National Conference on Man-Machine Speech Communications (*NCMMSC6*), pp.K51-64, 20-22 Nov 2001, Shenzhen **(Invited Keynote Speech, also an accepted paper by JCST)**

MANDARIN PRONUNCIATION VARIATION MODELING

Fang Zheng (郑方) ¶*, *Zhanjiang Song* (宋战江) ¶*, *Pascale Fung*†*, *William Byrne* ‡

¶ Center of Speech Technology, State Key Lab of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University
† Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology
‡ Center for Language and Speech Processing, The Johns Hopkins University
* Weniwen Technologies Limited, Hong Kong

fzheng@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn/~fzheng/

ABSTRACT

The pronunciation variability is an important issue that must be faced with when developing practical automatic spontaneous speech recognition systems. In this paper, the factors that may affect the recognition performance are analyzed, including those specific to the Chinese language. By studying the INITIAL/FINAL [2] (IF) characteristics of Chinese language and developing the Bayesian equation, we propose the concepts of generalized INITIAL/FINAL (GIF) and generalized syllable (GS), the GIF modeling and the IF-GIF modeling, as well as the context-dependent pronunciation weighting, based on a well phonetically transcribed seed database. By using these methods, the Chinese syllable error rate (SER) was reduced by 6.3% and 4.2% compared with the GIF modeling and IF modeling respectively when the language model, such as syllable or word N-gram, is not used. The effectiveness of these methods is also proved when more data without the phonetic transcription is used to refine the acoustic model using the proposed iterative forced-alignment based transcribing (IFABT) method, achieving a 5.7% SER reduction.

1. INTRODUCTION

For carefully produced read speech the current automatic speech recognition (ASR) systems can reach word accuracy over 90% while for casual and unplanned spontaneous speech its performance drops greatly [Fosler-Lussier 1999]. The difference in performance lies mainly in the difference of pronunciation style between the read and spontaneous speech, which can be detailedly studied at two levels, the phonetic level and the linguistic level.

At the phonetic level, the casual or spontaneous speech contains much more phone change (substituted, deleted, and inserted) phenomena and sound change (nasalized, centeralized, voiced, voiceless, more rounded, syllabic, pharyngealized, and aspirated) phenomena because of variable speaking rates, moods, emotions, prosodies, co-articulations and so on, even when the speaker is tending to utter in canonical pronunciations [Decker 1999][Greenberg 1999]. Other phenomena, such as lengthening, breathing, disfluency, lip smacking, murmuring, coughing, laughing, crying, modal/exclamation, silence, and noise, will also bring difficulties to ASR systems.

At the linguistic level, there are a lot of spoken language phenomena, such as repetitions, ellipses, corrections, hesitations, and so on, resulting from the fact that people are often thinking while speaking in daily life. This makes it difficult to make full use of the statistical language model, for example the N-Gram language model.

Compared with other languages such as English, Chinese has its own characteristics. Spoken language problems are made especially severe in casual Chinese speech since most Chinese are non-native standard Chinese speakers and are with complicated dialect and accent backgrounds. Some Chinese accents or dialects such as Cantonese are different from the standard Chinese as French is different from English. As a result, there is an even severe

[1] This work was a report for the project "*Mandarin pronunciation modeling*" supported by the National Science Foundation of USA under Grant No. #IIS-9820687, and carried out in the 2000 Summer Workshop on Language and Speech Processing, Center for Language and Speech Processing, Johns Hopkins University (*http://www.clsp.jhu.edu/ws2000/*), and a report of its further research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

[2] The uppercase words INITIAL and FINAL stand for two parts of Chinese syllable while lowercase the English words.

pronunciation variation due to the influence of speakers' native pronunciations. Additionally, the homonym issue, the homograph issue, the retroflex issue, the tone change issue, the Chinese syllable's short INITIAL/FINAL structure, and so on, that are specific to Chinese [Zheng 1999] make the spontaneous Chinese ASR more difficult than other languages.

To find a solution to the spontaneous speech and spoken language recognition, the following major aspects are mostly focused on in recent research.

1. Choosing the speech recognition unit (SRU) set

In the acoustic modeling stage, an SRU set should be well defined so that it can be used to well describe the phone changes and the sound changes, as well as the multi-pronunciation lexicon (MPL) to be discussed soon. Obviously, the definition of the SRU set and the MPL is an iterative procedure, the change of one may affect another. An annotated spontaneous speech corpus should also be available to train these SRUs, which at least has the base form (canonical) and surface form (actually observed) strings of SRUs.

The common used SRUs can be phonemes, sub-phonemes, or allophones; and for Chinese they can alternatively be syllables, semi-syllables, or INITIALs/FINALs. For a specific recognizer, the SRUs are often of one pre-selected type.

Two kinds of methods can be used to choose each SRU. One is based on the experts' knowledge [Finke 1997] [Byrne 2001], where a detailed phonetically transcribed spontaneous speech database and *a priori* knowledge/rules on phonetics and linguistics are used. Another one is the data-driven method. An initial acoustic recognizer should be trained at first. Then three approaches are used: (1) the SRU recognition confusion matrix method [Liu and Xu *et al*, 2000]; (2) the specific grammar based generation rules [Cremelie 1997][Cremelie 1999][Liu and Fung 2000]; and (3) the neural network [Fukada 1997] or the decision tree [Byrne 1998] [Riley 1999] to predict possible pronunciation variations and their probabilities given the canonical pronunciation sequence. As a matter of fact, the experts' knowledge based method and the data-driven method can be combined together [Ma 1998].

2. Constructing a multi-pronunciation lexicon

Normally, a lexicon entry is a sequence of SRUs. If each SRU is phonetically canonical, the lexicon will be a single-pronunciation lexicon (SPL), which is the situation in traditional speech recognition. To model pronunciation variation, the SRU set is specially defined so as to cover as many as possible pronunciation variations of each canonical SRU, which results in an SPL being expanded into an MPL. In this sense, the MPL can be regarded to be expanded from a canonical SPL by expanding each SPL entry into several actually observed pronunciations.

The MPL is definitely a superset of the SPL. So the introduction of MPL will obviously result in the confusion among lexicon entries. The choosing of a probabilistic MPL with a suitable size is a tradeoff between the description ability of multiple pronunciations and the increase of the inter-entry confusion.

3. Acoustically modeling spontaneous speech

The purpose here is to model the acoustic variations of spontaneous speech. The common used methods include: (1) using context-dependent modeling and Gaussian sharing technologies to model the pronunciation variation [Saraclar 1999]; (2) using the confusion matrix and possible pronunciation variation rules to model the intra-word and cross-word pronunciation variations [Liu and Xu *et al*, 2000]; (3) using fully data-driven maximum likelihood method to model pronunciation variations [Holter 1999]; and so on.

4. Customizing the decoding algorithm

After the MPL being introduced, the search space of the traditional acoustic decoding network will be greatly enlarged, especially when the context-dependent modeling is applied. The research aims at speeding up the decoder with the recognizer's performance kept. For example, Finke (1999) proposed an improved time synchronous search algorithm to reduce the path expansion scale by introducing intermediate shared nodes during the path expansion. Holter (1999) proposed an A* algorithm based tree-trellis search algorithm that can score multiple pronunciation variations simultaneously in the path.

5. Modifying the statistical language model

Suppose each lexicon entry is a word. For N-gram language modeling based ASR systems, the decoding is based on the following equation

$$\hat{W} = \arg\max_{W} P(X \mid W)P(W) \qquad (1)$$

where $\hat{W}$ is the recognized word sequence given the acoustic signal X and W is any possible word sequence. Once an MPL is introduced, Equation (1) should be modified to reflect the multiple pronunciations for each canonical SPL entry, because $P(X|W)$ is unable to reflect the pronunciation variations for word sequence W [Strik 1999]. One method is to train the surface form word N-Gram using the following equation

$$\hat{W} = \arg\max_{W = Baseform(V)} P(X \mid V)P(V) \qquad (2)$$

where V is one possible surface form word sequence of a base form word sequence and $Baseform(V)$ gives the

canonical word sequence of V. This model contains the pronunciation context information. But a much bigger surface form vocabulary $\{v\}$ results in a much bigger sparseness issue for language model, because each canonical word often corresponds to several surface form words. Another way is to introduce an intermediate term $P(V|W)$ such that

$$\hat{W} = \arg\max_{W=Baseform(V)} P(X|V)P(V|W)P(W) \quad (3)$$

But its disadvantage is that $P(V|W)$, which is the output probability of V given its canonical word sequence W, does not actually reflect the context dependency of W's pronunciation variant V.

In this paper, we will only focus on the pronunciation modeling techniques at the acoustic level, so all the method are acoustically proposed without considering the language model except when specifically stated.

The Chinese annotated spontaneous speech (CASS) corpus will be introduced in Section 2, which is a seed database. Based on the transcription and statistics of CASS corpus, the generalized INITIALs/FINALs (GIFs) are proposed to be the SRUs in Section 3 and therefore the MPL is established. In the following section, we construct the framework for the pronunciation modeling, where an adaptation method is used to refine the acoustic model and a context-dependent weighting method is used to estimate the output probability of any surface form given its corresponding base form. Section 5 lists the experimental results. In Section 6, a method is proposed to refine the model with more yet non-phonetically transcribed data. Summaries and conclusions are given in Section 7.

2. CASS CORPUS

The CASS corpus was created to collect samples of most of the phonetic variations in Chinese spontaneous speech caused by pronunciation effects, including allophonic changes, phoneme reduction, phoneme deletion and insertion, as well as duration changes [Li and Zheng *et al*, 2000]. The CASS corpus is necessary for the definition of the SRU set, the construction of the MPL, and the training of initial acoustic models. Therefore, the CASS corpus can be regarded as a seed database. In this section, the details of CASS corpus will be given briefly.

Made in ordinary classrooms, amphitheatres, or school studios without the benefit of high quality tape recorders or microphones, the recordings are of university lectures by professors and invited speakers, student colloquia, and other public meetings. The collection consists primarily of impromptu addresses, and was delivered in an informal style without prompts or written aids. As a result, the recordings are of uneven quality and contain significant background noises. The recordings were delivered in audiocassettes and digitized into single-channel audio files at 16 kHz sampling rate and with 16-bit precision. A subset of over 3 hours' speech was chosen for detailed annotation, which formed the CASS corpus. This corpus contains the utterances of 7 speakers at a speed as fast as about 4.57 syllables per second on an average [Li and Zheng *et al*, 2000], and in standard Chinese with slight dialectal backgrounds.

The CASS corpus was transcribed into a *five-level* annotation.

- *Character Level*. Canonical sentences (known as word/character sequences) are transcribed.
- *Toned Pinyin* (*or Syllable*) *Level*. A segmentation program was run to convert the character level transcription into word sequences, and then the word sequences were changed into sequences of toned pinyins through a standard word-to-pinyin lookup dictionary. After carefully checked, the canonical toned pinyin transcription was generated.
- *INITIAL/FINAL Level*. This semi-syllable level's transcription only includes the time boundaries for each (observed) surface form INITIAL/FINAL.
- *SAMPA-C Level*. This level contains the observed pronunciation in SAMPA-C [Chen 2000][Li and Chen *et al*, 2000], which is a labeling set of machine-readable International Phonetic Alphabet (IPA) symbols adapted for the Chinese language from the Speech Assessment Methods Phonetic Alphabet (SAMPA). In SAMPA-C, there are 23 phonologic consonants, 9 phonologic vowels and 10 kinds of sound change marks (nasalized, centralized, voiced, voiceless, rounded, syllabic, pharyngrealized, aspirated, inserted, and deleted), by which 21 INITIALs, 38 FINALs, 38 retroflexed FINALs as well as their corresponding sound variability forms can be represented. Tones after tone sandhi, or tonal variation, are attached to the FINALs.
- *Miscellaneous Level*. Several labels related to spontaneous phenomenon are used to independently annotate the spoken discourse phenomena, including modal/exclamation, noise, silence, murmur/unclear, lengthening, breathing, disfluency, coughing, laughing, lip smack, crying, non-Chinese, and uncertain segments. Information at this level can be used for garbage/filler modeling.

3. GENERALIZED INITIALS/FINALS AND GENERALIZED SYLLABLES

According to the Chinese language characteristics, the SRUs are to be chosen at the semi-syllable level, in other words, the Chinese INITIAL/FINAL (IF) level.

In spontaneous speech, there are two kinds of differences between the canonical IFs and their surface forms. One is the sound change or phone change from one canonical IF to a SAMPA-C sequence different from that of any canonical IF. For convenience, we refer to such a SAMPA-C sequence of an IF as one of its *generalized IFs* (GIFs). For example, INITIAL '*zh*' may be changed into

voiced '*z*'. The other is the change completely from one IF to another quite different IF, for example, INITIAL '*zh*' may be changed into '*z*'. Obviously, the IFs can be regarded as special GIFs and the GIF set is a superset of the IF set.

If we want to model the sound variability in the acoustic modeling and choose semi-syllable level units as SRUs, the first thing to do is to define the GIF set.

3.1 Definition of an Initial GIF Set

The canonical IF set consists of 21 INITIALs and 38 FINALs, totally 59 IFs. By searching in the CASS corpus, we initially obtain a GIF set containing over 140 possible SAMPA-C sequences; two examples are given in Table 1. However, some of them occur for only a couple of times which can be regarded as least frequently observed sound variability forms and will be removed from the GIF set in the subsequent step.

Once an initial GIF set is determined, a *GIF transcription* should be made for later use according to the *SAMPA-C Level transcription*. We call this kind of transcription a *dynamic transcription* in comparison with the original five levels' transcriptions. Dynamic transcriptions are useful in both the training and the testing procedures.

Table 1. Examples for IFs and their possible pronunciations in IPA & SAMPA-C format.

IF (Pinyin)	Pronunciation		Comments
	(IPA)	(SAMPA-C)	
z	ts	/ts/	Canonical
z	ts v	/ts_v/	Voiced
z	t©	/ts`/	Changed to '*zh*'
z	t© v	/ts`_v/	Changed to voiced '*zh*'
e	°	/7/	Canonical
e	°r	/7`/	Retroflexed, or changed to '*er*'
e	È	/@/	Changed to /@/ (a GIF)

Table 2. A canonical Chinese syllable and its possible pronunciations in IPA & SAMPA-C with output probabilities.

Syllable (Pinyin)	INITIAL		FINAL		Output Probability
	(IPA)	(SAMPA-C)	(IPA)	(SAMPA-C)	
chang	t©H	/ts`_h/	§Ð	/AN/	0.7850
chang	t©H v	/ts`_h_v/	§Ð	/AN/	0.1215
chang	t© v	/ts`_v/	§Ð	/AN/	0.0280
chang	<deletion>	<deletion>	§Ð	/AN/	0.0187
chang	¸	/z`/	§Ð	/AN/	0.0187
chang	<deletion>	<deletion>	i§Ð	/iAN/	0.0093
chang	tsH	/ts_h/	§Ð	/AN/	0.0093
chang	t©H	/ts`_h/	uÐ	/UN/	0.0093

3.2 Generation of an Initial Generalized Syllable (*GS*) Set

Because we focus only on the acoustic level pronunciation modeling and no Chinese word information will be made available in this research, our lexicon will not consist of Chinese words. Instead, Chinese syllables will be taken to form the lexicon entries. To generate an MPL for the recognizer, the surface form syllables should be found.

Similar to GIF, we refer to any possible pronunciation of a given canonical syllable as one of its *generalized syllables* (GSs). A GS consists of a generalized INTIAL (a GIF) followed by a generalized FINAL (another GIF). All GSs form the GS set, a superset of the canonical syllable set having about 408 toneless syllables. According to the CASS corpus as well as the dynamic GIF transcription, it is easy to find all possible GSs. Then the GS-to-GIF MPL can be generated by expanding the syllable-to-IF SPL.

Table 2 lists 8 MPL entries expanded from the canonical syllable '*chang*', where each entry has an output probability $P((GIF_1, GIF_2) \mid Syllable)$, defined as the probability of the GS or GIF pair (GIF_1, GIF_2) given its corresponding canonical syllable. These probabilities can be learned from the CASS corpus. This gives a

probabilistic GS-to-GIF MPL, where each entry is a Chinese GS with a probability.

3.3 Determining the Final GIF Set and the Final GS set

Now the method for fixing the GIF set (i.e. SRUs) and the GS set (and hence the MPL) is as follows.

For any canonical syllable *b*, all of its possible GSs $\{s_1, s_2, \cdots, s_K\}$ are listed in a descending order of output probability, that is to say, $P(s_i \mid b) \geq P(s_j \mid b)$ *for any* $i \leq j$, of course $\sum_{k=1}^{K} P(s_k \mid b) = 1$. A predefined accumulated output probability (ACP) threshold T_{ACP} is used to choose the first *R* GSs to be reserved such that $\sum_{i=1}^{R} P(s_i \mid b) \geq T_{ACP} > \sum_{i=1}^{R-1} P(s_i \mid b)$ while those thrown GSs are merged into the most similar reserved one. Afterwards, the final GS set is determined. According to the probability definition, the ACP can be thought of as a coverage percentage of pronunciation variations (CPPV) which determines the MPL size.

By collecting all the GIFs appearing in all the reserved GSs, the final GIF set is determined. Accordingly, these well-chosen GIFs are taken as SRUs and the dynamic *GIF Level* transcription should be modified so as to be used in the training procedure. In order to well model the spontaneous speech, additional garbage models are also built for lengthening, breathing, disfluency, lip smacking, murmuring, coughing, laughing, crying, modal/exclamation, silence, noise, and non-Chinese.

In the CASS corpus, the threshold is chosen as $T_{ACP} = 95\%$, and we finally have 86 GIFs and 576 GSs.

3.4 Probabilistic GIF N-Grams

From the statistics of the dynamic GIF transcription, the GIF output and transition probabilities are estimated for later use. The GIF output probability is defined as the probability of a GIF given its corresponding IF, written as *P* (*GIF* | *IF*). To include the GIF deletion, *P* (<*Del*> | *IF*) will also be estimated.

The GIF N-Grams, including unigram *P*(*GIF*), bigram $P(GIF_2 / GIF_1)$ and trigram $P(GIF_3 / GIF_1, GIF_2)$, give the GIF transition probabilities.

4. PRONUNCIATION MODELING

Given an acoustic signal *A* of spontaneous speech, the goal of the recognizer is to find the canonical syllable string *B* that maximizes the probability *P*(*B*|*A*). According to the Bayesian Rule, the recognition result is

$$B^* = \arg\max_B P(B \mid A) = \arg\max_B P(A \mid B) P(B) \qquad (4)$$

In Equation (4), $P(A \mid B)$ is the acoustic modeling part and $P(B)$ is the language modeling part. In this section we focus only on the acoustic modeling and propose some approaches to the pronunciation modeling.

4.1 Theory

Assume *B* is a string of *N* canonical syllables, i.e., $B = (b_1, b_2, \cdots, b_N)$. For simplification, we apply the independence assumption to the acoustic probability,

$$P(A \mid B) \approx \prod_{n=1}^{N} P(a_n \mid b_n) \qquad (5)$$

where a_n is the partial acoustic signal corresponding to syllable b_n. In general, by developing any term in right hand of Equation (5) we have

$$P(a|b) = \sum_s P(a|b,s) P(s|b) \qquad (6)$$

where s is any surface form of a canonical syllable b, in other words, s is one GS corresponding to b. Therefore, the acoustic model is divided into two parts, the first part $P(a|b,s)$ is the refined acoustic model while the second part $P(s|b)$ is the output probability of s given b. Equation (6) provides a solution to the sound variability modeling by introducing a surface form term. In the following subsections, methods for these two parts will be given.

4.2 IF-GIF Modeling

According to the characteristics of Chinese language, any syllable consists of an INITIAL and a FINAL. Because our speech recognizer is designed to take semi-syllables as SRUs, term $P(a|b,s)$ should be rewritten in terms of semi-syllables. Assume $b = (i_c, f_c)$ and $s = (i_g, f_g)$, where i_c and i_g are the canonical INITIAL and the generalized INITIAL respectively, while f_c and f_g the canonical FINAL and the generalized FINAL respectively. Accordingly, the independence assumption results in

$$P(a \mid b,s) \approx P(a \mid i_c, i_g) \cdot P(a \mid f_c, f_g) \qquad (7)$$

More generally, the key point of the acoustic modeling is how to model the IF and GIF related semi-syllable, i.e., how to estimate $P(a \mid IF, GIF)$. There are three different choices:

- *Use* $P(a \mid IF)$ *to approximate* $P(a \mid IF, GIF)$. This is the acoustic modeling based on IFs, named as the **independent IF modeling**.
- *Use* $P(a \mid GIF)$ *to approximate* $P(a \mid IF, GIF)$ [Liu and Xu *et al*, 2000][Saraclar 2000]. This is the acoustic modeling based on GIFs, referred to as the **independent GIF modeling**.

- *Estimate* $P(a \mid IF, GIF)$. This can be regarded as a refined acoustic modeling taking both the base form and the surface form of the SRU into consideration. Thus we refer to it as the **IF-GIF modeling** or **refined acoustic modeling**.

It is obvious that the IF-GIF modeling is the best choice among these three kinds of modeling methods if there are sufficient training data. This kind of modeling method needs a dynamic *IF-GIF transcription*.

The *IF transcription* can be obtained directly from the *Syllable Level transcription* via a simple syllable-to-IF lexicon, and this transcription is canonical. The *GIF transcription* is obtained by means of the method mentioned in Section 3.1 once the GIF set is determined. By comparing the *IF* and *GIF transcriptions*, an actual observed *IF transcription*, named as *IF-a transcription*, is generated, where the IFs corresponding to deleted GIFs are removed and the IFs corresponding to the inserted GIFs are inserted. Finally the *IF-GIF transcription* is generated directly from the *IF-a* and *GIF transcriptions*. Table 3 is an example to illustrate how the *IF-GIF transcription* is obtained.

Table 3. Steps for the generation of IF-GIF transcription.

Step	***Type***	***Transcription***							
0	*IF*	ic_1	fc_1	ic_2	fc_2	ic_3		fc_3	**...**
1	*GIF*	ig_1	fg_1	ig_2		ig_3	g_4	fg_3	**...**
2	*IF-a*	ic_1	fc_1	ic_2		ic_3	c_4	fc_3	**...**
3	*IF-GIF*	ic_1-ig_1	fc_1-fg_1	ic_2-ig_2		ic_3-ig_3	c_4-g_4	fc_3-fg_3	**...**

However, if the training data is not sufficient, the IF-GIF modeling will not work well or even work worse due to the data sparseness issue.

A reasonable method is to generate the IF-GIF models from their associated models, where the adaptation techniques [Young 1999] can be used. There are at least two approaches. The IF-GIF models can be transformed either from the IF models or from the GIF models. The former method is called the base form GIF (B-GIF) modeling and the latter the surface form GIF (S-GIF) modeling.

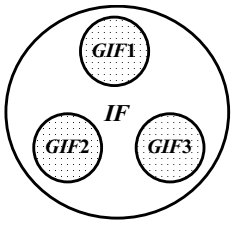


Figure 1. B-GIF modeling: adapting *P*(*a*|*b*) to *P*(*a*|*b*,*s*). Initial IF-GIF models are cloned from the associated IF model. In this example, base form *IF* has three surface forms *GIF*1, *GIF*2, and *GIF*3. Given model *IF*, three initial IF-GIF models, namely *IF-GIF*1, *IF-GIF*2 and *IF-GIF*3, can be generated from it.

The procedure for generating IF-GIF models using B-GIF method is:

Step 1. Train all K IF models $\{IF_k : 1 \le k \le K\}$ according to the *IF-a transcription*.

Step 2. For each k , generate the initial IF-GIF models by copying from model IF_k according to its corresponding M_k GIFs $\{GIF_{km} : 1 \le m \le M_k\}$. The resulting IF-GIF model set is $\{IF_k - GIF_{km} : 1 \le m \le M_k\}$. This procedure is illustrated in Figure 1.

Step 3. Adapt the IF-GIF models according to the corresponding *IF-GIF* transcription. We use the term '*adaptation*' here just for simplification; it is different from its original meaning.

The procedure for generating IF-GIF models using S-GIF method is:

Step 1. Train all M GIF models $\{GIF_m : 1 \le m \le M\}$ according to the *GIF* transcription.

Step 2. For each m , generate the initial IF-GIF models by copying from GIF_m model according to its corresponding K_m IFs $\{GIF_{mk} : 1 \le k \le K_m\}$. The resulting IF-GIF model set is $\{IF_{mk} - GIF_m : 1 \le k \le K_m\}$. This procedure is illustrated in Figure 2.

Step 3. Adapt the IF-GIF models according to the corresponding *IF-GIF* transcription, similarly to the B-GIF method.

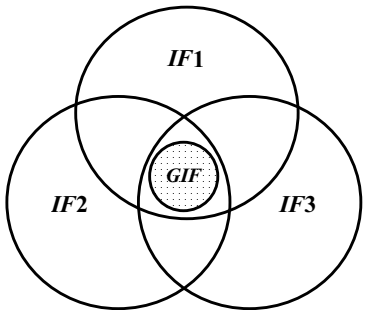


Figure 2. S-GIF modeling: adapting *P*(*a*|*s*) to *P*(*a*|*b*,*s*). Initial IF-GIF models are cloned from the associated

GIF model. In this example, three base forms *IF*1, *IF*2 and *IF*3 share the same surface form *GIF*. Given model *GIF*, three initial IF-GIF models, namely *IF*1-*GIF*, *IF*2-*GIF* and *IF*3-*GIF*, are generated.

The difference between the S-GIF method and the B-GIF method lies only in how we generate the initial IF-GIF models; the former method copies them from the base form models while the latter one from the surface form models. By comparing these two methods as illustrated in Figures 1 and 2, it is straightforward to deduce that the initial IF-GIF models using B-GIF method will have bigger within-model scatters than those using the S-GIF method. The theoretical analysis shows S-GIF method will outperform B-GIF method.

The IF-GIF modeling enables multi-pronunciation for each canonical syllable. Considering the syllable-to-IF_GIF MPL, each entry has the HTK-like form [Young 1999]

$$SYL \quad i_c - i_g \quad f_c - f_g \tag{8}$$

where $SYL = b = (i_c, f_c)$ is the base form and $s = (i_g, f_g)$ is its surface form.

4.3 Context-Dependent Weighting – A Kind of Pronunciation Weighting

In Equation(6), the second part $P(s|b)$ stands for the output probability of a surface form given its corresponding base form.

A simple way to estimate $P(s|b)$ is to directly learn from the database with both base form and surface form transcriptions. The resulting probability is referred to as the direct output probability (DOP). For comparison purpose, no probability appears in any HTK lexicon entry means setting $P(s \mid b) \equiv const$ and hence is called an equal output probability (EOP) scheme.

The problem is that the DOP estimation will not be so accurate if the training database is not big enough. Actually, what we are considering in the pronunciation probability $P(s|b)$ are the base form and surface form of Chinese syllables, and at the syllable level the data sparseness remains a problem, therefore many weights are often not well trained.

It is true that the syllable level data sparseness DOESN'T mean the semi-syllable (IF/GIF) level data sparseness, which suggests us to estimate the output probability via the semi-syllable level statistics instead.

According to the Bayesian Rule, the semi-syllable level output probability of a surface form, i.e. a GIF, given its corresponding base form, i.e. an IF, can be rewritten according to the context information as

$$P(GIF \mid IF) = \sum_{C} P(GIF \mid IF, C) P(C \mid IF) \tag{9}$$

where C is the context of the INITIAL/FINAL *IF*, it can be a bigram, a trigram or whatever related to *IF*. Supposing C includes the current INITIAL/FINAL IF and its left context IF_L, Equation (9) can be rewritten as

$$P(GIF \mid IF) = \sum_{IF_L} P\big(GIF \mid (IF_L, IF)\big) P\big(IF_L \mid IF\big) \tag{10}$$

In the sum on the right hand side of Equation (10), term $P\big(GIF \mid (IF_L, IF)\big)$ is the output probability given the context and term $P\big(IF_L \mid IF\big)$ is similar to the IF transition probability. These two terms can be learned from the database directly; hence Equation (10) is easy to be calculated offline. Based on the way of developing the output probability $P(GIF \mid IF)$, this method is called the context-dependent weighting (CDW) and the estimated probability is called the context-dependent weight (CDW). If we define

$$M_L(GIF \mid IF) = P\big(GIF \mid (L, IF)\big) P\big(L \mid IF\big), \tag{11}$$

Equation (9) can be rewritten as

$$P(GIF \mid IF) = \sum_{IF_L} M_{IF_L}(GIF \mid IF), \tag{12}$$

and according to Equation (10), we define another function as

$$Q(GIF \mid IF) = \max_{IF_L} M_{IF_L}(GIF \mid IF) \tag{13}$$

The above equations are focused on INITIALs and FINALs, and the IF pair (IF_L, IF) could be either a (INITIAL, FINAL) pair or a (FINAL, INITIAL) pair.

To give the syllable level output probability estimation $P(s|b)$ as in Equation (6), we have three different methods:

$$\text{CDW-M:} \quad P(s|b) \approx w_{s|b} = P\big(i_g \mid i_c\big) \cdot M_{i_c}\big(f_g \mid f_c\big) \tag{14}$$

$$\text{CDW-P:} \quad P(s|b) \approx w_{s|b} = P\big(i_g \mid i_c\big) \cdot P\big(f_g \mid f_c\big) \tag{15}$$

$$\text{CDW-Q:} \quad P(s|b) \approx w_{s|b} = Q\big(i_g \mid i_c\big) \cdot Q\big(f_g \mid f_c\big) \tag{16}$$

where $b = (i_c, f_c)$ and $s = (i_g, f_g)$ are as in Section 4.2. Obviously Equation (14) considers the intra-syllable constrains, which is believed to be more useful. Because of Equation (7) especially when using Equation (14) or (16), the sum of approximated $P(s|b)$ over all possible s for b is often not 1.0, that's the reason we call it *weight* instead of *probability*.

If we do not consider the IF-GIF modeling, instead we assume that in Equation (6) $P(a|b,s) \approx P(a|s)$, in other words the acoustic modeling is exactly the independent *GIF* modeling. In this case the use of the CDW results in that the probabilistic syllable-to-GIF MPL will have

entries in the form of

$$SYL \quad w_{s|b} \quad i_g \quad f_g \tag{17}$$

where the weight $w_{s|b}$ can be taken as any one from Equations (14), (15), and (16). Nothing taken for $w_{s|b}$ means the equal probability or weight.

4.4 Integrating IF-GIF modeling and Context-Dependent Weighting

When considering both the CDW and the IF-GIF modeling, we can combine Equations (8) and (17) together and have the probabilistic syllable-to-IF/GIF MPL with entry in the form of

$$SYL \quad w_{s|b} \quad i_c - i_g \quad f_c - f_g \tag{18}$$

4.5 Measuring the Pronunciation Lexicon Intrinsic Confusion

Though the introduction of MPL is useful to describe the pronunciation variation, it also enlarges the among-syllable confusion. From Figure 2, it is obvious that we still cannot judge the original canonical IF given only the observed GIF without a language model or GIF level context information even if the recognizer can achieve 100% acoustic accuracy, because the observed GIF might be generated from several different IFs. Only $\arg\max_{IF} P(GIF \mid IF)$ will be chosen as the final result no mater which *IF* generates this *GIF*. This is an intrinsic feature of the introduced MPL related to a specific weighting scheme. But there are enough reasons to think that the CDW weighting will be better than either the EOP weighting or the DOP weighting because it contains GIF level context information. In this section, we will theoretically analyze the confusion extent of the MPL related to different weighting schemes [Song 2001].

The pronunciation lexicon intrinsic confusion (PLIC) is to be defined as a function of a given MPL *L* and a weighting scheme *W* on *L* based on the following two assumptions:

1. The acoustic model is ideal with accuracy 100% for any testing set; and

2. No word level or syllable level language model is being used.

Assume $B = \{b\}$ is the canonical syllable set and $S = \{s\}$ is the generalized syllable set, and the observation mapping between any $b \in B$ and its possible surface form $s \in S$ is given in *L*, with a joint probability $P(s,b) = P(s \mid b) \cdot P(b)$ forming the weighting scheme $W = \{P(s \mid b), P(b) \mid b \in B, s \in S\}$.

PLIC is designed to reflect the syllable level intrinsic confusion extent for a given *L* and a given *W* on *L*, and is defined as the lower bound of syllable error rate (SER) under the above two assumptions, as follows.

$$PLIC(L,W) = \sum_{s \in S} P(s) \cdot \left(1 - \max_{b \in B} P(b \mid s)\right) \tag{19}$$

where $P(s)$ is the probability of the syllable observation *s*, and $P(b \mid s)$ is the *a posteriori* probability of *s* belonging to *b*, and $\max_{b \in B} P(b \mid s)$ is the probability of *s* being recognized as the *b*' with a maximum *a posteriori* probability. According to the Bayesian equation, Equation (19) can be rewritten as

$$PLIC(L,W) = \sum_{s \in S} \left\{ \sum_{b \in B} P(s \mid b) \cdot P(b) - \max_{b \in B} P(s \mid b) \cdot P(b) \right\} \tag{20}$$

Based on CASS corpus and the choosing of MPL as stated in Section 3.3, the PLIC values for different weighting schemes, EOP, DOP, CDW-P, CDW-Q, and CDW-M, are compared and illustrated in Figure 3.

From Figure 3, we can conclude that PLIC is a decreasing function of CPPV hence that of the lexicon size. The CPPV value of 100% means the MPL contains all possible pronunciations of any canonical syllables, and with the CPPV value decreases to some extent (about 60%) the lexicon becomes an SPL. A tradeoff should be made between the lexicon's confusion extent and the description ability of pronunciation variations.

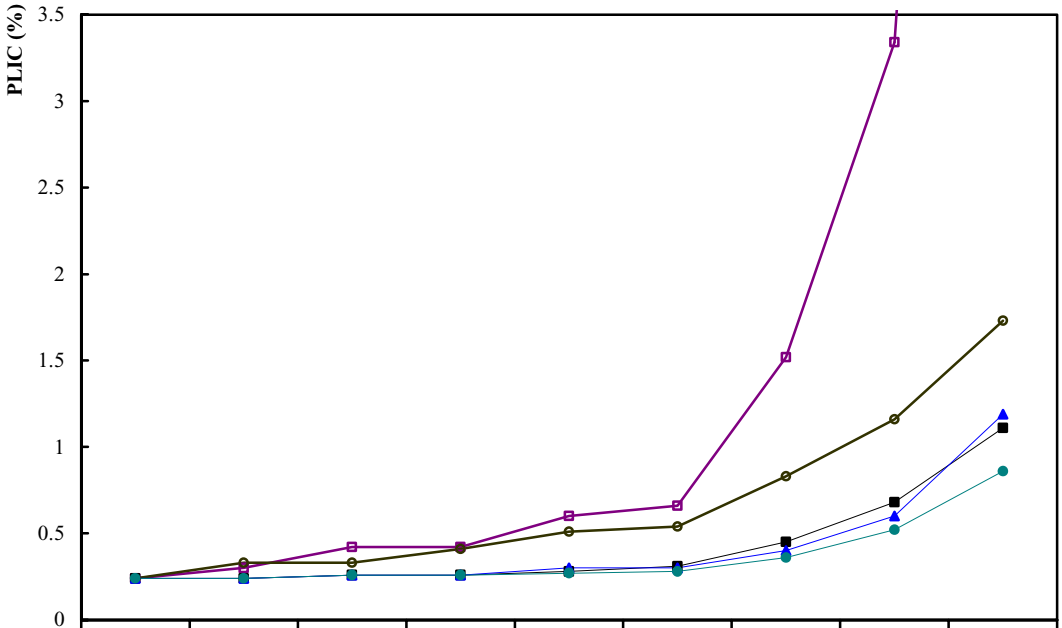
Though the PLIC is not strictly proportional to the SER, lower PLIC values will statistically correspond to higher recognition accuracy. From Figure 3 it is seen that, no matter how big the CPPV value is, the CDW-M weighting scheme always reaches a lowest PLIC value among those five weighting schemes. So it is straightforward that CDW-M will achieve the best recognition performance, theoretically.

5. EXPERIMENTAL RESULTS

All experiments are done across the CASS corpus. The CASS corpus is divided into two parts, the first part is the training set with about 3.0 hours' spontaneous speech data and the second is the testing set with about 15 minutes' spontaneous speech data. The HTK is used for the training, adaptation and testing [Young 1999]. A 3-state 16-gaussian HMM is used to model each IF, GIF or IF-GIF. The feature used here is 39-dimension MFCC_E_D_A_Z. The feature extraction frame size is 25 ms with 15 ms overlapped between any two adjacent frames.

Experimental results include (1) UO: unit (IF, GIF or IF-GIF) level comparison without the syllable lexicon constraint; (2) UL: unit level comparison with the syllable lexicon constraint; and (3) SL: syllable level comparison

with the syllable lexicon constraint. The listed percentages are percent correct *%Cor=%Hit=Hit/Num*100%=(Num-Del-Sub)/Num*100%* and percent accuracy *%Acc=(Hit-Ins)/Num*100%=(Num-Del-Sub-Ins)/Num*100%*, where *Num* is the total number of SRUs in the reference transcriptions, and *Hit*, *Del*, *Sub* and *Ins* indicate the numbers of hit, deletion errors, substitution errors and insertion errors respectively [Young 1999].



CPPV (%)	60	65	70	75	80	85	90	95	100
EOP	0.24	0.3	0.42	0.42	0.6	0.66	1.52	3.34	12.29
DOP	0.24	0.33	0.33	0.41	0.51	0.54	0.83	1.16	1.73
CDW-P	0.24	0.24	0.26	0.26	0.28	0.31	0.45	0.68	1.11
CDW-Q	0.24	0.24	0.26	0.26	0.3	0.3	0.4	0.6	1.19
CDW-M	0.24	0.24	0.26	0.26	0.27	0.28	0.36	0.52	0.86

Figure 3. The Pronunciation Lexicon Intrinsic Confusion (PLIC) curves as a function of the weighting scheme and the syllable level coverage percentage of pronunciation variation (CPPV).

Experiment 1. **Independent IF modeling**. The first experiment is done to test the canonical IF modeling and the result is listed in the second big column of Table 4. The lexicon used here is a single-pronunciation syllable-to-IF lexicon with equal weight, because each syllable corresponds to a unique canonical INITIAL and a unique canonical FINAL. This is just for comparison.

Table 4. Results of Independent IF/GIF Modeling.

Item	IF		GIF	
	%Cor	%Acc	%Cor	%Acc
UO	46.28	41.70	44.62	40.02
UL	50.34	42.30	47.55	39.95
SL	34.92	30.48	33.91	29.14

Experiment 2. **Independent GIF modeling**. This is the baseline system where an equal probability or weight is provided for the syllable-to-GIF MPL. Experimental result is shown in the third big column of Table 4. By comparing the two experiments, we find that in general the performance of independent GIF modeling is worse than that of independent IF modeling if no more pronunciation method is adopted. This is obvious because the GIF set is bigger than the IF set and obviously the PLIC of the GIF set is bigger than that of the IF set, which results in that GIFs are not better trained than IFs on the same training database.

Experiment 3. **IF-GIF Modeling**. This experiment is designed to test the IF-GIF modeling, $P(a|b,s)$. Except the acoustic models themselves, the experiment condition is similar to that in Experiment 2. The B-GIF and S-GIF modeling results are given in Table 5. We have tried the mean updating, MAP adaptation and MLLR adaptation

methods for both the B-GIF and the S-GIF modeling, and listed are the best results.

From this table, it is seen that S-GIF outperforms B-GIF; the reason can be seen in Figures 1 and 2 and is explained in Section 4.2. Compared with the GIF modeling, the S-GIF modeling achieves an SER reduction of 3.6%.

Table 5. Results of the IF-GIF Modeling.

Item	B-GIF		S-GIF	
	%Cor	%Acc	%Cor	%Acc
UO	43.31	38.67	41.36	36.83
UL	46.67	38.25	46.47	38.85
SL	36.07	31.39	36.63	31.67

Table 6. Effects of the use of IF-GIF modeling and syllable bi-gram.

Item (SL)	w/ GIF modeling		w/ IF-GIF modeling		w/ IF-GIF modeling & Syllable Bi-Gram	
	%C	%A	%C	%A	%C	%A
DOP	35.85	31.15	-	-	-	-
CDW-P	36.00	31.31	-	-	-	-
CDW-Q	35.71	31.29	-	-	-	-
CDW-M	37.25	32.76	37.87	33.39	40.90	36.75

Experiment 4. **Pronunciation Weighting**. This experiment is designed to find a best way to estimate the pronunciation weight $P(s|b)$. To avoid the influence from the IF-GIF modeling, we use GIF modeling only, in other words we assume $P(a|b,s) \approx P(a|s)$. The EOP and $P(a|b,s) \approx P(a|b)$ are not considered because they are much worse. In the syllable lexicon, two kinds of pronunciation weighting schemes, i.e. DOP and CDW, are used for each entry. The results for DOP and CDW methods are listed in second big column of Table 6. Though for CDW $\sum_S P(S|B) \le 1$ and mostly it does not meet $\sum_S P(S|B) = 1$ as DOP does, CDW performs better than DOP. Compared with the GIF modeling, the pure pronunciation weighting method CDW achieves a SER reduction of 5.1%.

Experiment 5. **Integrated Pronunciation Modeling**. Either IF-GIF modeling or CDW pronunciation weighting improves the system performance individually; we have reason to believe that the integration of CDW and IF-GIF modeling will improve the performance much better. The result is given in the third big column of Table 6. The SER reduction is 6.3% totally compared with the GIF modeling.

Experiment 6. **Integration of syllable N-gram**. Though language modeling is not the focus of pronunciation modeling, to make Equation (4) a complete one, we borrow a cross-domain syllable language model. This syllable bigram is trained using both read texts from *Broadcast News* (*BN*) and spontaneous texts from CASS, the amount of texts from BN is much bigger than those from CASS, and therefore we call it a borrowed cross-domain syllable bigram. From the result listed in the fourth big column of Table 6, it is not difficult to conclude that this borrowed cross-domain syllable N-gram is helpful. The SER reduction is 10.7%.

Figure 4 gives an outline of all above experimental results. The overall SER reduction compared with GIF modeling and IF modeling is 6.3% and 4.2% (all without syllable N-gram).

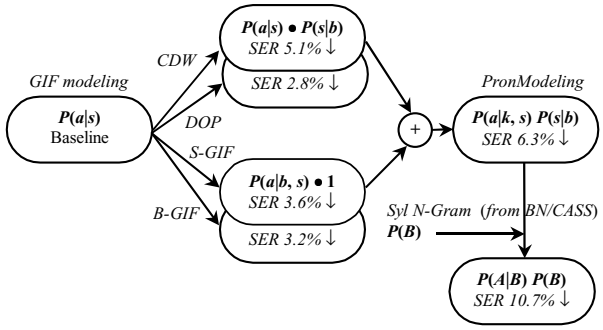


Figure 4. A summary of experimental results.

6. USING MORE DATA WITHOUT PHONETIC TRANSCRIPTION

The above experiments have proved that the proposed methods, including the concept of GIF, the refined acoustic modeling (IF-GIF modeling), and the context-dependent weighting (CDW), are effective, but they seem much dependent on the phonetically transcribed database, the CASS corpus. A question is whether these methods are still effective when more data without phonetic transcription is used to refine the acoustic model. The solution is given in this section to the raised question.

We regard the CASS corpus or a similar phonetically transcribed database as a seed database. The seed database is mainly used to define the GIF set and to initially train the CDW weights. A well designed seed database with IF N-Grams and GIF N-grams (*N* can be 1 through 3) well balanced and with most spontaneous phenomena well covered will be definitely helpful to the definition of the GIF set.

The seed database cannot be very big because the transcription effort is extremely great. When more data with syllable level transcription yet without GIF level transcription is available, the data-borrowing and deleted interpolation [Jelinek 1998][Huang 1996][Kim 1997] can be useful ideas to refine the existing acoustic model. The

problem is that the syllable level transcription is often a canonical syllable level transcription instead of a phonetic transcription. To solve the problem, we propose an iterative forced-alignment based transcribing (IFABT) method which is a data-driven one. The IFABT procedure can be described as follows.

Step 1. Using the seed database to define a GIF set and a syllable-to-GIF MPL, to train the context-dependent weights, and to train the IF-GIF model.

Step 2. Using the forced-alignment technique [Young 1999] and the MPL to decode both the seed database and the given bigger database with syllable level transcription so that an IF-GIF transcription can be generated.

Step 3. Using the two databases with the IF-GIF transcription to redefine the MPL and to retrain the context-dependent weights and the IF-GIF models.

Step 4. If the overall recognizer performance does not achieve a predefined performance threshold across a supervising set (which is another set different from either the training set or the testing set), go back to Step 2, otherwise stop.

We establish another three-hour database called CASS-II under almost the same condition as that of CASS corpus. CASS-II is only transcribed at the Chinese syllable and character level. By using the IFABT method and combining the two corpora into a six-hour database, we reduce the SER by about 5.7% across the same testing set as that used in the previous experiments.

7. SUMMARIES AND CONCLUSIONS

In order to model the pronunciation variability in spontaneous speech, firstly we propose the concept of generalized INITIAL/FINAL (GIF) and generalized syllable (GS) with or without probabilities, secondly we propose the GIF modeling and the IF-GIF modeling aiming at refining the acoustic models, thirdly we propose the context-dependent weighting method to estimate the pronunciation weights, and then we integrate the cross-domain syllable N-gram into the whole system. An iterative forced-alignment based transcribing (IFABT) method is finally proposed and verified for use in the case that only a small portion of database is phonetically transcribed.

The purposes of the above method can be summarized as follows. (1) The definition of the GIF set and the GS set is to cover more pronunciation variations in spontaneous speech. (2) The refined acoustic modeling (the IF-GIF modeling) is to introduce the difference of a GIF generated from different IFs so that the acoustic modeling is more refined. (3) The context-dependent weighting is to reduce the intrinsic confusion of the MPL and to improve the estimation accuracy of the lexicon entry probabilities (weights) by using the GIF level context information, and to solve the sparseness problem in the IF-GIF modeling. (4) The IFABT method is for use to better train the recognizer using an extra non-phonetically transcribed database.

It can be seen that although the introduction of the IF-GIF modeling and the pronunciation weighting leads to performance reduction at the unit level compared with the IF modeling, the syllable level overall performance for IF-GIF modeling greatly outperforms the IF modeling. From the experimental results, we conclude that

- The overall GIF modeling is better than the IF modeling.
- By refining the IF and GIF, the resulting IF-GIF modeling $P(a \mid b,s)$ is better than both the IF modeling $P(a \mid b)$ and the GIF modeling $P(a \mid s)$, even if data is sparse, when the S-GIF/B-GIF adaptation techniques are used to provide a solution to data sparseness.
- The S-GIF method outperforms the B-GIF method because of the well-chosen initial models for adaptation.
- The context-dependent weighting (CDW) is more helpful for sparse data than direct output probability (DOP) estimating.
- The cross-domain syllable N-Gram is useful.
- The above methods are still effective even when only a small portion of the database is transcribed at the phonetic level by applying the proposed IFABT method to the seed database and the non-phonetically transcribed database.

8. ACKNOWLEDGEMENTS

The authors would like to thank Prof. Aijun Li and her colleagues with the Institute of Linguistics, Chinese Academy of Social Sciences, for their efforts in transcribing the CASS corpus and their comments on the phonetic analysis of the CASS corpus. Thanks are also given to other members of the *Mandarin Pronunciation Modeling* team, Dr. Yi Liu, Hong Kong University of Science and Technology, Dr. Veera Venkataramani, the Johns Hopkins University, Mr. Umar Ruhi, University of Toronto, and Miss Terri Kamm, Department of Defense of USA, all the discussions in the 2000 Summer Workshop are quite valuable for this paper.

9. REFERENCES

[1] Byrne W, Finke M, Khudanpur S *et al*. "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 313-316, May 1998, Seattle.

[2] Byrne W, Venkataramani V, Kamm T *et al*. "Automatic generation of pronunciation lexicons for Madarin spontaneous speech," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. I, May 2001, Salt Lake City.

[3] Chen X-X, Li A-J, *et al*, "An application of SAMPA-C for standard Chinese," *International Conference on Spoken Language Processing*, 4: 652-655, Oct. 2000,

Beijing.

[4] Cremelie N and Martens J P. "Automatic rule-based generation of word pronunciation networks," *European Conference on Speech Communication and Technology* (*EuroSpeech'97*). 1997, 5: 2459-2462.

[5] Cremelie N and Martens J P. "In search of better pronunciation models for speech recognition," *Speech Communication*, 29: 115-136, 1999.

[6] Decker A M and Lamel L. "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, 29: 83-98, 1999.

[7] Finke M and Waibel A. "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," *European Conference on Speech Communication and Technology* (*EuroSpeech'97*), 5: 2379-2382, 1997.

[8] Finke M, Fritsch J, Koll D, *et al*. "Modeling and efficient decoding of large vocabulary conversational speech," *European Conference on Speech Communication and Technology* (*EuroSpeech'99*), 1: 467-470, 1999.

[9] Fosler-Lussier E and Morgan N. "Effect of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, 29: 137-158, 1999.

[10] Fukada T and Sagisaka Y. "utomatic generation of a pronunciation dictionary based on a pronunciation network," *European Conference on Speech Communication and Technology* (*EuroSpeech'97*), 5: 2471-2474, 1997.

[11] Greenberg S. "Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29: 159-176, 1999.

[12] Holter T and Svendsen T. "Maximum likelihood modelling of pronunciation variation," *Speech Communication*, 29: 177-191, 1999.

[13] Huang X-D, Hwang M-Y, Jiang L, *et al*. "Deleted interpolation and density sharing for continuous hidden Markov models," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 885-888, Atlanta, GA, 1996.

[14] Jelinek F. "Statistical methods for speech recognition," The MIT Press, Cambridge, MA, 1998.

[15] Kim N S and Un C K. "Statistically reliable deleted interpolation," *IEEE Trans. SAP*, 5: 292-295, 1997.

[16] Li A-J, Zheng F, Byrne W, Fung P, *et al*, "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech," *International Conference on Spoken Language Processing* (*ICSLP*'2000), 1: 485-488, Oct. 2000, Beijing

[17] Li A-J, Chen X-X, Sun G, *et al*, "The phonetic labeling on read and spontaneous discourse corpora," *International Conference on Spoken Language Processing* (*ICSLP*'2000), 4: 724-727, Oct. 2000, Beijing

[18] Liu M-K, Xu B, Huang T-Y, *et al*, "Mandarin accent adaptation based on context-independent / context-dependent pronunciation modeling," *International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*'2000), 4: 1025-1028, June 2000, Istanbul

[19] Liu Y and Fung P. "Rule-based word pronunciation networks generation for Mandarin speech recognition," *International Symposium of Chinese Spoken Language Processing*, 35-38, Oct. 2000, Beijing.

[20] Ma K, Zavaliagkos G, and Iyer R. "Pronunciation modeling for large vocabulary conversational speech recognition," *International Conference on Spoken Language Processing*, 6:2455-2458, Nov. 1998, Sydney.

[21] Riley M, Byrne W, Finke M, *et al*. "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, 29: 209-224, 1999.

[22] Saraclar M, Nock H, and Khudanpur S. "Pronunciation modeling by sharing Gaussian densities across phonetic models," *European Conference on Speech Communication and Technology* (*EuroSpeech'99*), 1:515-518, 1999.

[23] Saraclar M, Nock H and Khudanpur S. "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, 14: 137-160, 2000.

[24] Song Z-J. "Research on pronunciation modeling for spontaneous Chinese speech recognition," Ph.D. Dissertation: Beijing, Tsinghua University. Apr. 2001

[25] Strik H and Cucchiarini C. "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, 29: 225-246, 1999.

[26] Young S, Kershaw D, Odell J, Ollasen D, Valtchev V, and Woodland P. "The HTK Book: Version 2.2," Entropic Ltd., 1999.

[27] Zheng F. "A Syllable-Synchronous Network Search Algorithm for Word Decoding in Chinese Speech Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, II: 601-604, March 1999, Phoenix.