONTINUOUS SPEECH

Variation in pronunciation in continuous speech occurs due to a number of factors including speaking style, gender, dialect, etc. Existing attempts to account for this variation can be categorized broadly into two classes: Modeling at the acoustic level (front-end signal processing and acoustic model parameters) and at the lexical level (Chap. 2). The current work falls into the latter category.

Traditional lexicons used in speech recognition are first order Markov chains of the acoustic modeling units. The lexicon (or the pronunciation dictionary) is a representation of the recognizer's vocabulary in terms of its acoustic modeling units. However, several sources of variation in continuous speech affect the pronunciation of a word. These include (1) variations that depend on word level features of lexical items (e.g. part-of-speech, tense etc.), (2) variations particular to certain lexical entries, (3) variations that depend on the stress and syllable position of the phonemes and (4) variations that depend on the local

phonemic or phonetic context [47]. To account for these variations, several approaches have been proposed such as including multiple pronunciation variants [26], multi-words [17] and other hybrid techniques. More complex lexicons using articulatory feature information [48] and those using tree based dictionaries [47] [39] [16] have also been attempted yielding promising results. A thorough survey of literature in this context can be found in [31].

4.3 RELATED WORK ON PRONUNCIATION GRAMMAR NETWORKS

Ideally, though raw phonetic decoding is expected to reveal the underlying surface form, the poor performance of phonetic decoding makes it necessary to constrain the candidate surface-form hypotheses. Most approaches to add pronunciation variants to the lexicon (including techniques presented in Chap. 3) follow the conventional ASR pipeline. They comprise three phases- 1) Variant candidate selection/generation, 2) Forced Viterbi alignment, 3) Rescoring using other knowledge sources and 4)Best candidate(s)/rule selection. In this chapter, the candidate selection, Viterbi alignment and rescoring phases are integrated using a grammar-based decoder framework with token passing. Another advantage of using such an integrated framework is the time taken to infer the pronunciation variant used in the acoustics.

Most variant generation/ phonetic graph construction approaches are implemented as phonetic decoding constrained by a weighted finite state transducer [33] or by a rule-based generation criterion of alternatives given the baseform [35]. This chapter employs method similar to the weighted pronunciation network in [39] constructed via trained letter-to sound decision trees (Sec. 3.2.1). The advantage of using decision trees for phone graph construction for a continuous speech task is multi-fold. Firstly, each letter outputs a discrete probability density function (PDF) on the phones based on the modeling context (number of letters on either side of the letter in question); if a huge dictionary is used for training the trees, the probabilities tend to the reliable likelihoods of the phone distribution; another useful property of the trees is that each PDF may contain an optional ϵ (null) production. This can account for phone deletions, the kind of variation in continuous speech

that context-dependent phone modeling fails to capture [15]. Following from Chapter 3 the phone likelihoods at the leaf nodes capture an important relation between the spelling and the pronunciation. It has been established that orthographic information is more reliable in determining the underlying variant than the spoken exemplar itself.

While the choice is similar to [39], one of the earliest works employing decision trees for pronunciation modeling, the trees in [39] have been used as the pronunciations (entries in the lexicon) themselves. They were used within the framework of context dependent phone HMM (CDHMM) decoding. Since CDHMM modeling is known to account for a good deal of pronunciation variation itself [15], it is not suited to infer the actual surface form used in the acoustics. Hence, this chapter uses context independent acoustic models to decode through the phonetic graph.

4.3.1 Multiple Acoustic Examples

Since the current chapter deals with continuous speech databases, multiple instances of each word are evidenced in the training data. The best set of variants for words seen in the training task can be generated in the two following ways:

- Generating the best pronunciation variant(s) closely matching all the spoken instances of the word considering them all at once.
- Inferring each instance as a stand-alone variant followed by application of selection criteria to identify only genuine variations in pronunciation.

While the first method may be faster, it is at a risk of ignoring the nuances within each instance that may be important. This thesis employs the latter method, inference from each instance followed by spurious variant rejection. Each instance is individually processed as in the Chap. 3, thus justifying the use of spelling as an information source.

The Figure 4.2 shows a sample pronunciation network used to model the different realizations of the word ‘above’. The solid lines correspond to the transition from phone alternatives of one letter to those of the next. The dotted lines indicate the transitions via the ϵ production (the deletion process). The probability of each phone alternative associated with

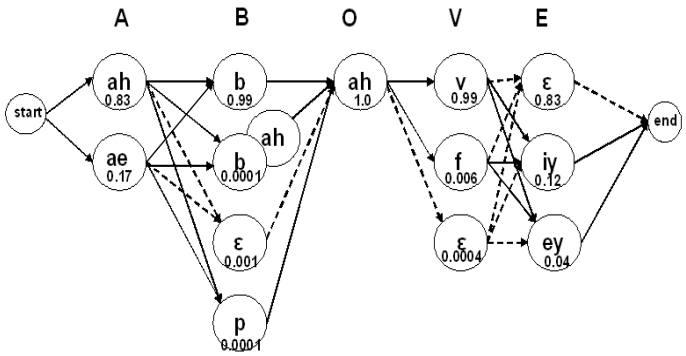

Fig. 4.2: Pronunciation network for the word ‘above’. Note the ϵ productions for the letters.

a grapheme is given under its phone label. This probability, output by the decision trees, is passed on as an additional token (along with the α score) during the forward computation along the trellis. As mentioned earlier, advantage of implementing such a pronunciation network is that a number of variants can be evaluated in a single pass of the decoder. Figure 4.2, for instance, represents 72 different variants of pronunciation for the word ‘above’.

Such representations of the pronunciation network are different from allophone networks in that the latter are built from existing phonological knowledge in the language, which is only qualitative. Also, allophone networks do not necessarily provide for ϵ productions, consequently rendering them incapable to handle phoneme deletion. The networks as in Figure 4.2 are built from phone predictions of alphabet decision trees trained on a huge pronunciation dictionary. Hence they automatically capture, and to a good extent, quantify the phonology of the language besides potentially providing for more variation. Also, they can be easily altered to any prior knowledge about the task. For example, if it is known that the speech is sloppy and all kinds of phone deletions are common, an ϵ can be inserted among all the alternatives and smoothing can be done to adjust the probabilities.

4.4 DECODING ALONG THE PRONUNCIATION NETWORK

Sequential connections are made from the phone alternatives of one grapheme to the next as illustrated in the figure 4.2. Wherever an ϵ is encountered, the connections are made to the productions of the next grapheme and so forth, iteratively. At initialization of the inference process, phone alternatives of the first grapheme are initialized with the respective α scores. These scores are propagated along all the valid connections from the current node. This is recursively done for all frames of the data. Where E is the emission probability, L is the probability assigned by the LTS decision tree, and P is the penalty given by a phone bigram LM to that particular transition, α_s^t, the α score of a state s at the time frame t is computed as

$$\alpha_s^t = E \times L^\eta \times P^\gamma \tag{4.1}$$

The exponentiation weights η and γ are scaling factors responsible for balancing the dynamic ranges of the three factors. This is the same as the rescoring criterion Eqn. 3.2 of the previous chapter. While in Chapter 3, the combination is applied at the utterance level, for experiments in this chapter it is applied at the frame level along with the forward computation. The state scoring the maximum α_s^t at the time instant $t = T$, the final frame, is selected from among the candidate ending states as suggested by the pronunciation network. A traceback from $t = T$ to $t = 0$ is performed along the network to infer the hidden state sequence. This hidden sequence is the surface form of the pronunciation in that word instance.

4.5 DATA AND EXPERIMENTAL SETUP

This chapter has two broad phases- Baseform inference and evaluation. The resources used in each phase are described below.

4.5.0.1 Grapheme-to-Phoneme trees

The Grapheme-to-Phoneme decision trees used for the phone graph generation are built from the 130K word vocabulary CMUdict [43] using the CART based letter-to-sound (LTS) module within the FESTVOX [40] framework. A phoneset of 39 phones is used in all experiments reported in this chapter. A 3-letter context on either side of a grapheme is used and the trees are built accordingly. The trees are the same as those in chapter 3. While in Sec. 3.2.1 the trees are used for *n*-best list generation, in this chapter they are used for phonetic graph construction.

4.5.0.2 Speech data and preprocessing

In this work, the evaluation is carried out on two tasks- single speaker ARCTIC and multi-speaker TIMIT speech databases. The two databases are phonetically segmented using the EHMM labeler [49] in FESTVOX. In the variant inference phase, 5 state context independent HMMs with 2 Gaussian components per state are used as the acoustic models, and variance normalized MFCCs are used as features. Using the labels and the initial dictionary used for training, features for words are extracted from each utterance.

Since cross-word effects cause additional variations that alter pronunciation in continuous speech, word boundaries may be confounding for any pronunciation inference technique. This is more severe if CI models are employed. To avert this, the inference is done at word level in this chapter. Table 4.1 shows the number of training words and the instances dealt within each training set.

Table 4.1: Number of tokens and unique words in the training sets

Task	Tokens	Unique words
ARCTIC	7997	2366
TIMIT	39687	4896

4.5.1 Evaluation

As will be discussed in section 4.6.2, all inferred variants are not suitable for addition to the lexicon. So, criteria are proposed for selection of promising candidates. Evaluation of the inferred variants is done by computing WER of a continuous speech recognizer with and without the data-derived baseforms. The recognizer's acoustic models are 5-state context dependent continuous HMMs with 8 Gaussians per state. Since this work aims to study the effect on performance only due to changes in the lexicon, all other sources of errors [10] are minimized to the extent possible. To this end, the test transcript is also included into the language model training, and the WER is computed at the best set of language weight and word insertion penalty parameters.

4.6 EXPERIMENTS AND RESULTS

4.6.1 Baselines

- The acoustic models are built using the officially designated training set (4620 utterances) of the TIMIT corpus. The performance is reported on the test set (1680 utterances) for the TIMIT baseline. The language model for this test set is built on the entire transcription (training and test) transcript. As mentioned earlier, this is a deliberate setting employed to isolate the changes in performance only due to the baseforms in the pronunciation dictionary.
- For the single speaker (American English) ARCTIC data, 80% (904 utterances) of the data is used as the training set for baseform inference. Testing is done on the remaining 20% of the data using TIMIT CD acoustic models. The language model in this case is built from the entire ARCTIC transcription. CMUdict is used as the lexicon in both cases, an LTS suggested pronunciation is used for the OOV words.

Table 4.2 presents the baseline performance of the recognizer on ARCTIC and TIMIT test sets.

Table 4.2: Baseline WERs on TIMIT and ARCTIC test sets

Task	Vocabulary	Performance (WER %)
ARCTIC	2770	9.006
TIMIT	6122	6.057

Interestingly, though ARCTIC seems to be a smaller (vocabulary wise) and simpler (single speaker) task than TIMIT, the WER is significantly higher. The justification to this seemingly aberrant behavior lies in the perplexity values of the test data. Table 4.3 shows that the perplexity of the ARCTIC test set with respect to a model built on the whole ARCTIC data is much more than that of the TIMIT test set.

Table 4.3: Perplexities of the testing transcripts

Modeled data	Testing Data	Perplexity
TIMIT (test+train)	TIMIT test	5.81
ARCTIC(test+train)	ARCTIC test	24.94

4.6.2 Baseform Inference and selection

The inference technique described in section 4.4 is applied on all word instances of the training set utterances of ARCTIC and TIMIT. Table 4.4 shows the number of words and the number of unique variants identified by the inference algorithm[1].

Table 4.4: Number of unique surface forms

Speech Corpus	#words	#unique surface forms
ARCTIC	2366	3584
TIMIT	4896	12784

It is noteworthy from the numbers in Table 4.4 that a fair amount of agreement (1.5 variants/word) exists among the inferred variants in ARCTIC data while a relatively much

[1]The parameters $\eta = 0$ and $\gamma = 1$ are used, effect of higher values are discussed in later sections

lower agreement (2.6 variants/word) is seen in the TIMIT data. This is justified by the fact that ARCTIC is a single speaker database and the variations in pronunciation are consistent with the training set. TIMIT data, on the other hand, is a collection of speech from 630 speakers (10 utterances per speaker) of 8 different dialects. So, it has more variation and little representative data to train on.

The inferred surface forms are variants of pronunciation in specific instances of words. They cannot substitute the canonical baseforms present in the dictionary, but only be added as alternative baseforms [50]. They may not all be genuine pronunciation variants and cannot directly qualify to be enrolled as baseforms. Some of them could have been 'accidentally' inferred due to an unknown algorithmic artifact or due to a disfluency or unintended mispronunciation by the speaker himself. Since, it is a known fact that increasing the number of baseforms may hurt the system performance due to added confusibility among the lexical entries, it is essential to filter out unwanted (unlikely) variants among the inferred surface forms.

The auto-generated dictionaries are randomly sampled and the kind of changes undergone to the baseform are noted. Appendix B presents the observation from this study. As mentioned earlier, all the changes noted in the Appendix B are not genuine. So, techniques are needed to select the best variants. Briefly, selection is done using the following criteria-

1. Frequency of a surface form.
2. Levenshtein distance of a surface form from the canonical baseform.

As a first stage elimination, only the most frequent variants are considered as candidates for further evaluation, since absolute frequency of a variant is the most reliable rule for its acceptability [35]. Among the frequent variants, those that are phonetically similar to the canonical baseform are selected for addition into the lexicon. The similarity is computed using the Levenshtein distance measure, a dynamic programming based string matching algorithm. The Levenshtein distance[2] is a measure of the minimum number of operations needed to transform one string into another. In the current context, the operations involve the

[2]Levenshtein edit distance comes from Information Theory and is widely applied in Computer Science

insertion, deletion or substitution of a single phone (e.g., The Levenshtein distance between phone strings [t ow m aa t ow] and [t ow m ey t o] is 1).

Figure 4.3 illustrates the decrease in the error rate with pronunciation alternatives (variants) added per increasing Levenshtein distance. The dotted line is the baseline WER (9.006%) on the ARCTIC test set. The rate of decrease in WER gradually comes down with increasing number of variants before reaching an optimum (7.88% due to 2100 additional variants). This decrease in improvement rate is because of the added confusibility and due to the fact that with increasing Levenshtein distance, the new variants selected are farther from the canonical baseform.

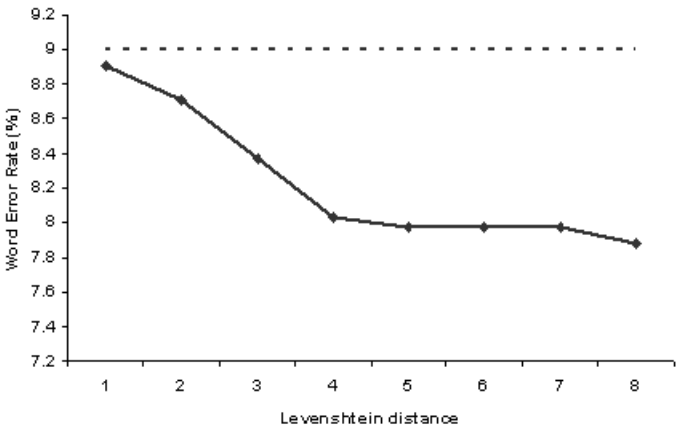

Fig. 4.3: Recognizer performance on ARCTIC test data with variants added as per increasing Levenshtein distance

4.6.2.1 Analysis of the improvement

The effect of additional baseforms on the ASR output is as follows:

- Of all the changes (improvements + degradations) caused to the new decoder hypotheses, 75.4% improved the hypotheses and the rest 24.6% detrimental with respect to the reference transcription.

- 82.5% of the improvements were caused due to the use of an inferred variant (known from the Viterbi traceback). This may be compared to [35], where only about 33% of the improvements are due their data-derived rules.
- 46.2% of the degradations caused did not use a variant during decoding. This means that a sizable portion of the newly introduced errors are due to confusibility than the variants themselves.

To analyze the specific kind of variations that are causing the 1.12% absolute (12.5% relative) improvement, the Viterbi traceback path of the improved hypotheses are examined. The variant causing the improvement in each case is identified. This is done for each improved hypothesis of the test set. The results of this analysis are as summarized in Table 4.5,

Table 4.5: Phone changes involved in the improvement-causing variants

Phone change	Relative incidence (%)
Vowel Substitution	70.5
Consonant Deletion	17.6
Consonant Substitution	5.8
Vowel Deletion	5.8

Accordingly, the precedence order of phone change for genuine pronunciation variation is :

$$VD >=< CS < CD < VS \tag{4.2}$$

Vowel substitutions are the most likely genuine variations followed in order by consonant deletions, consonant substitutions and vowel deletions (VD and CS are found to be equally likely). Where CS, VD, CD, VS stand for Consonant Substitution, Vowel Deletion, Consonant Deletion and Vowel Substitution, respectively. This conforms with intuition about pronunciation variation except for vowel deletions [3]. It is to be mentioned here that since a 39-phoneme English phoneset is used in the experiments, reduced vowels (ax, ix

[3]vowel deletion may be phonologically likely in conversational speech but not in carefully read speech.

etc) are not among the phones. The lack of baseforms with vowel reductions has forced the inference technique to accept an unlikely phone change, Vowel Deletion. This is arrived at by examining the word instances undergoing by the VD phone change. Table 4.6 shows examples of VD affected words. The variant is also compared against reduced form pronunciation provided in TIMIT dictionary:

Table 4.6: Example instances of words undergoing vowel deletion.

Word	Inferred variant	Reduced baseform
PROBABLY	p r aa b b l iy	p r aa b ax b l iy
ARRIVAL	r ay v ah l	ax r ay v el

This ability of the framework to automatically discover behavior patterns of the data is useful to compare phonological processes across different tasks, although this characterization is not attempted in this chapter. Since VD is merely an implementation artifact of this setup, the equivalence of VD and CS in the precedence order is discounted in the experiments below.

4.6.2.2 Modified Levenshtein distance metric

Using the incidence order suggested in the previous section, a modified distance metric is developed to penalize each phone change in order of its precedence. This is in contrast with the standard Levenshtein distance metric which equally penalizes all phone changes, regardless of the kind of change. The modified metric is used to calculate the distance of a variant from the canonical baseform, and lexicons with the selected variants are again tested on the test set.

Although not as a significant WER reduction, the use of modified metric did prove to be judicious. Since the new distance metric accepts only reasonable phone changes, the number of selected variants is reduced, at no cost of performance loss. This is important considering that the size of the lexicon directly increases the confusibility and also the decoding time of the system. For the ARCTIC test set, while Levenshtein distance required about 2100

additional variants to bring about the 12.5% relative improvement shown in the Fig. 4.3, the use of modified distance metric has given the same improvement selecting 1900 variants.

The plot in Fig. 4.4 shows the number of selected variants with increasing distance from the canonical baseform. The broken line plot is that of the Levenshtein distance and the solid line represents modified Levenshtein distance modified according to the precedence order.

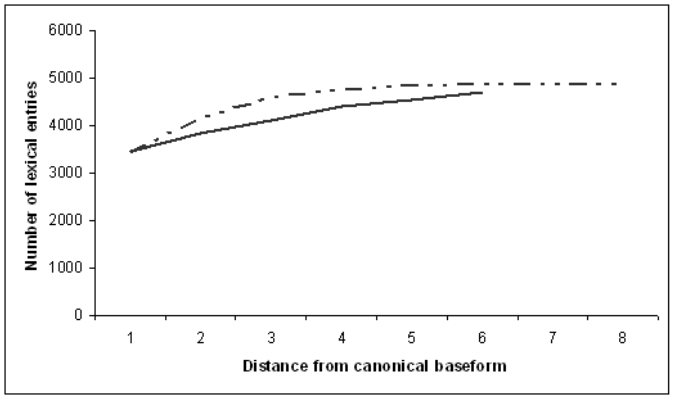

Fig. 4.4: The number of selected variants with increasing distance from canonical baseform.

4.6.2.3 Tuning the Inference decoder

This section presents tuning techniques to improve the variants inferred and reducing their number. For all experiments presented so far, inference has been done with $\eta = 0$ and $\gamma = 1$ in Equation 4.1, making it equivalent to inference in a standard finite state transducer framework. In this section, the effect of η (orthographic exponentiation weight in Eqn. 4.1) on the inferred variants and on the recognizer performance is shown. Conceptually, increasing the parameter η corresponds to averaging out certain acoustic detail and relying more on the spelling (the decision tree probability) for decoding the hidden surface form. Each value of η gives a unique set of variants inferred from the training data. Adding these variants to the baseline dictionary (after frequency and distance based selection), the best possible WER on the test data is empirically obtained. Figure 4.5 shows the best performance of different variant sets inferred using increasing values of η. The dotted line is the best performance of

variants generated with no explicit information from the spelling, $\eta = 0$ in Eqn. 4.1, 7.88% on the ARCTIC test set.

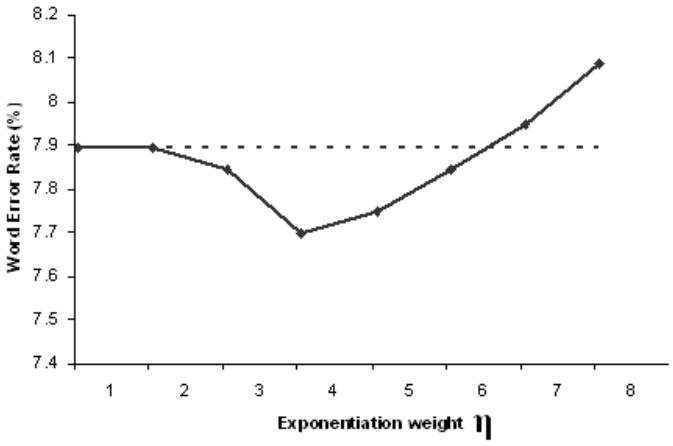

Fig. 4.5: Recognizer performance on ARCTIC test data with increasing η

It can be seen from figure 4.5 that spelling does help improving the inference of the right kind of variants. The least error rate (7.68% WER) is obtained by augmenting the variants inferred with $\eta = 3$ in equation 4.1. As η increases further, there is an undue bias of score towards the decision tree probabilities, downplaying the acoustic scores to the extent of being neglected. This makes the variants biased to the information from the spelling and not reflective of the true variation in pronunciation. This explains the increasing WERs beyond $\eta = 4$ in figure 4.5.

Another advantage due to the averaging out of certain acoustic detail mentioned above is the increasing agreement among inferred variants, amounting to a decrease in the number of variants to be added to the lexicon. In section 4.6.2.1, a relative improvement of 12.5% is shown by the addition of 2100 variants (at $\eta = 0$). While at $\eta = 3$, only 1200 variants are added to get a 14% relative improvement.

In summary, Table 4.7 compares the performances on the ARCTIC and TIMIT systems in each of the schemes described above (Sections 4.6.1, 4.6.2, 4.6.2.2 and 4.6.2.3). As can be seen, relatively, there is only a modest improvement on the TIMIT test set. This may

be attributed to the wide variation in the TIMIT database and insufficient training data to exhibit all of it. Table 4.7 summarizes the improved performances on TIMIT and ARCTIC.

Table 4.7: Final WERs on TIMIT and ARCTIC test sets

Technique of Variant addition	ARCTIC		TIMIT	
	Vocabulary	WER	Vocabulary	WER
Baseline	2770	9.006	6122	6.05
Frequency+ Levenshtein	4870	7.88	9005	5.84
Frequency+ Modified Levensh.	4670	7.88	8928	5.85
Spelling+Frequency+ Modified Levensh.	2970	7.68	7831	5.63

4.7 SUMMARY

This chapter has introduced a grammar based decoder framework for inferring surface forms of words seen in the training data. The framework is realized by construction of a pronunciation grammar network along which to search for the pronunciation variant of the word used in the acoustics. The network itself is constructed via a statistical model (here decision trees) built to predict likely candidate pronunciations of the word. The scores from other information sources are integrated into the network as tokens to be passed on during the forward computation of the Viterbi alignment. The framework provides for adjusting the reliability of each source of information about the pronunciation- the acoustic evidence and the orthography. These are effectively combined to score each candidate in the search space. The hidden variant of the pronunciation is inferred from the network via a Viterbi traceback. The proposed framework has an advantage of computation over the standard procedure employed for this problem. The framework is capable to discover processes in

continuous speech, like vowel deletion. The chapter also proposes distance based selection criteria to identify the genuine variants in pronunciation for augmenting to the ASR lexicon.

CHAPTER 5

Summary and Conclusions

To improve the performance of speech recognition systems, pronunciation modeling is one of the important issues to be addressed. Typically it involves inferring pronunciation variants and careful selection of variants into the speech recognition system. This thesis has explored data-driven methods for improving pronunciation modeling in ASR.

A systematic study has been performed on different information sources such as orthography, acoustics, phone language model and their usefulness for inferring pronunciation variants. A method to combine these information sources has been proposed to generate appropriate pronunciation variants and has been validated on the task of isolated word recognition. It was observed that the orthography carried significant information about the pronunciation than the other sources.

In order to further evaluate the combination of different information sources for pronunciation inference on large vocabulary ASR tasks, the proposed scheme was optimized with respect to the time complexity with out any loss of accuracy. The modified method involves generation of a pronunciation network for each word using orthographic information. The various sources of information about pronunciation are effectively combined within the standard Viterbi decoding framework to infer the variants from the pronunciation network.

To improve the performance of an ASR system, it is not only sufficient to infer variants but it is also important to perform careful selection of variants. Different criteria for selecting the right kind of variants have been studied for continuous speech. The performance of decoders with different variant sets are compared to understand which knowledge source is more informative about the word pronunciation. Detailed performance analysis of the best recognizer hypotheses has been done to understand the phonological processes in continuous speech. The conclusions from this study are corroborated by improved performance of

variants thus obtained. The propositions made in this thesis are validated by significant WER reductions on continuous speech tasks.

The following are the important conclusions of this work:

- Lexicons generally used for speech recognition are suboptimal, and can be improved with just the training data provided for acoustic model training.
- Our experiments have shown that orthography is significant information source for inferring the pronunciation variants than the other information sources such as acoustics and phone language model.
- To improve the performance of large vocabulary continuous speech recognition systems it is not only sufficient to infer pronunciation variants, but it also requires a careful selection of variants. In the selection process the baseforms can be used.
- The usually employed Levenshtein distance metric is not generic for computing phonetic distance between baseforms; knowledge of the pertinent task could be exploited to device a more appropriate distance measure.

5.1 DIRECTIONS FOR FUTURE WORK

- In the current work, the initial bootstrapping dictionary has to be large for reliable grapheme-to-phoneme likelihood capture by the decision trees. Future work can focus on circumventing this problem by devising alternative approaches to efficient pronunciation network generation.
- The techniques in this thesis *learn* the variations only when they are frequent and expressibly large (as a phone change). There is no other means of prioritizing the variations. Future work may focus on establishing the hierarchy of phonological events that occur within pronunciation that describe pronunciation variation at the phoneme level.
- It would be also useful to see how the pronunciation variants could be applicable in the case of conversational speech synthesis.

APPENDIX A

OGI Names

The subset of the OGI names corpus used for the inference in Chapter 3 is presented below-

abby	alicia	annette
arthur	barron	becky
beverly	bill	billy
bogg	bradshaw	bryant
bunches	campbell	canzee
carl	carol	carolyn
carter	catherine	cecilia
chris	cindy	collins
cortijos	cramer	cruise
curtis	cyndie	dana
daniel	danny	dan
darmal	darr	davis
diane	dowling	ekblad
eletto	elizabeth	emily
federson	ferrell	fujimura
gail	garito	garner
garry	gene	glenn
grace	graham	halkowiez

hall
hansen
harman
harper
harris
hartfield
harvey
hasler
hattie
heath
hill
hollins
holton
homes
horton
hughes
inman
jackie
janet
jan
jason
jay
jeanette
jeanne
jennifer
jensen
jerry
jessen
jessica
john
jonathan
joseph
joyce
karen
kathy
kaye
kelly
kelton
kendall
leslie
lichens
lisa
lori
louvier
mahoney
majorie
margaret
mary
mayorga
melanee
melanie
melissa
michelle
montgomery
moore
morgan
muir
mulholland
nagel
nancy
neddy
neilson
newman
olds
packham
patrick
paul
pearl
peggy
polly
porter
purtyman
rachel
ralph
randy
raymond
reichman
richard
roberts
rogers
ronald
rosell
rosemary
rose
rousche
sam
sanchez

sara	scott	shane
shanna	smithman	snow
stephanie	steth	stevens
steven	stonehawker	sue
suggs	summers	suzanne
suzy	tannenbaum	terri
thompson	torres	tracy
trobinino	vaughan	vicki
von	waltmeirs	waters
wells	wempy	wilbur
wilmede	wood	

APPENDIX B

Observations on Random sampling of Auto-generated dictionaries

The following are the phone changes observed in the inferred surface forms of TIMIT database. They are compared against the canonical baseforms of the TIMIT dictionary.

Type of change	Sample inferred variants		Frequency	Comments
Consonant Deletion	BEHIND(1) LARGE(2)	b ih hh ay n l aa zh	Often	Plosives, semivowels, word ending consonants most affected.
Consonant Substitution	VIRTUE(1) ASIDE	f r ch y uw ah s ay t	Often	Usual substitutions are within the voiced/unvoiced pairs of stops and fricatives; substitutions among nasals.
Consonant Substitution	LONG(1) WITH(11)	l ao ng g hh w iy t th	Rare	-
Vowel Deletion	SHUFFLED	sh ah f l d	Rare	The phoneset used does’nt have reduced phones (IX, AX) leading to some vowels being omitted.

Contd...

Vowel Insertion	ABOUT(2) CLAIM	ah b aa aw t k l ey iy m	Often	Frames seem to be shared between more than one vowels where unwanted.
Vowel Substituted	ADDED(2) ACROSS	ae d ih d ah k r aa s	Often	Usual substitutions between short/long vowels. Possible substituions within front/mid/back vowels.

REFERENCES

[1] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, pp. pp. 1–8, 1970.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39(1), pp. pp. 1–39, 1977.

[3] P. J. Jang and A. G. Hauptmann, "Hierarchical cluster language modeling with statistical rule extraction for rescoring n-best hypotheses during speech decoding," in *Proc. of ICSLP*, (Sydney, Australia), 1998.

[4] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. IT-13, pp. pp. 260–269, 1967.

[5] B. T. Lowerre, *The harpy speech recognition system*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1976.

[6] M. Ravishankar, *Efficient Algorithms for Speech Recognition*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1996.

[7] M. Woszczyna, *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition*. PhD thesis, University of Karlsruhe, Germany, 1998.

[8] V. Venkataramani, *Code breaking for automatic speech recognition*. PhD thesis, Johns Hopkins University, Baltimore, MD, USA, 2005.

[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[10] L. L. Chase, "Blame assignment for errors made by large vocabulary speech recognizers," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 1563–1566, 1997.

[11] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data," in *Proc. Eurospeech*, (Madrid), 1995.

[12] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch," in *Proc. ICSLP '98*, (Sydney, Australia), 1998.

[13] S. Greenberg, "Speaking in shorthand– a syllable-centric perspective for understanding pronunciation variation," in *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, (Kekrade, Netherlands, May 1998. ESCA.), 1998.

[14] M. Saraclar, *Pronunciation Modeling for Conversational Speech Recognition*. PhD thesis, Johns Hopkins University, Baltimore, MD, USA, 2000.

[15] D. Jurafsky, "What kind of pronunciation variation is hard for triphones to model," in *Proc. ICASSP*, (Salt Lake City, UT, May 2001), 2001.

[16] Eric John Fosler-Lussier , "Dynamic Pronunciation Models for Automatic Speech Recognition," Tech. Rep. TR-99-015, University of California, Berkeley, Berkeley, CA, 1999.

[17] Finke Michael and Waibel Alex, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. Eurospeech*, 1995.

[18] J. M. Kessens, M. Wester, and H. Strik, "Improving the performance of a dutch csr by modeling within-word and cross-word pronunciation variation," *Speech Commun.*, vol. 29, no. 2-4, pp. 193–207, 1999.

[19] H. J. Nock and S. J. Young, "Detecting and improving poor pronunciations for multiwords."

[20] M. Ravishankar and M. Eskenazi, "Automatic generation of context-dependent pronunciations," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 2467–2470, 1997.

[21] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP '96*, (Philadelphia), pp. 2328–2331, 1996.

[22] A. Xavier and D. Christian, "Improved acoustic-phonetic modeling in Philips' dictation system by handling liaisons and multiple pronunciations," in *Proc. Eurospeech '95*, (Madrid), pp. 767–770, 1995.

[23] P. S. Cohen and R. L. Mercer, "The phonological component of an automatic speech-recognition system," *Reddy, D.R. (Ed) Speech Recognition. Invited Papers Presented at the 1974 IEEE Symposium.*, pp. 275–319, 1975.

[24] N. Cremelie and J.-P. Martens, "In search of better pronunciation models for speech recognition," *Speech Communication*, vol. 29, no. 2-4, pp. 115–136, 1999.

[25] B. C.S. and Y. S.J., "Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from x-ray data," in *Proc. ICSLP '96*, (Philadelphia), pp. 969–972, 1996.

[26] Adda-Decker M. and Lamel L., "Pronunciation variants across system configuration," *Speech Communication*, 1999.

[27] I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Joint pronunciation modeling of non-native speakers using data-driven methods," in *Proc. ICSLP '00*, (Beijing, China), pp. 622–625, 2000.

[28] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Commun.*, vol. 29, no. 2-4, pp. 99–114, 1999.

[29] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks," *Speech Commun.*, vol. 27, no. 1, pp. 63–73, 1999.

[30] T. Holter and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," *Speech Commun.*, vol. 29, no. 2-4, pp. 177–191, 1999.

[31] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," *Speech Commun.*, vol. 29, no. 2-4, pp. 225–246, 1999.

[32] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Commun.*, vol. 29, no. 2-4, pp. 209–224, 1999.

[33] Grace Chung and Staphanie Seneff and Chao Wang and I. Hetherington, "A dynamic vocabulary spoken dialogue interface," in *Proc. ICSLP*, (Jeju Island), 2004.

[34] H. T.Svendsen, F.K.Soong, "Optimizing baseforms for HMM-base speech recognition," in *Proc. Eurospeech*, 1995.

[35] J. M. Kessens, C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation," *Speech Commun.*, vol. 40, no. 4, pp. 517–534, 2003.

[36] G. Chung, C. Wang, S. Seneff, E. Filisko, and M. Tang, "Combining linguistic knowledge and acoustic information in automatic pronunciation lexicon generation," in *Proc. of ICSLP*, (Jeju Island, Korea), 2004.

[37] S. Deligne and L. Mangu, "On the use of lattices for automatic generation of pronunciation," in *Proc. ICASSP*, (Hongkong, China), 2003.

[38] B.Ramabhadran, L.R.Bahl, P. DeSouza, and M. Padmanabhan, "Acoustics-only based automatic phonetic baseform generation," in *Proc. ICASSP*, (Seattle, USA), 1998.

[39] M. Riley and A. Ljolje, "Automatic generation of detailed pronunciation lexicons," *Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer.*, 1995.

[40] A. W. Black, "Festvox: Building New Synthetic Voices." http://www.festvox.org.

[41] S. Deligne and B. Maison and R. Gopinath, "Automatic generation and selection of multiple pronunciations for dynamic vocabularies," in *Proc. ICASSP*, (Salt Lake City, UT), 2001.

[42] I. Bazzi and J. Glass, "A multi-class approach for modelling out-of-vocabulary words," in *Proc. ICSLP*, (Denver, Colorado), 2002.

[43] "CMUDICT: CMU pronunciation dictionary." http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[44] John S Garofolo, "TIMIT acoustic-phonetic continuous speech corpus," 1993. Linguistic Data Consortium, Philadelphia.

[45] "CMUsphinx, The Carnegie Mellon Sphinx Project." http://cmusphinx.sourceforge.net.

[46] "Names v1.3, the CSLU OGI names corpus." http://cslu.cse.ogi.edu/corpora/names/.

[47] T. Hazen, I. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," PMLA, 2002.

[48] Deng L. and Sun D., "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Amer.*, 1994.

[49] K. Prahallad, A. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toulouse, France), 2006.

[50] M. Magimai.-Doss and H. Bourlard, “On the adequacy of baseform pronunciations and pronunciation variants,” IDIAP-RR 27, IDIAP, 2004.

LIST OF PUBLICATIONS

The work done during my masters has been disseminated to the following conferences/journals:

1. Gopala Krishna Anumanchipalli, Mosur Ravishankar and Raj Reddy, ”Improving Pronunciation Inference using N-Best list, Acoustics and Orthography ”, in Proceedings of *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, USA, 2007.
2. D. Bohus, S. Grau, D. Huggins-Daines, V. Keri, G. Krishna, R. Kumar, A. Raux, and S. Tomko, ”Conquest - an Open-Source Dialog System for Conferences”, in Proceedings of *HLT-NAACL ’07*, Rochester, NY, USA, 2007.
3. Gopala Krishna Anumanchipalli, Kishore Prahallad and Mosur Ravishankar, ”Pronunciation Modeling Using Lexical and Acoustic Information for Speech Recognition”, Under Review at *IEEE Intl. Conf. of Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, 2008.
4. Gopala Krishna Anumanchipalli, Kishore Prahallad and Alan W Black, ”Significance of Early Tagged Contextual Graphemes in Grapheme Based Speech Synthesis and Recognition Systems”, Under Review at *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, 2008.
5. Gopala Krishna Anumanchipalli, Kishore Prahallad and Mosur Ravishankar, ”Data-driven Lexical Modeling for Continuous Speech Recognition”, In Preparation for *Speech Communication*.

CURRICULUM VITAE

1. **NAME:** Gopala Krishna Anumanchipalli

2. **DATE OF BIRTH:** 24 June 1984

3. **PERMANENT ADDRESS:**
   Gopala Krishna Anumanchipalli
   S/O: A.V.S. Narayana
   43-9-9 Rly. New Colony
   Visakhapatnam 530016
   Andhra Pradesh, India

4. **EDUCATIONAL QUALIFICATIONS:**

   - June 2006: Bachelor of Technology in Computer Science and Engineering (Hons.), IIIT Hyderabad

   - November 2007: Master of Science (by Research) in Computer Science and Engineering, IIIT Hyderabad

THESIS COMMITTEE

1. **GUIDES:**

   - Mr. S. Prahallad Kishore
   - Dr. Ravishankar Mosur

2. **MEMBERS:**

   - Prof. B. Yegnanarayana
   - Dr. G. Ramamurthy