PHONE DELETION MODELING IN SPEECH RECOGNITION

by

KO, YU TING

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Computer Science and Engineering

August 2010, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

KO, YU TING

PHONE DELETION MODELING IN SPEECH RECOGNITION

by

KO, YU TING

This is to certify that I have examined the above M.Phil. thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

DR. BRIAN KAN-WING MAK, THESIS SUPERVISOR

PROF. MOUNIR HAMDI, HEAD OF DEPARTMENT

Department of Computer Science and Engineering

16 August 2010

ACKNOWLEDGMENTS

I would like to express my sincere thankfulness to Dr. Brian Mak for his supervision throughout my MPhil study. He taught me not only the knowledge on speech recognition, but also helped sharpen my analytical and presentation skills. Thank Dr. Manhung Siu for introducing me to Dr. Brian Mak and thank Dr. Dit-Yan Yeung and Dr. Tan Lee for being my panel.

I would like to express my gratitude to my colleagues including Benny Ng and Ye Guoli. I learnt a lot from them in the past 2 years.

I would also like to thank my mother and my wife for their patience and consistent support for my study. They granted me great freedom in my career.

Last but not least, thank God for answering my prayer to have the opportunity to experience a research life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PHONE DELETION MODELING IN SPEECH RECOGNITION

by

KO, YU TING

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

ABSTRACT

In a paper published by Greenberg in 1998, it was said that in conversational speech, phone deletion rate may go as high as 12%. On the other hand, Jurafsky reported in 2001 that phone deletions cannot be modeled well by traditional triphone training. These findings motivate us to model phone deletions explicitly in current ASR systems. In this thesis, phone deletions are modeled by adding skip arcs to the word models. In order to cope with the limitations of using whole word models, context-dependent fragmented word models(CD-FWMs) are proposed. Our proposed method is evaluated on both read speech (Wall Street Journal) and conversational speech (SVitchboard) task. In the read speech evaluation, we obtained a word error rate reduction of about 11%. Although the improvement in conversational speech is modest, reasons are given and relevant analyses are carried out.

CHAPTER 1

INTRODUCTION

1.1 Background

Pronunciation variations are one of the major reasons that make automatic speech recognition (ASR) a hard task. People pronounce the same word in different ways. There is a reason for this phenomenon. For example, if a group of people migrate to another place, after a certain amount of time, their pronunciation may be different from the original group. This happens because of the lack of contact between the two groups and also the influence from the local people. It is not surprising there are a great deal of variations in English pronunciation, seeing as it is the most widely used language throughout the world. Even the people who live in the southern parts of the U.S. pronounce words in a slightly different way from the people who live in the northern parts. It is highly improbable that people would check their dictionaries everyday to see what the canonical pronunciation of certain words are. Eventually, they may forget about the canonical pronunciation and pronounce in a very different way. In the U.S., there are 80 ways of pronouncing the word "AND" [17]. In Hong Kong, people often pronounce the word "CAT" as [k ae][1], missing a [t] sound from the canonical pronunciation [k ae t]. Also, they often pronounce the word "PAPER" as [p ey p ah] of which the canonical pronunciation is [p ey p er].

State-of-the-art recognizers can easily attain a word error rate (WER) of less than 10% in read speech, which could be accurate enough for some applications. However, they can only attain a WER of 30%-40% in conversational speech. One of the major reasons for the drop in accuracy may be attributed to pronunciation variations in speech. People speak very differently in dictation and conversation. In dictation, people try to pronounce the words in a standard way, but this is unnatural in conversation. Therefore, researchers try to investigate the best ways to model pronunciation varia-

[1]The pronunciation of the whole phone set is listed in Table A.1

tions in order to narrow the gap in the accuracy between read speech and conversational speech.

Given that most ASRs consist of three components, there are three levels at which variations can be modeled [1]: the lexicon, the acoustic models and the language model (LM). Pronunciation variation can be modeled at different levels simultaneously.

Most pronunciation variation modeling techniques are attempted at the lexicon level [2, 3, 4]. At this level, pronunciation variation is usually modeled by adding pronunciation variants (phonetic transcriptions) to the lexicon (dictionary). Substantial gains are achieved by adding these new entries to the dictionary until it reaches a point where the gains are offset by the increase in confusability of words. This means that adding pronunciation variation entries to the dictionary may avoid some old errors but may creates some new errors at the same time. In order to determine which set of pronunciation variants leads to the largest gain, different criteria are used, such as the frequency of occurrence of the variants [2] or a maximum likelihood criterion [3].

Although adding pronunciation variants to the lexicon is the most common way, it does not include the probabilities of the variants. For the sake of precise modeling, the statistical behavior of the variants can be captured at the level of the LM [10] or a third party component [6, 7].

Modeling pronunciation variations at the lexicon level, for example, looking for a better dictionary, has resulted in a substantial improvement in terms of WER. On the other hand, people have tried to further improve the system by modeling variations at the acoustic model level. The question this has brought about is: Are there any acoustic units better than conventional phone units in modeling pronunciation variations? S. Greenberg tried to answer this question with a syllable-centric perspective [17]. He carried out a systematic analysis of pronunciation variations in 1998 with the use of a conversational English speech corpus, the Switchboard [31]. His paper concluded that syllables are more stable linguistic units for pronunciation modeling than phones. These findings prompted a new research direction in the automatic speech recognition (ASR) community to investigate the modeling of syllables and other long units as the acoustic units for ASR. To date, the long unit approach has not yet fulfilled its promise. The disappointing results can be attributed to the following factors:

- The exponentially increased number of units compared to phone modeling.
- The data sparsity problem due to the huge amount of units.

As the number of parameters generally increases with the number of acoustic units, more data are needed for their reliable estimation. Since data are always limited and unbalanced, the advantages of these long unit systems were offset by some poorly trained models. State tying is proven to be a good way to address this problem, however, it is well developed on phone level models only. Owing to these constraints, the advantages of using long-span units are not obvious and the improvement is small.

In [23], they used a fragmented unit approach to limit the growth in number of units but at the same time keep the context dependence between the units. For example, the unit "p^ey^p^er" is cut into three segments "p", "er^p" and "er"so that only the head and the tail phones are exposed as context.

In this thesis, we try to model phone deletions explicitly by implementing skipping arcs in the acoustic model. In practice, we have to choose a linguistic unit larger than a phone to hold the skipping arcs. In order to place as many skip arcs as possible, we choose to perform word modeling. At the same time, we use a fragmented unit approach similar to [23] so as to cope with the limitations in long unit modeling.

1.2 Thesis Outline

This thesis is organized as follows: A review of ASR basics is given in chapter 2. This includes a review of hidden Markov model (HMM) and phone-based acoustic modeling.

In chapter 3, existing methods of modeling pronunciation variation are reviewed.

In chapter 4, our proposed explicit modeling of phone deletions is presented. The need to construct context-dependent fragmented word models (CD-FWM) is explained.

In chapter 5, experimental evaluations are described in detail. Analysis is made and the effectiveness of our proposed method is investigated. The WER of our approach and the baseline are compared. Conclusion and future works are discussed in the last chapter.

CHAPTER 2

ASR BASICS

2.1 Phonemes and Phones

In spoken language, a phoneme is defined to be the smallest, abstract unit of sound that can distinguish words. For example, there is one phoneme different in the word pair "DOG" and "FOG" that makes them different. Phonemes are highly related to the sounds, but they are not the sounds. Phones are sounds. Phonemes and phones are not one to one mapping. Several phones may belong to the same phoneme and they are called allophones. Phonemes can be considered as the elementary unit occur in our brain when we speak.

In linguistics, phonemes and phones are totally different. But in engineering perspective, they are assumed to be the same. In phone-based modeling, each phoneme is treated as a unique phone and being modeled. The assumption behind is obviously wrong because the number of distinct sounds is far more than the number of phonemes. This results in a weak discriminative power among the models and therefore the recognition performance is bad. This also explains why triphone modeling is much better than monophone modeling in recognition performance because triphone modeling has covered much of the distinct sounds by largely increasing the number of units.

In this thesis, phonemic transcriptions are enclosed by slashes (/ /) and phonetic transcriptions are enclosed by square brackets ([ ]).

2.2 Major Components in ASR System

A common ASR system usually consists of three major components: a dictionary, an acoustic model and a language model. Their functionalities are described as follows:

- ***Dictionary***: It defines the pronunciation of words by listing out their phonetic transcription. An example showing how the dictionary looks like is shown in Table 2.1.

Table 2.1: An example of dictionary.

Word	Phonetic Transcription
ABOUT	ah b aw t
CONSIDER	k ah n s ih d er
CAT	k ae t
DOG	d ao g
HUNDRED	hh ah n d r ah d

- ***Acoustic model***: It describes the statistical behavior of the acoustic signal in the feature space. It consists of a set of hidden Markov models representing each of the basic units. This is going to be discussed in more detail in the coming section.

- ***Language model***: It describes the relationship between words and it normally encapsulates the information of English grammar. For example, "IN ORDER" is usually followed by the word "TO"

2.3 Hidden Markov Model

For ease of description, let us define:
λ: an HMM model (normally means all the parameters in the model),
a_{ij}: the transition probability from state i to state j,
J: the total number of states in the HMM λ,
T: the total number of frames in the observation vector sequence $\mathbf{X}$.
x_t: an observation vector at time t,
X: a sequence of T observation vectors, $[x_1, x_2, \ldots, x_T]$,
q_t: the state at time t,
W: the state sequence, $[q_1, q_2, \ldots, q_T]$,

The hidden Markov model is a finite state machine. In case of a continuous HMM, each state is associated with a probability density function (pdf), which is usually a mixture of Gaussians. Transitions among the states are associated with a probability a_{ij} representing the transition probability from state i to state j. HMM is a generative statistical model. In each time step t, the system transits from a source state q_{t-1}

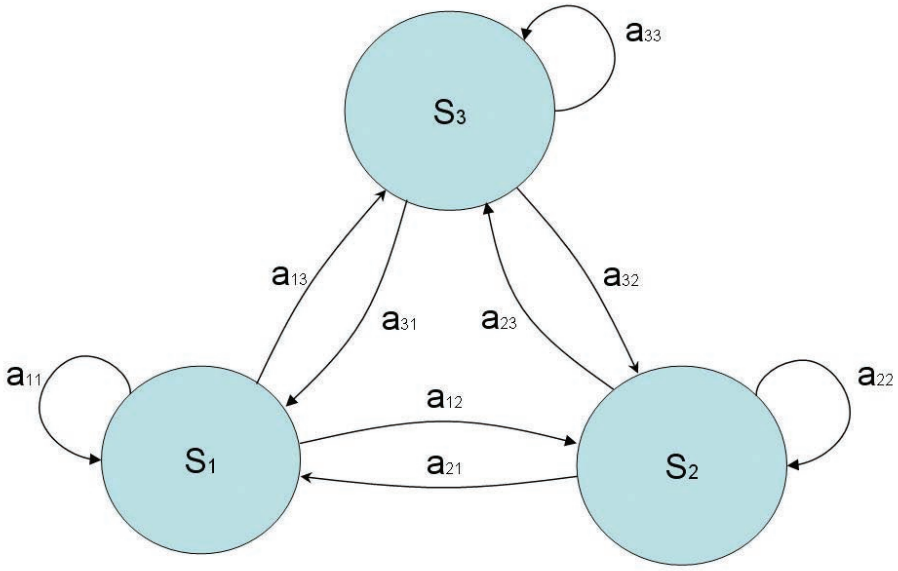

Figure 2.1: An example of HMM with 3 states.

to a destination state q_t and an observation vector x_t is emitted. The distribution of this emitted x_t is governed by the probability density function in the destination state. The model parameters are the transition probabilities and the parameters of the set of probability density functions. The model complexity is used to measure the amount of parameters in the model. An example of a first-order HMM is shown in Fig. 2.1.

In a hidden Markov model, the state sequence is not observable whereas only the observations generated by the model is directly visible. The "hidden" Markov model is so named because of the hidden underlying state sequence.

There are three major issues in hidden Markov modeling:

- ***The Evaluation issue*** : From a generative perspective, any sequences of observations can be generated by a model within a certain time duration. Given the HMM parameters λ, it is possible to determine the probability $P(X|\lambda)$ that a particular sequence of observation vector X is generated by the model. In this case, the model parameters λ and the observation vector X are the inputs, and the corresponding probability is the output.

- ***The Training issue*** : From a training/learning perspective, the sequence of observation vector X is given whereas the model parameters λ are unknown. The observed data give us some information about the model and we can use them

to estimate the model parameters λ. The given data used for estimation are regarded as the training data. In this case, the observed data X is the input, and the estimated model parameters λ are the outputs.

- ***The Decoding issue*** : In a decoding process, the model parameters λ and the sequence of observation vector X is given where the sequence of states W is unknown. The goal is to look for the most likely sequence of underlying states W which maximizes $P(W|X, \lambda)$. In this case, the model λ and the observation vectors X are the inputs, and the decoded sequence of states W is the output.

2.3.1 Assumptions in the Theory of HMM

There are two major assumptions made in the theory of first-order HMMs:

- ***The Markov assumption***: It is assumed that in first-order HMMs the transition probabilities to the next state only depend on the current state and not on the past state history. Given the past k states,

$$P(q_{t+1} = j|q_t = i_1, q_{t-1} = i_2, \ldots, q_{t-k+1} = i_k) = P(q_{t+1} = j|q_t = i_1), \qquad (2.1)$$

  where $1 \leq i_1, i_2, \ldots, i_k, j \leq J$.

  On the other hand, the transition probabilities of a k^{th}-order HMM depend on the past k states.

- ***The output independence assumption***: It is assumed that given its emitting state the observation vector is conditionally independent of the previous vectors as well as the neighbouring states. Hence, we have

$$P(X|W, \lambda) = \prod_{t=1}^{T} P(x_t|q_t, \lambda). \qquad (2.2)$$

  If the states are stationary, the observations in a given state are assumed to be independently and identically distributed (i.i.d.).

2.4 Phone-based Acoustic Modeling

In phone-based acoustic modeling, the basic modeling units are phones. Each distinct phone in the phone set is modeled by an HMM. The acoustic model consists of a set of phone HMMs. HMMs are used because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal.

2.4.1 Usage of HMM as a Phone Model

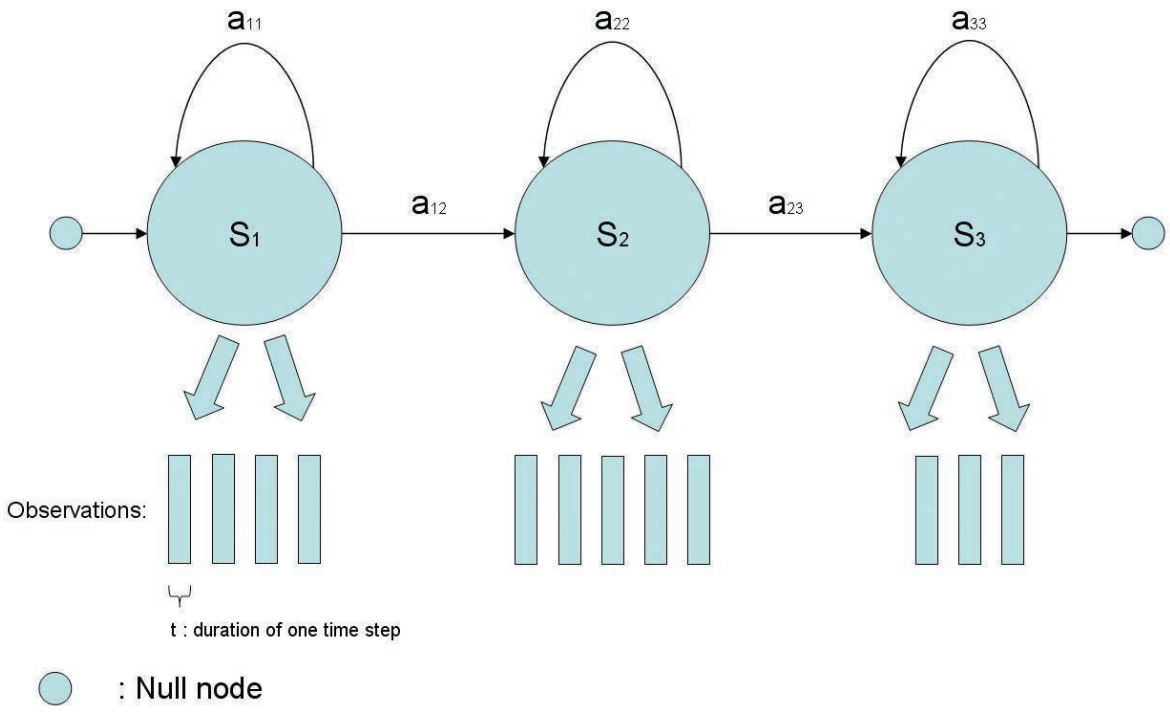

Figure 2.2: An example of a 3-state left-to-right HMM.

An example of HMM which is most commonly used to model a phone is shown in Fig. 2.2. It can be treated as a special form derived from the general form in Fig. 2.1 by setting $\{a_{13}, a_{21}, a_{31}, a_{32}\}$ to zero. It is a 3-state straightly left-to-right HMM in which only left-to-right transitions are allowed in order to capture the sequential nature of speech. The first and the last nodes are null nodes, they are non-emitting states which will not generate any observations and are used to indicate the entry and exit states. This specific structure makes it easy to connect with another HMM to form a longer HMM. For example, several phone HMMs may connect together to form a syllable HMM or a word HMM.

In Fig. 2.2, the rectangle blocks are the acoustic observations emitted by the HMM. In this thesis, each of the acoustic observations is a 39-dimensional vector in the feature space.

The continuous state observation probability density functions are commonly estimated as mixtures of Gaussian densities with diagonal covariances due to its simplicity and trainability.

2.5 Context Dependence

It is observed that the acoustic behavior of a certain phone is highly influenced by its neighbouring phones due to coarticulation. In 1985, context-dependent phone models were proposed and the idea was to replace a single phone model by a number of detailed models which are different from each other with different neighbouring phones. As a result, there are two major kinds of units: context-independent (CI) units and context-dependent (CD) units:

- ***Context-independent (CI) units***: In phone-based modeling, context-independent units are also called monophones. Each phone is modeled by a single HMM. It is assumed that each phone is acoustically independent from their neighbouring phones in an utterance. This assumption obviously violates the fact that neighbouring units do affect each other (the co-articulatory effects) and therefore the performance of using CI units is modest.

- ***Context-dependent (CD) units***: In order to model the co-articulatory effects in speech, context-dependent units are developed. In phone-based modeling, the most common context-dependent unit is triphone of which the realization depends on its preceding and following phones. For example, both of the models "p-er+t" and "b-er+m" are modeling the phone [er], but they are different from each other with different preceding and following phones. Here, the phone before '-' is the preceding phone and the phone following '+' is the following phone.

  Since we do not know the sequence of phones in the testing utterances, we have to consider all combinations of tri-units during recognition. If there are N mono-units, there will be altogether N^3 tri-units. This procedure is called tri-unit expansion.

One thing we have to pay attention is the number of total units. Since the number of model parameters increases with the number of modeling units during the tri-unit expansion, we need more training data to estimate the parameters. As the amount of training data is always limited, data sparsity will occur if N is too large. In a typical phone-based modeling system[1], N is about 40.

Normally, triphone modeling can reduce more than 60% of the errors in speech recognition compared to monophone modeling.

2.6 Parameter Tying

Although the acoustic models should contain enough fine acoustic details so as to increase the discriminative power for decoding, finer details usually result in more model parameters and decrease their trainability. While using context-dependent units can better model the nature of speech, the large number of units (about 125,000 triphones in English) makes it almost impossible to reliably model each of them. One common solution is to use the well trained models to help those badly trained models by tying their parameters together. The technique of parameter tying is proven to be successful in reducing the number of parameters in the acoustic model and, at the same time, preserving its performance in recognition. Below are some of the most typical HMM parameter tying approaches at various levels:

- At the phone level: A monophone model can be regarded as tying all the context-dependent models of a monophone into one model. This results in a context-independent system and its model complexity is small as the number of phones in a typical phone set is about 40. However, the accuracy is also modest. In brief, it ties too much.

- At the HMM level: In this case, the whole HMM are tied. For example the whole model of "ay-t+f" and "ay-t+hh" are being tied, meaning that they are referring to the same HMM.

- At the state level: As described in the previous section, each phone is represented by a 3-state HMM. Since co-articulatory effect is more prominent at the beginning

[1]In a phone-based modeling system, N is equal to the number of monophones in the phone set.

and ending of a phone, it is suggested that tying should be done on a particular state rather than the whole phone HMM. States are tied means that they share the same observation probability density function.

CHAPTER 3

REVIEW OF EXISTING METHODS IN MODELING PRONUNCIATION VARIATION

In the past decades, the type of speech used in ASR research has gradually progressed from isolated words to connected words, carefully read speech, and finally conversational or spontaneous speech. It is no doubt that the pronunciation variation increases when going from isolated words to conversational speech. As a matter of fact, there had been an increase in the amount of research on modeling pronunciation variations.

In 1998, the ESCA(now ISCA) workshop on modeling pronunciation variation for automatic speech recognition [1] carried out an in-depth analysis on a large amount of pronunciation variation modeling techniques. It concluded with several major characteristics existing in the techniques:

- Target of pronunciation variation,
- Information sources,
- Information representation and
- Level of modeling.

With these characteristics, different methods for pronunciation variation modeling can be distinguished and it is easier for us to perform an objective comparison of the methods. In this chapter we are going to illustrate the major characteristics of different pronunciation variation modeling techniques described above and at the same time do a review on these techniques.

3.1 Target of Pronunciation Variation

The target of pronunciation variation to be modeled can be divided into two main categories: within-word variation and cross-word variation. Making use of a lexicon

with phonetic transcription entries of words, modeling within-word variation is an easier choice because the variants can be simply added to the lexicon. On the contrary, modeling cross-word pronunciation involves context dependency for each word. For example, a variation of a word A often occurs when it is preceded by a particular word B, while the same variation hardly occurs in other cases.

As a matter of fact, there are relatively more attempts in modeling within-word variation [8, 3, 2] than in modeling cross-word variation [6, 7].

A compromise between the ease of modeling at the level of the lexicon and the need to model cross-word variation is to use multi-words [10]. In this case, sequences of words (usually frequently occurring phrases) are treated as one single entity in the lexicon. For example, the two words "THEY" and "ARE" are combined to form a new entity "THEY'RE" in the lexicon. Therefore the variants of these phrases can be added to the lexicon.

3.2 Information Sources

The information sources define how the pronunciation variation is obtained. According to this criterion, the methods can be divided into knowledge-based methods and data-driven methods.

3.2.1 Knowledge-based Methods

In knowledge-based methods [9, 10], the pronunciation variation is derived from the prior knowledge of a particular language. In this case, phonological rules and linguistic studies of a particular language take an important role. Pronunciation variants are generated according to the characteristics of the language. For example, it is well known that schwas and liaisons are frequently occurred in French and the pronunciation variants can be generated by making use of this information. Therefore, a drawback of knowledge-based methods is that the methods are not portable across different languages. Furthermore, simply making use of the phonological rules may cause too much generalization where the benefit is canceled out by the ambiguity introduced by the variants. As a result, people are becoming more interested in data-driven methods.

3.2.2 Data-driven Methods

A data-driven method means that the pronunciation variation is estimated from the training data. A typical way of doing this is by aligning two phonetic transcriptions of each training utterances, a standard transcription (base form) and a more realistic auditory transcription (surface form).

The standard transcription is directly obtained from the dictionary which is composed of the canonical pronunciations of words. This represents how the words should be pronounced. The auditory transcription can be obtained either by human expert (hand-labeled) [5, 6] or by automatic generation [7]. This transcription is highly related to the acoustic signals as it represents how the words are really pronounced by the speakers.

Hand-labeled transcriptions are believed to be better in phone error rate, but it is extremely time-consuming. Automatic ways in generating transcriptions like using a phoneme recognizer [8] or flexible forced alignment decoding [7] still work well when manual transcriptions are not supported.

Table 3.1: Phonetic transcription alignment of the word "DOCUMENTATION".

Base form	d	aa	k	y	ah	m	eh	n	t	ey		sh	ah	n
Surface form	d	aa	k	y	uw	m	eh		t	ey	t	sh	ah	n

Using the word "DOCUMENTATION" as an example, the standard and the auditory transcriptions are aligned with each other in Table 3.1. In this example, [ah] is substituted by [uw]; [n] is deleted and [t] is inserted. Pronunciation variation can be learnt from the differences between the two transcriptions. The question now is how we represent the pronunciation variation.

3.3 Information Representation

As mentioned in the previous section that pronunciation variation can be known from either a data-driven or knowledge-based source. The choice now is to decide the way to represent this information, which depends on whether the information is being formalized or not and how to formalize the information.

3.3.1 Formalization Methods

In general, formalization means using a more abstract and compact representation, where the focus is on general variation rules that are learnt from the surface form transcription rather than on the variation of individual words. The rules learnt are then used to generate the variants from the base form pronunciation of the words. The formalization can be done by various approaches like decision trees [5], neural network [8] or rewrite rules [6, 7].

3.3.1.1 Rewrite Rules

In [6, 7], pronunciation variation is captured by a number of rewrite rules which are in statement form. After aligning the surface form transcription to the base form transcription, any conflict between the two transcriptions leads to one or more rewriting rules. Using the case in Table 3.1 as an example, a rewrite rule is generated for the substitution of [ah] meaning that [ah] is very likely to be pronounced as [uw] when it is on the right of [m] and on the left of [y]. Similarly, two more rewrite rules are generated for the deletion of [n] and the insertion of [t]. The rewrite rules are then used to construct the pronunciation network of the words. The probabilities within the pronunciation network is estimated by counting the number of times the rule had to be applied on the training utterances.

3.3.2 Enumeration Methods

In contrast to the above-mentioned methods of formalization, one can simply list out all the pronunciation variants of the same word from the surface form transcription according to the word boundaries, and then prune the pronunciation entries of the same word using a threshold value. This kind of methods is regarded as enumeration.

At this point, it is difficult to judge which kind of methods will certainly work better than the others. The drawback of enumeration is that it cannot predict the pronunciation variation of new words, and may work badly if there are few training samples of individual words. On the other hand, methods of formalization usually suffer from overgeneration and undergeneration.

3.4 Level of Modeling

Modern ASR systems usually consist of three components: the lexicon, the acoustic model and the language model. Different pronunciation variation modeling techniques are developed at these three different levels. The good thing is that variation modeling can be done on different levels simultaneously in order to capture different aspects of pronunciation variation.

3.4.1 Lexicon

Pronunciation variation at the lexicon level may be modeled by simply adding pronunciation variants to the lexicon. The probability of a correct word being selected is increased because of a better match of the incoming signal with the newly added transcription entries of the word. As a result, this fixes some errors and leads to a better performance.

However, adding pronunciation variants to the lexicon brings ambiguity into the system at the same time. This is because the added variants may be confused with some existing pronunciation entries of other words and therefore new errors are introduced. Recognition performance starts to degrade by adding more variants when the improvement is cancelled out by the increasing confusability. Therefore, different criteria are suggested to determine which set of pronunciation variants leads to the largest gain in the performance of the system. For example, frequency of occurrence of the variants [2], degree of confusability between the variants [4], and a maximum likelihood approach [3] have been tried. Because of its simplicity, frequency of occurrence of the variants in the training data is the most popular criteria in selecting the variants. That is, the variants of which the number of occurrence is greater than a threshold are added to the lexicon.

3.4.2 Acoustic Model

3.4.2.1 Iterative Forced-Alignment Training

A general way to optimize an acoustic model is to retrain the model with a better transcription of the training data. A better transcription may be obtained through

generating forced alignments between the acoustics and the words. In the generation of forced alignments, the words of the utterances are given and the recognizer may choose the one which gives the highest likelihood to the acoustic among several pronunciation variants of the words. This way of training is usually used together with the lexicon level pronunciation modeling described in the previous section. In general, the iterative forced-alignment training is done in the following steps:

1. Train the first set of acoustic models using a canonical lexicon.
2. Obtain a multiple-pronunciation dictionary using any approach.
3. Generate forced alignments to improve the transcription of the training corpus.
4. Train new models using the improved transcriptions.
5. Repeat step 4 and 5 until the transcription no longer changes.

This has become a standard procedure in modern ASR training.

3.4.2.2 Other Basic Units

Most ASR systems use phones as the basic modeling units, however there are attempts at using various modeling units other than phone. It is suggested that using a long unit like a syllable is a better choice since it should be able to encapsulate all the short term pronunciation variations. However, using long units is not trivial as it usually suffers from data sparseness. A more detailed description about the pros and cons of using long units will be given in the next chapter.

3.4.3 Language Model (LM)

While the previous section describes how the pronunciation variants are generated and listed in the lexicon normally the probabilities of these variants are generated at the same time. For example, if the rewrite rules are used to generate the variants, they can produce the probabilities of the variants according to the number of times the rules have been applied to the training data. Since the dictionary does not contain any

probabilistic information, the probabilities of variants can be handled in the language model.

As mentioned in the previous chapter, given a speech signal X, decoding is to find a string of words W which maximizes $P(X|W)P(W)$ where $P(X|W)$ is the acoustic score and $P(W)$ is the language score calculated by a language model. Now, each variant may be treated as a distinct unit and it is used to estimate a new language model [10]. Decoding then becomes finding a string of variants V which maximizes $P(X|V)P(V)$ where $P(X|V)$ is the acoustic score and $P(V)$ is a language score calculated by the new language model. One drawback of this approach is that the number of units in the LM is highly increased and aggravates the data sparseness problem.

A modified approach would be to introduce an intermediate level $P(V|W)$ so that the decoding criteria becomes $P(X|V)P(V|W)P(W)$ where $P(X|V)$ is the acoustic score, $P(W)$ is the original language model and $P(V|W)$ is the pronunciation score by a third party component named pronunciation network [6, 7].

CHAPTER 4

EXPLICIT MODELING OF PHONE DELETIONS WITH CONTEXT-DEPENDENT FRAGMENTED WORD MODELS (CD-FWM)

4.1 Different Kinds of Pronunciation Variation

First of all, pronunciation variations are divided into two types: phonemic variations and phonetic variations. Phonemic variations mean that the word has a totally different phoneme transcription, for example, the word "A" should be transcribed as phoneme /ah/ when it is pronounced in sentence like "... is a boy ...", but it should transcribed as phoneme /ey/ when it is pronounced in sentence like "... the A. B. C. company ...". This kind of pronunciation variation is expected to have a very different phonetic (sound) behavior, so it is well modeled at the lexicon level (by adding entries in the dictionary).

Phonetic variations are defined as any difference between the canonical (dictionary) phone sequence and the surface (hand-labeled) phone sequence. This can be further divided into three categories: Phone deletions, vowel reductions and phone substitutions. Their definitions are as follows:

- Phone deletions are cases in which the canonical pronunciation has a phone missing in the surface pronunciation. For example, the canonical transcription of the word "AND" has three phones ([ae n d]) but the surface transcription of some "AND" tokens has only two phones ([ae n]) where the [d] at the end is missing.

- Vowel reductions are cases in which the vowels in unstressed syllables are reduced to vowels which are shorter in duration and weaker in loudness. For example the word "IT" may be pronounced as its reduced form [ix t] whereas its canonical form is [ih t].

- Phone substitutions are cases where a certain phone in the surface pronunciation is distinct from the canonical pronunciation. For example, the standard pronunciation of the word "ENVIRONMENT" should be [ih n v ay r ah n m ah n t], but most Hong Kong people pronounce it as [ae n v ay r ah n m ah n t], the [ih] at the front is substituted by [ae].

4.2 Why Phone Deletions?

In order to enhance the research and development of ASR on conversational speech, a data corpus named Swichbaord was collected in 1992. Switchboard is a corpus of conversational speech which recorded through the telephone channel. It includes conversations made by paid volunteers of both sexes from every major dialect of American English. Till now, Switchboard is still one of the benchmark corpus when the ASR community reports recognition performance of conversational speech.

Jurafsky designed an interesting experiment in 2001 using the Switchboard corpus to investigate what kind of pronunciation variations were hard for traditional triphone modeling [14]. It turned out that the current method of triphones training could model phone substitution and vowel reduction quite well by using more training data, but had a problem with modeling syllable deletions. On the other hand, before Jurafsky's experiment, Greenberg had carried out an in-depth analysis in 1998 using the Switchboard corpus. He reported that the phone deletion rate in conversational speech may be as high as 12%. Inspired by both Greenberg's and Jurafsky's findings, we would like to model phone deletions explicitly in the acoustic model.

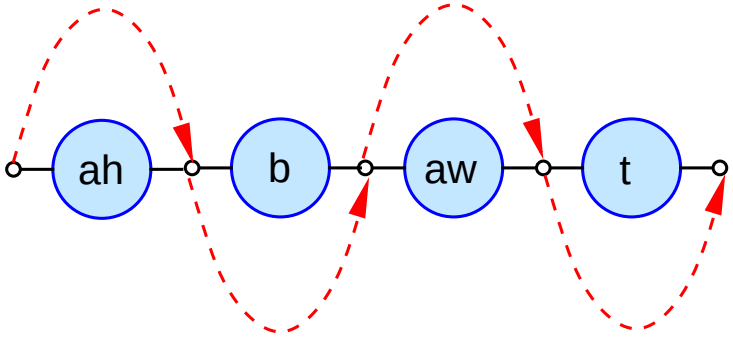

Figure 4.1: An example of adding skip arcs to allow phone deletions.

Our idea of modeling phone deletion by skipping arcs is simple as shown in Fig. 4.1. In practice, since we use the existing trainer and decoder (and in our case, HTK's), we have to choose a linguistic unit larger than a phone to implement phone deletions. In order to place as many skip arcs as possible, we choose to perform word modeling. However, word modeling is not trivial as it suffers from a few limitations, which are discussed in the following sections.

4.3 The Choice of Basic Units

The acoustic model in ASR consists of a set of HMMs associated with a linguistical meaning. These linguistic units should eventually be able to make up the words in the ASR task. In other words, the choice is how we break down the words and model them. For example, the word "PAPER" can be viewed as a concatenation of four phones ([p], [ey], [p] and [ah]), or two syllables[1] ("p^ey" and "p^ah"), or one single unit ("p^ey^p^ah"). In order to strike the balance between trainability and resolution, the choice of a modeling unit is important. For example, a word model is good in capturing word-specific information that leads to a strong discriminant power, but in most cases, there are insufficient training data for each word. Phone modeling has struck a good balance and is still the most popular approach till now.

4.4 Long Units Modeling

Both words and syllables are regarded as long-span units as they are longer than a phone in nature and duration. Over the past decades, long-span unit modeling, especially syllable modeling, has drawn great interest.

4.4.1 Supporting Arguments

Greenberg's seminal papers [18, 17] presented a syllabic-centric perspective for understanding pronunciation variation. One of his major findings from a systematic analysis of manually transcribed conversations from the Switchboard [31] corpus is that the syllable is a more stable linguistic unit for pronunciation modeling than the phone:

[1]The notation '^' means that they are the same acoustic unit.

- in Switchboard, the phone deletion rate is about 12% whereas the syllable deletion rate is about 1%.
- syllable onsets are well preserved; syllable nuclei may change; syllable coda are frequently disposed.

Also, most researchers [19, 20, 22] believed that:

- a syllable seems to be an intuitive unit for representation of speech sounds [12, 13]. Listeners can identify the number of syllables in a word easily, even the syllable boundaries.
- syllable and other units longer length should be able to capture long-span temporal dependencies and spectral variations [11] in speech.

The findings prompted a new research direction in the automatic speech recognition (ASR) community to investigate the modeling of syllables as the acoustic units for ASR [26, 27, 19, 20, 22, 28, 23], or to incorporate syllable information to improve speech recognition [24, 25].

4.4.2 Difficulties in Long Units Modeling

To date, the long span modeling approach has not yet fulfilled its promise, this can be attributed to the following reasons:

- Compared to phone modeling, the number of units in long-span units modeling is highly increased. We investigated the number of units using an English read speech corpus named WSJ. We found that there are around 5K distinct syllables and 13K distinct words in the corpus. In considering context-dependent units, the number of tri-syllables and tri-words system is $(5K)^3$ and $(13K)^3$ respectively. The number of phones in a traditional phone-based system is around 40, and the number of tri-phones is about $40^3 \simeq 64$K. Therefore, the number of units is highly increased in long-unit modeling.
- The data sparsity problem arises due to the huge number of units. With more models, more data are needed for a reliable estimation because the number of

parameters increases with the number of models. Therefore, data sparsity is always a problem for a system with huge number of units. The training data available is always limited and unbalanced. From the research results of syllable modeling, we learn that the advantages of using long units are often offset by a group of poorly trained models.

- State tying is usually a good way to cope with data sparsity as it shares the parameters among the sets of models in the system. With this technique, the parameters can be greatly reduced, and reliable training can still be performed with a limited amount of data. State tying uses a decision tree-based approach with rules based on linguistic knowledge of phonemes. However, the tying rules are only well developed on phone units, but not on syllable units or word units.

In order to cope with these difficulties, research efforts in the past mainly focused on:

- determining the mixing of phone units with syllable units [20, 21]?
- modeling context dependency in syllable modeling without an explosion of units [28, 23]?
- solving the data sparsity problem due to the explosion of syllable units, especially when context-dependent syllables were used [20, 28, 23]?

This thesis is not another attempt at long-unit modeling. Instead, we would like to seek a platform to explicitly model phone deletion.

4.4.3 Solutions to the Limitations

As mentioned above, state tying is a way to cope with data sparsity, but this needs to be done on a phone level. We therefore construct a word model that is bootstrapped from a state-tied cross-word triphone model. An example of the construction of a context-independent word model from word-internal triphones is shown in Fig. 4.2. Suppose we want to construct a model for the word "ABOUT", of which the phone transcription consist of four phones ([ah b aw t]). We hereby represent this model as

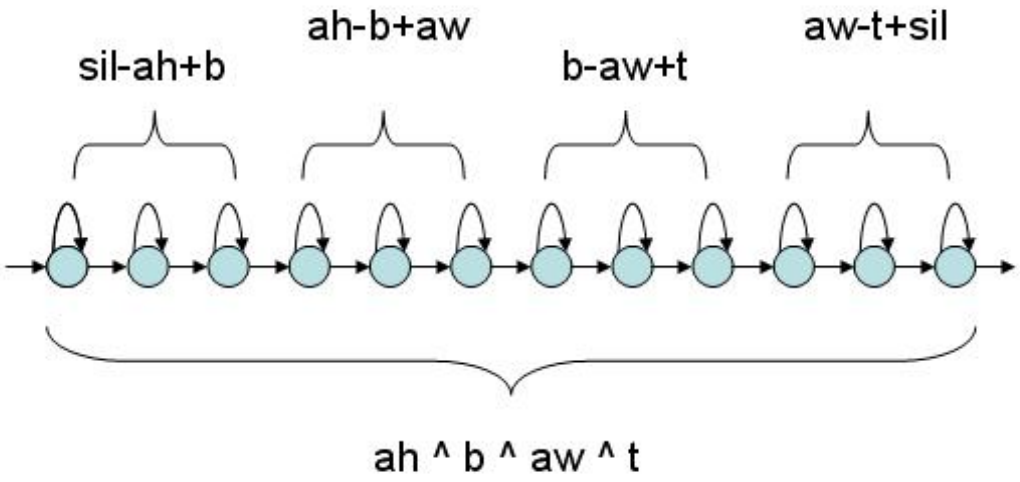

Figure 4.2: An example of the construction of a context-independent word model from word-internal triphones.

"ah^b^aw^t". Four triphone models are needed for this construction, they are "sil-ah+b", "ah-b+aw", "b-aw+t" and "aw-t+sil". Now we have a word model "ABOUT" of which its states are tied according to the rules at the phone level. Please note that the recognition performance of CI word models constructed in this way should be the same as that of the original word-internal triphones. They are exactly the same in nature (the distribution in states and the transition probabilities). They are only different in representation.

Next, we hope to reduce the number of tri-units while not sacrificing the context dependency modeling. Assume we are working on a task of which the vocabulary size is 5K. The number of tri-units will be astronomical. Inspired by the work [23], we propose context-dependent fragmented (whole) word models to be the platform to implement phone deletions.

4.5 Context-dependent Fragmented Word Models (CD-FWM)

A context-independent (CI) word model may be easily constructed from word-internal triphones as shown in Fig. 4.2 for the word "ABOUT". However, modeling contextual

word models is not easy, and a naive approach of "tri-word modeling" is infeasible even for a modest task with a few hundred words in its vocabulary.

Table 4.1: An example of showing how fragmented word models can reduce the number ot units using the word "CONSIDER" (where '?' means any phone).

Non-fragmented Version	
CI mono-unit	k^ah^n^s^ih^d^er
CD tri-unit	?-k^ah^n^s^ih^d^er+?
#units	40 x 5k x 40= 8M
Three-segment Version	
CI mono-unit	k ah^n^s^ih^d er
CD tri-unit	?-k+ah^n^s^ih^d k-ah^n^s^ih^d+er ah^n^s^ih^d-er+?
#units	40 x 5k + 5k + 40 x 5k $\simeq$ 0.4M
Four-segment Version	
CD mono-unit	k ah n^s^ih^d er
CD tri-unit	?-k+ah k-ah+n^s^ih^d ah-n^s^ih^d+er n^s^ih^d-er+?
#units	5k + 5k + 40 x 5k $\simeq$ 0.2M

Following the approach of fragmented context-dependent syllable models in [23], we propose the *context-dependent fragmented word models (CD-FWM)* and split a word into three or more segments so that the center segment is not influenced by cross-word contexts. This will greatly reduce the number of possible context-dependent units. Table 4.1 shows an example of how fragmented word models can serve this purpose.

In the example, the word "CONSIDER", of which the phone transcription is [k ah n s ih d er], is used to demonstrate the way of fragmentation. Here we model context-dependency by including the previous and the following phones. In the non-fragmented version, the whole word is treated as one single unit ("k^ah^n^s^ih^d^er"), and it will expand into a group of tri-units: "?-k^ah^n^s^ih^d^er+?" where '?' means any phone. Assume we are working in a 5K-vocabulary task and there are 40 base phones, the total number of CD units will be 40 x 5K x 40 = 8M.

However, if we split the word into three segments: "k", "ah^n^s^ih^d" and "er", they will expand into three groups of tri-units: "?-k+ah^n^s^ih^d", "k-ah^n^s^ih^d+er" and "ah^n^s^ih^d-er+?". The middle group of CD units are regarded as context-dependency subword units (CD-SWUs) as they contain multiple phones. Since the combinations of the phones in the CD-SWUs are almost unique for each word, the total number of units is equal to the vocabulary size, 5K. For the first and the last

groups of CD units, since they are affected by the phones of the neighbouring words, the total number of CD units in each group will be 40 x 5K = 200K. As a result, the total number of CD units needed in a three-segment version is 200K+5K+200K $\simeq$ 0.4M. We can see that 95% of the units have been reduced.

Furthermore, if we split the word into four segments: "k", "ah", "n^s^ih^d" and "er", they will expand into four groups of tri-units: "?-k+ah", "k-ah+n^s^ih^d", "ah-n^s^ih^d+er" and "n^s^ih^d-er+?". In this case, the first group of units are no longer word-specific and they are actually normal triphones, therefore no extra unit is needed to construct for this group. The fact that the first group of units are not word-specific will not affect the implementation of phone deletion skip arcs as we are not going to allow skipping the first phones of the words. Since the combinations of the phones in the second and the third groups are almost unique for each word, the total number of units in each group is equal to the vocabulary size, 5K. For the last group of CD units, since they are affected by the first phone of the following word, the total number of CD units will be 40 x 5K = 200K. As a result, the total number of CD units needed in a four-segment version is 200K+5K+5K $\simeq$ 0.2M.

Table 4.2: Examples of context-dependent fragmented word model (where '?' means any phone).

Word	Modified Transcription
Context-dependent Fragmented Word Model	
ABOUT	ah b^aw t
?-ah+b^aw	ah-b^aw+t b^aw-t+?
CONSIDER	k ah n^s^ih^d er
?-k+ah	k-ah+n^s^ih^d ah-n^s^ih^d+er n^s^ih^d-er+?
HUNDRED	hh ah n^d^r^ah d
?-hh+ah	hh-ah+n^d^r^ah ah-n^d^r^ah+d n^d^r^ah-d+?

We can see in the example that splitting the words into four segments can reduce the total number of CD units even more than splitting the words into three segments. However, for the words with only four or five phones, splitting them into four segments will result in SWUs with only one or two phones and the SWUs may not be unique for each word. In order to balance the number of total units and the uniqueness of the SWUs, we design the fragmentation scheme in a way that the number of segments depends on the word length L, which is defined as the number of phones in its canonical

pronunciation, as follows:

- $L \leq 3$: the word is represented by the original cross-word triphones instead of a word model, and no phone deletions are allowed.
- $L = 4$ or 5: the word is split into three segments with the first and the last segment consisting of a single phone. Table 4.2 gives an example of a 3-segment CD-FWM for the word "ABOUT".
- $L \geq 6$: the word is split into four segments with the first two segments and the last segment consisting of a single phone. Table 4.2 gives an example of a 4-segment CD-FWM for the words "CONSIDER" and "HUNDRED".

Thus, in a CD-FWM, there are actually both CD phone units and CD subword units (SWU). In a 3-segment CD-FWM, both the first and the last segments are affected by cross-word contexts, and they are not the conventional triphones: the right context of the first segment, and the left context of the last segment is the center subword segment. (We call them CD phones as they are not the conventional triphones.) A '?' in the segment means it can be any phone, therefore they are a group of N units, where N is the number of monophones. The middle segment is the CD subword unit. In a 4-segment CD-FWM, only the first segment is a cross-word triphone, the remaining three segments are similar to a 3-segment CD-FWM but now only the last segment is affected by cross-word contexts. Conventional triphones are shared by many words and they contain global information. Skip arcs are not placed on these triphones as we only want to capture word-specific phone deletion behavior.

The important point is that for words with $L \geq 4$, the center SWU is almost unique for each word. For words being split into three or four segments, the number of acoustic units increases by $O(nV)$, where n is the number of phones and V is the size of the vocabulary. As a consequence, the overall number of acoustic units only increases by $O(nV)$ instead of $O(V^3)$ if "tri-words" are used.

4.5.1 Practical Implementation of CD-FWM

In the practical implementation of CD-FWM by HTK, right now we cannot skip two successive phones within an SWU. The reason is that an SWU is represented by an

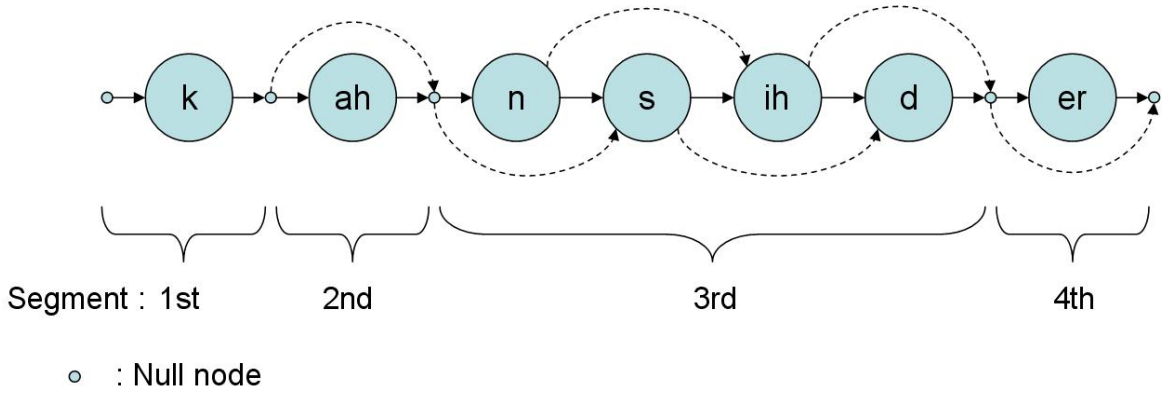

Figure 4.3: An example of adding skip arcs to allow phone deletions in the actual implementation of context-dependent fragmented word models (CD-FWM).

HMM unit, and there are no null nodes inside an HMM unit in HTK. To skip the first phone of an SWU, a skip arc is constructed from the previous null node to the first state of its second phone. To skip a middle phone in an SWU, a skip arc is added to jump from the last state of its previous phone to the first state of its following phone. To skip the last phone of an SWU, a skip arc is added to jump from the last state of its previous phone to the following null node. Fig. 4.3 shows an example with the word "CONSIDER". We further did not allow skipping the first phone in a CD-FWM. As mentioned above in Greenberg's findings that syllable onsets are well preserved. It is our conjecture in this thesis that the first phone of the word is very unlikely to be deleted.

CHAPTER 5

EXPERIMENTAL EVALUATION

The proposed phone deletion modeling with CD-FWM was evaluated first on read speech corpus then on conversational speech corpus.

5.1 Experiment on Read Speech

5.1.1 Data Setup: Wall Street Journal

The DARPA Wall Street Journal speech corpora (WSJ) consist of read speech with texts drawn from Wall Street Journal news. They were built in 1991 to support research on large-vocabulary Continuous Speech Recognition (CSR) systems. The first two WSJ corpora are often known as WSJ0 and WSJ1. They were collected at the Massachusetts Institute of Technology Laboratory for Computer Science (MIT-LCS), SRI International, and Texas Instruments (TI) in 1991-1993.

We designed the training set, development set and evaluation set as follows (each set is associated with a set ID for ease of reference):

- Training Set (si_tr_s): This is the standard speaker-independent SI-284 WSJ training set. It consists of 8,720 WSJ0 utterances from 101 WSJ0 speakers and 38,275 WSJ1 utterances from 201 WSJ1 speakers. Thus, there are a total of about 44 hours of read speech in 46,995 training utterances from 302 speakers. There are about 3-5 secs of silence at the beginning and the end of each recording. These silence frames were stripped off using an endpoint detection algorithm. A model was first trained on the raw training utterances. This pre-trained model was then used to generate forced alignments of the training data to locate the silence segments so that they can be stripped off.

  Table 5.1 shows the coverage of words of various phone lengths in the training set.

Table 5.1: Coverage of words of various phone lengths in the lexicon and word tokens of WSJ training set.

Word Length	Lexicon	Word Tokens
$L = 1$	11 (<0.1%)	21,760 (2.6%)
$L = 2$	216 (1.6%)	207,219 (25.1%)
$L = 3$	1091 (7.9%)	177,559 (21.5%)
$L = 4$	1881 (13.7%)	108,149 (13.1%)
$L = 5$	2426 (17.7%)	86,804 (10.5%)
$L = 6$	2347 (17.1%)	79,730 (9.6%)
$L = 7$	1943 (14.2%)	54,960 (6.7%)
$L = 8$	1506 (11%)	40,074 (4.8%)
$L = 9$	1026 (7.5%)	24,267 (2.9%)
$L \geq 10$	1278 (9.3%)	25,779 (3.1%)

Table 5.2: Information of various WSJ data sets.

Data Set	#Speakers	#Utterances	Vocab Size
train (si_tr_s)	302	46,995	13,725
dev1 (si_et_05)	8	330	1,270
dev2 (si_dt_05)	10	496	1,842
eval (si_et_h2)	10	205	998

- Development Set 1 (si_et_05): This is the standard Nov'92 5K non-verbalized WSJ benchmark test set. It consists of 330 utterances from 8 speakers (5 male and 3 female speakers), each with about 40 utterances. In this thesis, this set is used as one of the development set to tune the number of Gaussian components and tied states of the models.

- Development Set 2 (si_dt_05): This is the WSJ1 5K development set. The utterances containing out-of-vocabulary (OOV) words were removed. There are 496 utterances from 10 speakers in this set. We used this development set to tune the decoding parameters.

- Evaluation Set (si_et_h2): This set is extracted from the standard Nov'93 5K non-verbalized WSJ read speech HUB2 evaluation set. The goal of HUB2 evaluation was to improve basic speaker independent performance on clean data. The utterances containing OOV words were removed and there are 205 utterances from 10 speakers in this set.

A summary of these data sets is shown in Table 5.2.

5.1.2 Experimental Setup

The proposed method of explicit modeling of phone deletion using CD-FWM was evaluated first on the read speech WSJ corpus and then on the conversational speech Switchboard corpus. In order to evaluate the effectiveness of our method, the following setup was repeatedly used in the read speech experiments:

- Feature Extraction: The traditional 39-dimensional Mel Frequency Cepstral Coefficient (MFCC) [35] vectors were extracted at every 10ms over a window of 25ms. The 39 dimensions consist of 12 MFCCs and the normalized log energy as well as their first and second order derivatives.

- Dictionary: The Carnegie Mellon University (CMU) Pronouncing Dictionary version 0.7a [34] was used. It is a machine-readable pronunciation dictionary for North American English that contains over 125,000 frequently used words with their phonetic transcriptions. Many of the words have multiple pronunciation entries. The phone set contains 39 phones.

- Language Model: The standard WSJ '87-89 baseline bigram-backoff [16] language model was used in the experiment. It contains all the words in the test set lexicon with no verbal punctuation.

- Decoding: The recognition was performed using the HTK program HVite [33] with a beam search threshold of 500. HVite is a general-purpose Viterbi word recognizer. It will match a test utterance against a network of acoustic HMMs and outputs its words.

5.1.3 Training of the Baseline Cross-word Triphone Models

The SI baseline model consists of 62,402 virtual triphones and 17,107 real triphones based[1] based on 39 base phones. It was trained on the si_tr_s set. Each triphone model is a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of at most 16 components per state, and there are totally 5,864 tied states. In addition, there are a 1-state short pause model and a 3-state silence model.

Table 5.3: Recognition performance on the Nov'93 Hub2 5K evaluation task. All models have 12,202 tied states. The values of the grammar factor and insertion penalty are 13 and -10 respectively. The numbers in the brackets are the number of virtual units. (SWU = Sub-Word Units, PD = Phone Deletion)

Model	#CD Phones	#SWUs	#Skip arcs	Word Acc.
cross-word triphones	17,107 (62,400)	0	0	91.53%
CD-FWM for $L \geq 6$:				
without PD	39,763 (419,674)	7,256 (9,857)	0	91.55%
with PD	39,763 (419,674)	7,256 (9,857)	58,833 (404,562)	92.30%
CD-FWM for $L \geq 4$:				
without PD	58,581 (705,142)	11,075 (14,657)	0	91.58%
with PD	58,581 (705,142)	11,075 (14,657)	79,917 (542,877)	92.40%

5.1.4 Training of Context-dependent Fragmented Word Models (CD-FWM)

CD-FWM were derived from the baseline cross-word triphones as follows:

[1]By the default setting of our training tool(HERest), real triphones got at least 3 training samples.

STEP 1 : The canonical pronunciation of each word in the dictionary was modified: the original phonetic representation was replaced by the corresponding FWM segments. Note that the number of segments in the FWM of a word depends on its length as described in Section 4.5. The number of cross-word triphones, additional CD phones, and new CD subword units (SWU) in the CD-FWMs for different settings are shown in Table 5.3.

STEP 2 : The required models in the CD-FWM system: cross-word triphones, additional CD phones, and CD SWUs were then constructed from the cross-word triphones in the baseline system. At this point, the two systems are essentially the same — with the same set of tied states (and, of course, the same state-tying structure) — and have the same recognition performance.

STEP 3 : Skip arcs were added to the additional CD phones and CD SWUs to allow deletion of phones according to the rules described in Section 4.5.

STEP 4 : The new CD-FWMs with skip arcs were re-trained for four EM iterations.

As a sanity check for the efficacy of phone deletions, we also re-trained the models constructed from STEP 2 without adding the phone deletion skip arcs for four EM iterations in another experiment. Notice that although the underlying tied states in CD-FWMs are the same as those in the baseline cross-word triphones that derive them, due to the SWUs (which are represented by the center segments in the FWMs), after re-training the acoustic models that involve those center segments (e.g., "?-ah+b^aw" in Table 4.2) will have their own state transitions different from those in the original triphones, and they are almost word-dependent (because only a few words will share these units which have a context spanning over more than three phones). The state distributions might also be different after re-training.

5.1.5 Results and Discussion

The recognition performance of the cross-word triphone baseline and the various CD-FWM systems[2] are shown in Table 5.3. We first carried out the experiment of using CD-FWM only for words with $L \geq 6$. It means only words with $L \geq 6$ are represented

[2]The significant tests of the WSJ experiments are summarized in Table B.1.

by CD-FWMs with the addition of phone deletion skip arcs and the rest of the words are represented by normal triphones without implementation of phone deletion modeling. Then we extended the coverage of CD-FWM so that the words with L = 4 or 5 were also represented by CD-FWMs.

It can be seen that without the addition of phone deletion skip arcs, re-trained CD-FWMs give almost no recognition improvement over the baseline triphone system[3]. Although the new CD phones and CD SWUs in CD-FWMs may model some word-specific information through the re-estimated state transitions in those models, since state transitions are much less important than the state distributions in an HMM, the improvement is expected to be small, if any.

The biggest gain comes from the addition of skip arcs to allow phone deletions for words with $L \geq 4$; it is 0.87% absolute (10.27% relative). On the other hand, most of the gain comes from modeling phone deletion for words with $L \geq 6$ while further modeling phone deletion for words with $L = 4$ or 5 only gives an additional 0.1% gain.

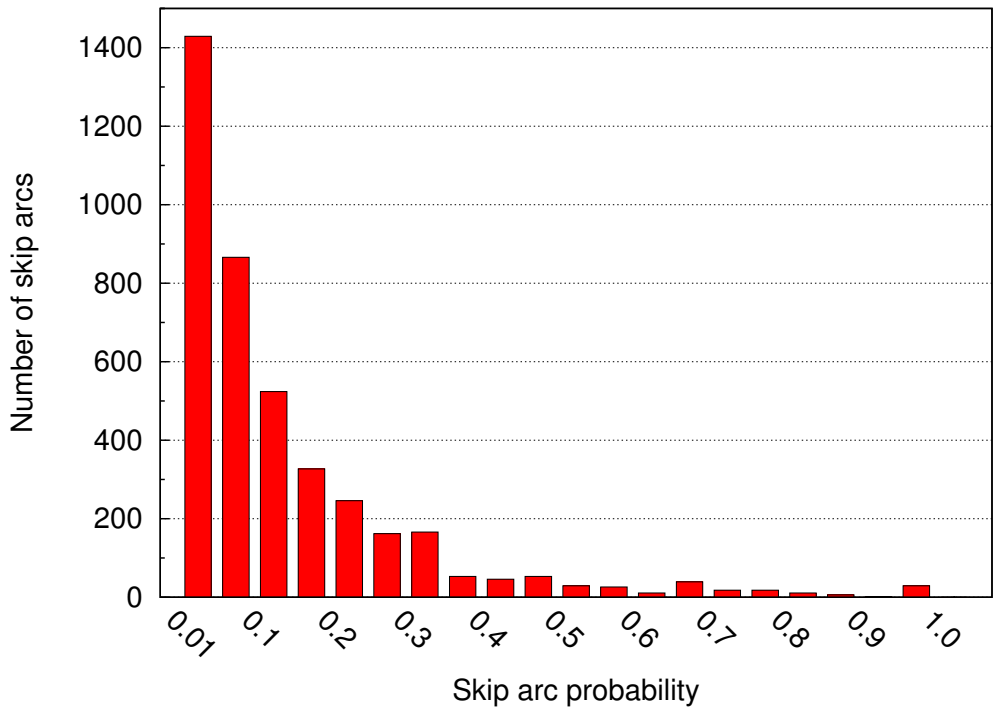

Figure 5.1: Distribution of phone deletion probabilities for the CD-FWM system with $L \geq 4$. Those with a probability less than 0.01 are removed from this plot.

[3]We had empirically verified, as expected, that CD-FWMs gave the same recognition performance as the baseline triphones which derived them if they were not re-trained.

5.1.6 Analysis of the Skip Arc Probabilities

We looked at the estimated probabilities of the skip arcs of the CD-FWM system that implemented phone deletions for words with four or more phones. A distribution of their probabilities is plotted in Fig. 5.1. Out of the total 79,917 phone deletion skip arcs, only 4,060 (5%) of them have a probability greater than 0.01 and they should have captured the phone deletion behavior of the training data (the rest are not included in the plot of Fig. 5.1). The remaining skip arcs got a very small probability after training and it would not affect the recognition performance even if they were set to zero. Thus, the proposed phone deletion modeling method does not give a load to the model complexity; yet, the recognition improvement is relatively substantial.

5.1.7 Experiment with Single-pronunciation Dictionary

As mentioned in Chapter 3, pronunciation variation modeling can be done at different level simultaneously. The experiments in previous section was actually using a dictionary with multiple pronunciation variants. It means that pronunciation modeling at lexicon level was already applied.

Table 5.4: Recognition performance on the Nov'93 Hub2 5K evaluation task with the use of the single-pronunciation dictionary. The values of the grammar factor and insertion penalty are 13 and -10 respectively. The numbers in the brackets are the number of virtual units. (SWU = Sub-Word Units, PD = Phone Deletion)

Model	#CD Phones	#SWUs	#Skip arcs	Word Acc.
cross-word triphones	17,107 (62,402)	0	0	91.20%
CD-FWM for $L \geq 4$: with PD	56,134 (705,142)	10,101 (14,657)	79,917 (542,877)	91.69%

In order to investigate the effectiveness of our proposed method in the absense of other pronunciation modeling methods, we repeated the experiment with a dictionary without multiple pronunciation variants (that means each word has only one pronunciation entry). The single-pronunciation dictionary was modified from the CMU dictionary[4] in a way that all the alternative pronunciations of the words with $L \geq 4$

[4]In the CMU dictionary, the ratio between lexicon and pronunciation variants for words with $L \geq 4$ is 1:1.25.

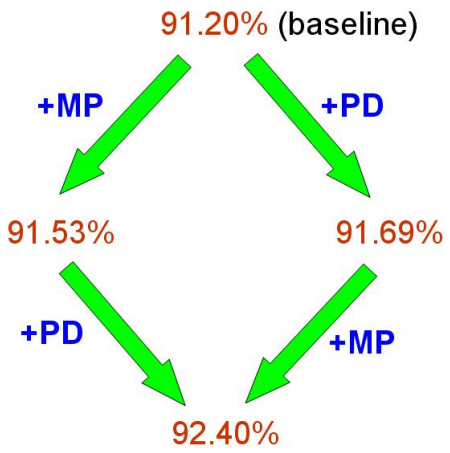

Figure 5.2: Recognition performance of PD vs. MP on the Nov'93 Hub2 5K evaluation task. PD stands for using our proposed phone deletion modeling method and MP stands for using the multiple-pronunciation dictionary. The baseline result is obtained using the single-pronunciation dictionary.

were removed. (The multiple pronunciation variants for the words with $L \leq 3$ were kept). Table 5.4 shows the results of using the single-pronunciation dictionary. and Fig 5.2 summarizes the results of both pronunciation modeling methods.

When only one pronunciation modeling method is used, our method gives a larger gain than using the multiple-pronunciation dictionary (absolute 0.49% vs. 0.33%). On the other hand, the gain when both pronunciation modeling methods are used simultaneously is even greater than the sum of gains when the methods are used alone. It is because our method relies on the pronunciation variants in the dictionary to generate the phone deleted variants. When some pronunciation variants are added, the corresponding phone deleted version of those variants can be modeled.

Therefore, it is suggested to implement our proposed method with the use of a multiple-pronunciation dictionary and the improvement gained by our proposed method is additive to that gained by existing pronunciation variation modeling at lexicon level.

5.2 Experiment on Conversational Speech

5.2.1 Data Setup: SVitchboard

SVitchboard [29] is a conversational telephone speech data set defined using subsets of the Switchboard-1 corpus [31]. It defines several small vocabulary data set ranging from 10 words to 500 words of which each task has a completely closed vocabulary. Each data set is further divided into 5 partitions so that they can be used as the training set, development set and evaluation set. The speakers of each partition do not overlap with speakers of other partitions. In this thesis, we use the SVitchboard 500-word subtask one for the evaluation on conversational speech. The training set, development set and evaluation set are described as follows:

- Training set: Partition A, B and C of the SVitchboard 500-word tasks were used as the training data. There are in total 13,597 utterances from 324 speakers. The duration of speech in this set is 3.69 hours in total.
- Development set: Partition D of the SVitchboard 500-word tasks was used as the development data. It consists of 4,871 utterances from 107 speakers. The duration of speech in this set is 1.32 hours in total.
- Evaluation set: Partition E of the SVitchboard 500-word tasks was used as the testing data. It consists of 5,202 utterances from 107 speakers. The duration of speech in this set is 1.43 hours in total.

Table 5.5: Information of various data sets in the SVitchboard 500-word subtask one.

Data Set	#Speakers	#Utterances	#Word Tokens	Duration of speech (hours)
training	324	13,597	51,324	3.69
development	107	4,871	18,075	1.32
evaluation	107	5,202	20,021	1.43

A summary of these data sets is shown in Table 5.5.

5.2.2 Experimental Setup

The following setup was used in the conversational speech experiments:

- Feature Extraction: The 39-dimensional Perceptual Linear Prediction (PLP) [36] vectors were extracted at every 10ms over a window of 25ms. The 39 dimensions consists of 12 PLP coefficients and the normalized log energy as well as their first and second order derivative.
- Dictionary: The lexicon produced by the Switchboard Transcription Project [30] was used. The number of base phones is originally 42 but it is reduced to 39 by converting [ax] to [ah]; [el] to [ah l] and [en] to [ah n]. This was done to reduce the number of triphones. Now the base phone set is exactly the same as the one in read speech experiment.
- Language Model: A bigram-backoff language model was constructed using the language modeling toolkit SRILM [32]. Only the training data set was used to train the LM.
- Decoding: Recognition was performed using the HTK program HVite [33] with a beam search threshold of 200.

5.2.3 Training of the Baseline Cross-word Triphone Models

The baseline triphone model consists of 62,402 virtual triphones and 4,558 real triphones based on 39 base phones. Each triphone model is a strictly left-to-right 3-state continuous-density hidden Markov model, with a Gaussian mixture density of at most 16 components per state, and there are totally 660 tied states. The model size was chosen to maximize development set accuracy. In addition, there are a 1-state short pause model and a 3-state silence model.

The training procedures of CD-FWM were the same as the one in the read speech experiment.

Table 5.6: Recognition performance on the SVitchboard 500-word E set. All models have 660 tied states. The values of the grammar factor and insertion penalty are 13 and -20 respectively. The numbers in the brackets are the number of virtual units. (SWU = Sub-Word Units, PD = Phone Deletion)

Model	#CD Phones	#SWUs	#Skip arcs	Word Acc.
cross-word triphones	4,558 (62,402)	0	0	44.17%
CD-FWM for $L \geq 6$:				
without PD	4,631 (65,599)	79 (79)	0	44.18%
with PD	4,631 (65,599)	79 (79)	567 (3,513)	44.23%
CD-FWM for $L \geq 4$:				
without PD	4,908 (78,679)	249 (250)	0	44.33%
with PD	4,908 (78,679)	249 (250)	1,549 (10,427)	44.43%

5.2.4 Results

From the recognition performance of various system in Table 5.6, we can see that the addition of phone deletion skip arcs gives only small recognition improvement (absolute 0.1%) in the conversational speech task and the results are all statistically insignificant[5]. In the following, we would like to investigate the modest improvement by studying the coverage of long words in the conversational speech corpus. Furthermore, we would like to investigate the confusions induced by phone deletion modeling.

5.2.5 Analysis of Word Tokens Coverage

Table 5.7: Coverage of words of various phone lengths in the lexicon and word tokens of the training set of the SVitchbaord 500-word subtask one.

Word Length	Lexicon	Word Tokens
$L = 1$	5 (1%)	5480 (11%)
$L = 2$	66 (13%)	16,312 (32%)
$L = 3$	181 (36%)	19,112 (37%)
$L = 4$	117 (23%)	6,349 (12%)
$L = 5$	55 (11%)	2,261 (4%)
$L = 6$	32 (6%)	740 (1%)
$L = 7$	23 (5%)	463 (1%)
$L \geq 8$	21 (4%)	607 (1%)

[5]The significant tests of the SVitchboard 500-word subtask one are summarized in Table B.2.

In [17], it has been shown that words differ greatly in terms of their frequency of occurrence in spoken English. The most common words occur far more frequently than the least, and most of them are short words with few phones. A frequency analysis of the lexicon and word tokens[6] of the training set of the SVitchboard 500-word subtask one in Table 5.7 illustrates the magnitude of this effect. The short words ($L \leq 3$) account for approximately 80% of all the word tokens in the training data.

Table 5.8: Comparison of word tokens coverage of various lengths in read speech and conversational speech test set.

Word Length	Hub2 Eval Set	SVitchboard 500-word E Set
$L \geq 6$	942 (26%)	708 (3.5%)
$L \geq 4$	1,817 (50%)	4,130 (20.6%)
$L \geq 1$	3,647 (100%)	20,021 (100%)

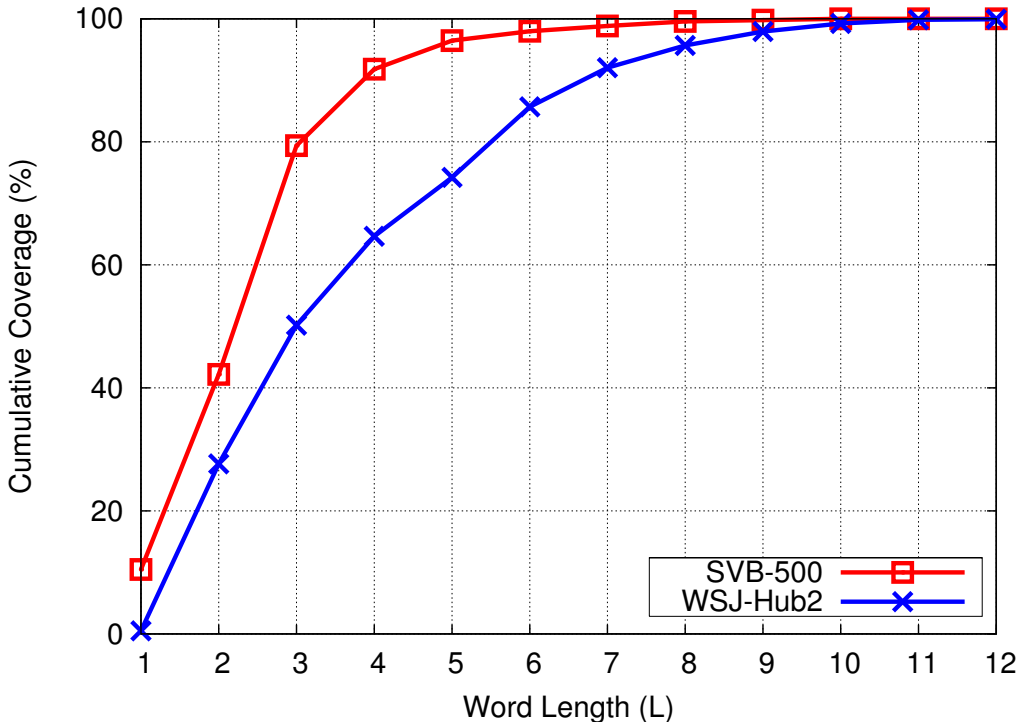

Figure 5.3: Cumulative coverage of word tokens as a function of word length in the WSJ Hub2 set and the SVitchboard 500-word E set.

From the results in the read speech experiment(Table 5.3), most of the gain comes from modeling phone deletion for words with $L \geq 6$ while further modeling phone deletion for words with $L = 4$ or 5 only gives an additional 0.1% gain. As illustrated in

[6]In this thesis, the term "word token" means multiple copies of the same word are counted repeatedly.

Table 5.8 and Fig. 5.3, the coverage of words with $L \geq 6$ in the SVitchboard 500-word E set is much smaller than in the WSJ-Hub2 set (26% vs. 3.5%). As a result, the improvement in the conversational speech experiment may not be as obvious as in the read speech experiment.

5.2.6 Analysis of Confusions Induced by Phone Deletion Modeling

In this section, we would like to investigate the confusions induced by our proposed phone deletion modeling method. We carried out the analysis using the CD-FWM system with $L \geq 4$ in the SVitchboard 500-word task.

Let us first denote the CD-FWM without phone deletion skip arcs as Model NP and CD-FWM with phone deletion skip arcs as Model P. We then investigate the confusion between the two models, NP and P, by doing the following analysis:

- For each test utterance, the recognized sentence produced by each model, NP or P, is aligned with the reference transcription.
- Thus, for each word in the reference transcriptions, we may know if each of the two models recognizes it correctly: wrong recognitions are caused by substitution or deletion errors; insertion errors are not taken into account in this analysis.
- Each word in the reference transcriptions may be classified into one of the following four categories:
    1. correctly recognized by both Model NP and P.
    2. correctly recognized by Model NP but wrongly recognized by Model P.
    3. wrongly recognized by Model NP but correctly recognized by Model P.
    4. wrongly recognized by both Model NP and P.

In Table 5.9, there are 342 words which are correctly recognized by model P but wrongly recognized by model NP and 108 of them are recognized as phone deleted. (The rest of the words have a chance that they are corrected due to a cascade effect of their neighbouring words getting phone deleted.) For example, the word "PERSONALLY"

Table 5.9: Breakdown of the number of words according to the recognition result of two models, NP and P, in the SVitchboard 500-word subtask one.

CD-FWM Without Phone Deletion	CD-FWM With Phone Deletion	
	Correct	Wrong
Correct	9,961	309
Wrong	342	9,407

is correctly recognized by model P with [n] deleted while it is wrongly recognized as "PERSON" by model NP. The word "USED" is correctly recognized by model P with [d] at the end deleted while it is wrongly recognized as "USE" by model NP.

On the other hand, there are 309 words which are wrongly recognized by model P but correctly recognized by model NP and 82 of them are recognized as phone deleted. These words are confused by adding phone deletion skip arcs. For example, the word "THING" is correctly recognized by model NP while it is wrongly recognized as "THINK" by model P with [k] at the end deleted. Another example is that the word "SOME" ([s ah m]) is correctly recognized by model NP while it is wrongly recognized as "SOMETHING" ([s ah m th ih ng]) by model P with [th] in the middle and [ng] at the end deleted. With two phones deleted, "SOMETHING" only differ from "SOME" in having a [ih] at the tail and the system wrongly recognized the signal following "SOME" to be [ih]. Therefore, "SOME" is confused with "SOMETHING" in this particular example.

5.3 Phone Deletion Modeling on Context-independent System

Let us take the word model of "ABOUT" ([ah b aw t]) as an example to illustrate the context-mismatch of units occurred in our proposed method. While no phones is deleted, the state distribution sequence for "ABOUT" is "sil-ah+b, ah-b+aw, b-aw+t, aw-t+sil" (Fig. 4.2). If [aw] is deleted, the state distribution sequence becomes "sil-ah+b, ah-b+aw, aw-t+sil" (Fig. 5.4). There is a context-mismatch[7] between the two units, "ah-b+aw" and "aw-t+sil".

[7]In this example, the context-mismatch does not hold if the distributions of "ah-b+aw" is exactly the same as "ah-b+t" and the distributions of "aw-t+sil" is exactly the same as "b-t+sil".

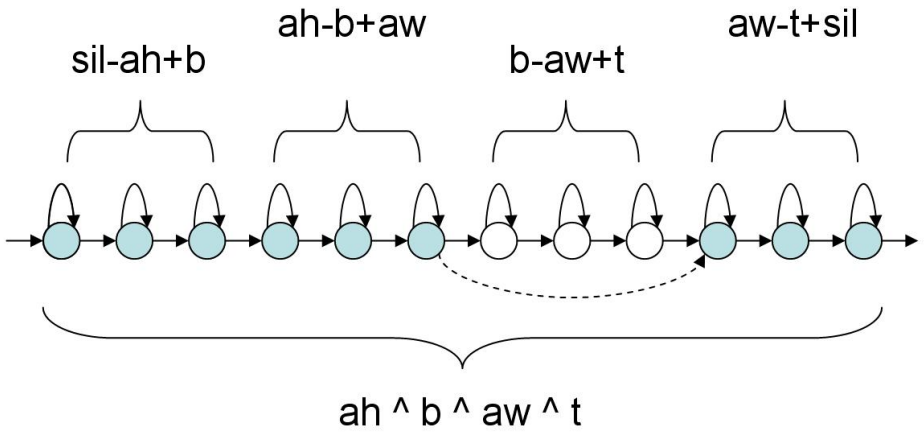

Figure 5.4: The state sequence of the word model of "ABOUT" while [aw] is deleted.

In order to investigate the effect of the context-mismatch to our proposed phone deletion modeling method, we implemented our method on a context-independent system. This experiment was done using the SVitchboard 500-word subtask one. The experiment settings such as data set, feature extraction, dictionary, LM were the same as the the settings in Section 5.2.2, and only now the baseline acoustic model was changed to 39 monophone HMMs. Each monophone model is a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of at most 64 components per state, and there are totally 120 states including the 3-state silence model.

It is not necessary to fragment the word models in a context-independent system as there is no tri-unit expansion. As a result, context-independent whole word models (CI-WWMs) were constructed from the monophones in the baseline system. Skip arcs were added to the CI-WWMs in the same way as how the skip arcs were added to the CD SWU according to the rules described in Section 4.5. The new CI-WWMs with skip arcs were then re-trained for four EM iterations. In this experiment, words with $L \geq 4$ were represented by CI-WWMs and the remaining words were represented by monophone models.

5.3.1 Results and Discussion

The recognition performance of the monophone baseline and the various CI-WWM systems are shown in Table 5.10. It can be seen that the addition of phone deletion

Table 5.10: Recognition performance on the SVitchboard 500-word E set. All models have 120 tied states. The values of the grammar factor and insertion penalty are 10 and -10 respectively. (WU = Word Units, CI-WWM = Context-independent Whole Word Model, PD = Phone Deletion)

Model	#WUs	#Skip arcs	Word Acc.
monophones	0	0	34.08%
CI-WWM for $L \geq 4$:			
without PD	251	0	34.08%
with PD	251	1,293	33.44%

Table 5.11: Breakdown of the number of words according to the recognition result of two models, CI-WWM without phone deletion modeling and CI-WWM with phone deletion modeling, in the SVitchboard 500-word subtask one.

CI-WWM Without Phone Deletion	CI-WWM With Phone Deletion	
	Correct	Wrong
Correct	7,355	890
Wrong	793	10,983

skip arcs degrades the recognition performance.

We would like to look at the number of words which are corrected or confused by adding phone deletion skip in the CI system. Table 5.11 was generated with the steps described in Section 5.2.6. In Table 5.11, there are 890 words confused by adding phone deletion skip arcs and 614 of them are recognized as phone deleted although the phone deleted words are not correct. On the other hand, there are 793 words that are corrected by adding phone deletion skip arcs and all of them are recognized as phone deleted. The number of confused words in the context-independent system is much larger than in the context-dependent system. The reason for the greater confusion in the context-independent system could be explained by the following analysis.

5.3.2 Analysis of Confusability of Phone Deletion Modeling in Context-independent System and Context-dependent System

Let us take the two words, "USE"([y uw z]) and "USED"([y uw z d]), as an example. Their word units are "y^uw^z" and "y^uw^z^d" respectively. In a context-independent system, the state distributions of these two word units are tied to corresponding mono-

phones and their state distribution sequences are " y, uw, z" and "y, uw, z, d" respectively. If [d] is allowed to be deleted from "y^uw^z^d", the state distribution sequences of these two word models will be the same. In this case, the only difference between the two models are their word-specific state transition probabilities. Since state transitions are much less important than the state distributions in an HMM, the discriminative power between these two word models is expected to be small after the addition of phone deletion skip arcs.

In contrast, the state distributions of the two word units are tied to corresponding triphones in a context-dependent system. The state distribution sequence for "USE" is "i-y+uw, y-uw+z, uw-z+f" and for "USED" is "i-y+uw, y-uw+z, uw-z+d, z-d+f". (Here we assume the phones before and followed by are [i] and [f].) If "z-d+f" is allowed to be deleted, these two state distribution sequences still differ from a "uw-z+f" to a "uw-z+d". Still, the word-specific state transition probabilities of the two models are different. Therefore, the discriminative power between these two word models in this case should be greater than in the context-independent case. This explains why the confusability of modeling phone deletions in context-independent system is greater than in context-dependent system.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This thesis investigates the effectiveness of modeling phone deletion explicitly for automatic speech recognition. First, we present our motivation: phone deletions frequently occur in human speech and it cannot be modeled well by traditional triphone training. Second, we model phone deletions by adding skip arcs and the practical reasons why we propose the Context-dependent Fragmented Word Models (CD-FWM) are explained. Finally, we conduct experiments to demonstrate the improvement resulted from phone deletion modeling and carry out relevent analyses. In the following sections, I will summarize the contributions of this thesis to the ASR community and suggest possible future work.

6.1 Conclusion

Although it is generally expected that phone deletions are more common in spontaneous speech, we hypothesize in this thesis that they may also occur in read speech. Our hypothesis is supported by the recognition improvement gained from phone deletion modeling and the analysis of the skip arc probabilities in the read speech experiment. Moreover, only 5% of the skip arcs have a probability greater than the threshold; this suggests that phone deletion is a word-specific phenomenon rather than a random process.

Right now, the effectiveness of our method is limited in conversational speech recognition as we are modeling phone deletion only on words with more than three phones. If we can relax the limitation on the number of total units, whole word models can be applied on the short words and the performance of our method can be verified more effectively in conversational speech. On the other hand, it is expected that phone deletion modeling on short words may lead to an further increase of confusion. Therefore, some discriminative training method should be applied to reduce the confusion while phone deletion is modeled on short words.

6.2 Contributions

In this thesis, the CD-FWM with the addition of phone deletion skip arcs is proposed. It improves the acoustic model by explicit modeling of phone deletions. Our experiments show that the proposed method improves the word recognition accuracy from the baseline 91.53% (given by cross-word triphones) to 92.40% in read speech data. Furthermore, we have verified that the improvement gained by our method is additive to the gain obtained by existing pronunciation variation modeling at the lexicon level using multiple pronunciations.

We analyse the skip arcs probabilities and find that only a small proportion of skip arcs has a probability larger than 0.01. This suggests that our method of modeling phone deletion does not increase the model complexity. From a pronunciation variation modeling perspective, the pronunciation weights are captured naturally by the skip arc probabilities in the acoustic model. For the words with no phone deletion, the skip arcs automatically vanish (probabilities become zero) after re-estimation.

The word tokens coverage of the SVitchboard 500-word task are analysed. We compare the frequency distributions of word tokens of different lengths in WSJ and SVitchboard 500-word task. This analysis helps explain why the gain of our proposed method is modest in conversational speech.

The source of confusions induced by phone deletion modeling is investigated. Phone deletion modeling was implemented on a context-independent system and we show that the context-mismatch of units occurred in our proposed method actually help reduce the confusion.

I have published three papers [37, 38, 39] throughout my MPhil study.

6.3 Future Work

Although CD-FWM system with the addition of phone deletion skip arcs gives substantial recognition improvement over the baseline triphone system in the read speech experiment, the gain in conversational speech is modest. From our analysis, the modest gain in conversational speech experiment is attributed to the small coverage of long words in conversational speech corpus.

Currently, CD-FWM is not applied on the words with three phones as there is only one phone left in the center subword unit after the fragmentation. The fragmented units will mix with other global triphones and would become no longer word-specific. According to the findings in [17], phone deletions indeed occur in short words like "AND" ([ae n d]). Moreover, Table 5.7 shows that words with three phones ($L = 3$) is the largest group in both lexicon and word tokens of conversational speech. Future work will deal with this limitation and capture the phone deletion information in short words.

On the other hand, it is worth to investigate which set of skip arcs can lead to largest gain. From our analysis, it can be seen that some skip arcs fix old errors but some produce new errors. If those skip arcs which lead to confusions more than improvement are removed, the recognition performance of the system can be further improved. Future work will deal with this as well.

REFERENCES

[1] Strik, H. and Cucchiarini C., "Modeling pronunciation variation for ASR: Overview and comparison of methods," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998.

[2] J. Kessens and M. Wester, "Improving recognition performance by modelling pronunciation variation," in *Proceedings of the CLS opening Academic Year '97-'98*, 1997, pp. 1–20.

[3] T. Holter and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998 .

[4] D. Torre, L. Villarrubia, L. Hernandez and J.M. Elvira, "Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Munich, vol.2, pp. 1463–1466, 1997.

[5] M. Riley, W. Byrne, M. Finke, S. Khudanpur and A.Ljolje, "Stochastic pronunciation modeling from hand-labelled phonetic corpora," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998.

[6] N. Cremelie and J.P. Martens, "On the use of pronunciation rules for improved word recognition," in *Proceedings of EuroSpeech-95*, Madrid, Spain, Vol. III, pp. 1747–1750.

[7] N. Cremelie and J.P. Martens, "Automatic rule-based generation of word pronunciation networks," in *Proceedings of EuroSpeech-97*, pp. 2459–2462.

[8] T. Fukada, T. Yoshimura and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks and language," in *Proceedings of the ESCA*

Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, The Netherlands, 4-6 May, 1998.

[9] M. Adda-Decker, "Pronunciation variants across systems, languages and speaking style," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998.

[10] M. Wester, J.M. Kessens and H. Strik, "Improving the performance of a Dutch CSR by modelling pronunciation variation," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998.

[11] H. Gish and K. Ng, "Parameter trajectory models for speech recognition," in *Proceedings of IEEE International Conference on Spoken Language Processing, Philadelphia, PA*, pp. 466–469, Oct, 1996.

[12] Dominic W. Massaro, "Perceptual units in speech recognition," *Journal of Experimental Psychology*, 102(2):199–208, Oct, 1974.

[13] Dogulas O'Shaughnessy, "Speech Communication," *Addison-Wesley Publishing Company, Reading, Massachusetts*, chapter 5:pages 164–203, 1987.

[14] D. Jurafsky, W. Ward, J. P. Zhang, K. Herold, X. Y. Yu, and S. Zhang, "What kind of pronunciation variation is hard for triphones to model?," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[15] Vaibhava Goel, Peder A. Olsen; IBM T.J. Watson Research Center, USA, "Acoustic Modeling Using Exponential Families," in *Proceedings of Interspeech*, 2009.

[16] Slava M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3), 400-401, 1987.

[17] S. Greenberg, "Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 4-6 May, 1998.

[18] S. Greenberg, "Understanding speech understanding towards a unified theory of speech perception," in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, W.A. Ainsworth and S. Greenberg, Eds. 1996, pp. 1–8, Keele University, UK.

[19] A. Ganapathiraju, J. Hamaker, J. Picone, M. ordowski, and G. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no.4, pp. 358–366, May 2001.

[20] A. Sethy and S. Narayanana, "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2003, vol. I, pp. 772–775.

[21] A. Sethy, B. Ramabhadran, and S. Narayanana, "Improvements in English ASR for the MALACH project using syllable-centric models," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* , 2003.

[22] A. Hämäläinen, L. Bosch, and L. Boves, "Construction and analysis of multiple paths in syllable models," in *Proceedings of Interspeech*, 2007, pp. 882–885.

[23] K. thambiratnam and F. Seide, "Fragmented context-dependent syllable acoustic models," in *Proceedings of Interspeech*, 2008, pp. 2418–2421.

[24] Su-Lin Wu, M. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 2, pp. 987–990.

[25] Su-Lin Wu, E. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1998, vol. 2, pp. 721–724.

[26] A. Hauenstein, "Using syllable in a hybrid HMM-ANN recognition system," in *Proceedings of Eurospeech*, 1997, vol. 3, pp. 1203–1206.

[27] R. J. Jones, S. Downey, and J. S. Mason, "Continuous speech recognition using syllables," in *Proceedings of Eurospeech*, 1997, vol. 3, pp. 1171–1174.

[28] Hao Wu and Xihong Wu, "Context dependent syllable acoustic model for continuous Chinese speech recognition," in *Proceedings of Interspeech*, 2007, pp. 1713–1716.

[29] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small Vocabulary Tasks from Switchboard 1," in *Proceedings of Interspeech*, 2005.

[30] Greenberg, S., "The Switchboard Transcription Project," in *Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD*, 1997.

[31] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Mar. 1992, pp. 517–520.

[32] Andreas Stolcke, "SRILM an Extensible Language Modeling Toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing*, 2002.

[33] Steve Young et al., *The HTK Book (Version 3.4)*. University of Cambridge, 2006.

[34] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[35] Davis, S. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980, 28(4), pp. 357–366.

[36] Hynek Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," in *Journal of the Acoustical Society of America*, 1990, 87(4), pp. 1738–1752.

[37] Tom Ko and Brian Mak, "Improving Speech Recognition by Explicit Modeling of Phone Deletions," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 4858-4861, March, 2010, Dallas, Texas, USA.

[38] Brian Mak and Tom Ko, "Automatic Estimation of Decoding Parameters Using Large-Margin Iterative Linear Programming," in *Proceedings of Interspeech*, pages 1219-1222, Sept, 2009, Brighton, U.K.

[39] Brian Mak and Tom Ko, "Min-max Discriminative Training of Decoding Parameters Using Iterative Linear Programming," in *Proceedings of Interspeech*, pages 915-918, Sept, 2008, Brisbane, Australia.

APPENDIX A

PHONE SET IN THIS THESIS

Table A.1: The phone set and their examples.

Phoneme	Example	Transcription
aa	ODD	aa d
ae	AT	ae t
ah	HUT	hh ah t
ao	OUGHT	ao t
aw	COW	k aw
ay	HIDE	hh ay d
b	BE	b iy
ch	CHEESE	ch iy z
d	END	eh n d
dh	WEATHER	w eh dh er
eh	BEAR	b eh r
er	HURT	hh er t
ey	ATE	ey t
f	FREE	f r iy
g	GREEN	g r iy n
hh	HE	hh iy
ih	IT	ih t
iy	EAT	iy t
jh	JANE	jh ey n
k	KEY	k iy
l	LIGHT	l ay t
m	ME	m iy
n	SON	s ah n
ng	PING	p ih ng
ow	NO	n ow
oy	TOY	t oy
p	PIG	p ih g
r	RIGHT	r ay t
s	SEA	s iy
sh	SHE	sh iy
t	TEA	t iy
th	THETA	th ey t ah
uh	FOOT	f uh t
uw	TWO	t uw
v	VERY	v eh r iy
w	WET	w eh t
y	YET	y eh t
z	ZOO	z uw
zh	VISION	v ih zh ah n

APPENDIX B

SIGNIFICANT TESTS

In the significant tests, the cross-word triphone baseline and various CD-FWM systems are compared. The abbreviations of various systems and the tests are summarized as follows:

TRIPHONE: cross-word triphones system.

CD-FWM6-NP: CD-FWMs for $L \geq 6$ without addition of phone deletion skip arcs.

CD-FWM6-P: CD-FWMs for $L \geq 6$ with addition of phone deletion skip arcs.

CD-FWM4-NP: CD-FWMs for $L \geq 4$ without addition of phone deletion skip arcs.

CD-FWM4-P: CD-FWMs for $L \geq 4$ with addition of phone deletion skip arcs.

MP: Matched Pair Sentence Segment (Word Error) Test.

SP: Signed Paired Comparison (Speaker Word Accuracy Rate) Test.

WI: Wilcoxon Signed Rank (Speaker Word Accuracy Rate) Test.

MN: McNemar (Sentence Error) Test.

Table B.1: Significant tests of the WSJ experiments.

	CD-FWM6-NP	CD-FWM6-P	CD-FWM4-NP	CD-FWM4-P
TRIPHONE	MP: same SP: same WI: same MN: same	MP: CD-FWM6-P SP: same WI: same MN: same	MP: same SP: same WI: same MN: same	MP: CD-FWM4-P SP: same WI: same MN: same
CD-FWM6-NP		MP: CD-FWM6-P SP: same WI: same MN: same	MP: same SP: same WI: same MN: same	MP: CD-FWM4-P SP: same WI: same MN: same
CD-FWM6-P			MP: CD-FWM6-P SP: same WI: same MN: same	MP: same SP: same WI: same MN: same
CD-FWM4-NP				MP: CD-FWM4-P SP: same WI: same MN: same

Table B.2: Significant tests of the SVitchboard 500-word subtask one.

	CD-FWM6-NP	CD-FWM6-P	CD-FWM4-NP	CD-FWM4-P
TRIPHONE	MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same
CD-FWM6-NP		MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same
CD-FWM6-P			MP: same SP: same WI: same MN: same	MP: same SP: same WI: same MN: same
CD-FWM4-NP				MP: same SP: same WI: same MN: same