International Conference on Spoken Language Processing 2002, pp. 2461-2464, Sep. 16-20, 2002, Colorado, USA

REDUCING PRONUNCIATION LEXICON CONFUSION AND USING MORE DATA WITHOUT PHONETIC TRANSCRIPTION FOR PRONUNCIATION MODELING

Fang Zheng¶*, *Zhanjiang Song*¶*, *Pascale Fung*†, *and William Byrne*‡

¶ Center of Speech Technology, State Key Lab of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua Univ., Beijing, 100084, China
† Department of Electrical and Electronic Engineering, Hong Kong Univ. of Science and Technology
‡ Center for Language and Speech Processing, The Johns Hopkins Univ., USA
fzheng@sp.cs.tsinghua.edu.cn, *http://sp.cs.tsinghua.edu.cn/~fzheng/*

ABSTRACT

The multiple-pronunciation lexicon (MPL) is very important to model the pronunciation variations for spontaneous speech recognition. But the introduction of MPL brings out two problems. First, the MPL will increase the among-lexicon confusion and degrade the recognizer's performance. Second, the MPL needs more data with phonetic transcription so as to cover as many surface forms as possible. Accordingly, two solutions are proposed, they are the context-dependent weighting method and the iterative forced-alignment based transcription method. The use of them can compensate what the MPL causes and improve the overall performance. Experiments across a naturally spontaneous speech database show that the proposed methods are effective and better than other methods.

1. INTRODUCTION

The performance degrading of an ASR system for spontaneous speech is mainly caused by the difference of pronunciation styles between read and spontaneous speeches, either at the phonetic level or the linguistic level.

At the phonetic level, the spontaneous speech contains much more phone change and sound change phenomena because of variable speaking rates, moods, emotions, prosodies, co-articulations and so on. Other phenomena, such as lengthening, breathing, disfluency, lip smacking, murmuring, coughing, laughing, crying, modal/exclamation, 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.

Table 1. Terms and symbols used in this paper.

Term	***Meaning***
Syllable	The pronunciation of character used in written Chinese. Totally 408.
INITIAL	First part of Chinese syllable.
FINAL	Second/last part of Chinese syllable.
CIF or IF	Canonical INITIAL or FINAL. Totally 59.
GIF	Generalized INITIAL or FINAL, defined according to surface form transcriptions [6]. The GIF set is a superset of the IF set.
GS	Generalized syllable whose two parts are GIFs.
phonetic transcription	The phoneme level transcription that can be used to define the GIF set.
lexicon	A syllable-to-IF or syllable-to-GIF vocabulary, with or without the surface form output probability.

Symbol	***Meaning***
${}^{c}i$	A canonical INITIAL.
${}^{c}f$	A canonical FINAL.
${}^{g}i$	A generalized INITIAL.
${}^{g}f$	A generalized FINAL.
$b=({}^{c}i, {}^{c}f)$	A canonical (base form) syllable.
$s=({}^{g}i, {}^{g}f)$	A generalized (surface form) syllable.
a	An acoustic signal.
SYL	A Chinese syllable as a lexicon entry.
w	The output probability or weight of a surface form syllable given its canonical one in a lexicon entry [5].

For Chinese, pronunciation variations are made

* The authors are currently with Beijing d-Ear Technologies Co., Ltd.

especially severe in casual speech since most Chinese people are non-native standard Chinese speakers and are with complicated dialect and accent backgrounds. A syllable in an accent or dialect may correspond to a different one in another accent or dialect; this is an accent/dialect related pronunciation shift.

Though the Gaussian density sharing technology is useful to pronunciation modeling [2], it cannot actually provide a full and efficient solution to the above problems. Actually, the multiple-pronunciation lexicon (MPL) can be used to describe the pronunciation variations in different situations [4][1][6].

When only the acoustic model is focused on, the introduction of MPL leads to the following equation,

$$P(a \mid b) = \sum_{s} P(a \mid b,s)P(s \mid b)\,, \qquad (1)$$

where the definitions and meanings of symbols can be found in Table 1. Therefore, the acoustic model is divided into two parts, the first part $P(a|b,s)$ is the refined acoustic model (RAM) while the second part $P(s|b)$ is the output probability of s given b. This provides a solution to the variation modeling by introducing a surface form term.

Two problems arise. First, a sufficient enough database as well as both the base form and the surface form transcription should be established, where the surface form transcription is a time and manpower consuming procedure. Second, the introduction of surface form increases the pronunciation lexicon's intrinsic confusion (PLIC), that is to say, the confusion extent among surface form syllables in the lexicon.

In this paper, we will prove that the previously proposed context-dependent weighting (CDW) [6] can be used to reduce the PLIC. An iterative forced-alignment based transcription (IFBAT) is also proposed to meet the requirement of the phonetically transcribed database.

2. REDUCING LEXICON'S CONFUSION

In this section, the CDW method will be introduced, the PLIC will be defined, and then it will be proved that the CDW method is helpful to reduce the PLIC value.

2.1. Context-Dependent Weighting

A kind of adaptation method is adopted to get the RAM part in Equation (1), which efficiently provides a good solution to the data sparseness for the IF-GIF refined acoustic modeling [6]. And therefore each MPL entry has the following HTK-like form [5]

$$SYL \quad {}^{c}i\text{-}{}^{g}i \quad {}^{c}f\text{-}{}^{g}f \qquad (2)$$

Equation (2) specifies an equal output probability (EOP) for (${}^{c}i$-${}^{g}i$, ${}^{c}f$-${}^{g}f$) given SYL, which should be modified according to the second part $P(s|b)$ in Equation (1).

A simple estimation of $P(s|b)$ is the direct output probability (DOP) which is calculated from the database. However, the transcribed data might be too sparse to get an accurate estimation for it because of the large number of generalized syllables s'. It is straightforward to think that the sparseness problem is smaller at a lower level than at the syllable level. So we propose a context-dependent weighting (CDW) method to estimate it. The idea can be expressed by Equation (3).

$$P(GIF \mid IF) = \sum_{C} P(GIF \mid IF, C)P(C \mid IF) \qquad (3)$$

where C can be any context, for example, IF pairs, or GIF pairs. Considering the data sparseness, we use left bi-IF (IF_L, IF) as the context. If we define

$$M_L(GIF \mid IF) = P(GIF \mid (L, IF))P(L \mid IF) \qquad (4)$$

Equation (3) becomes

$$P(GIF \mid IF) = \sum_{L} M_L(GIF \mid IF) \qquad (5a)$$

We further define an alternative form as

$$Q(GIF \mid IF) = \max_{L} M_L(GIF \mid IF) \qquad (5b)$$

Thereafter, we have three kinds of estimations for $P(s|b)$,

$$\text{CDW-M:} \quad P(s \mid b) \approx w_{s|b} = P({}^{g}i|{}^{c}i) \cdot M_{c_i}({}^{g}f|{}^{c}f) \qquad (6a)$$

$$\text{CDW-P:} \quad P(s \mid b) \approx w_{s|b} = P({}^{g}i|{}^{c}i) \cdot P({}^{g}f|{}^{c}f) \qquad (6b)$$

$$\text{CDW-Q:} \quad P(s \mid b) \approx w_{s|b} = Q({}^{g}i|{}^{c}i) \cdot Q({}^{g}f|{}^{c}f) \qquad (6c)$$

For GIF and IF-GIF modeling, the MPL entry has the following form respectively.

$$SYL \quad w_{s|b} \quad {}^{g}i \quad {}^{g}f \qquad (7a)$$

$$SYL \quad w_{s|b} \quad {}^{c}i\text{-}{}^{g}i \quad {}^{c}f\text{-}{}^{g}f \qquad (7b)$$

If we replace $w_{s|b}$ with the directly calculated $P(s|b)$, we have the DOP form; if we remove $w_{s|b}$ or replace it with a constant, we have the EOP form.

2.2. Pronunciation Lexicon's Intrinsic Confusion

The introduction of MPL is useful to describe the pronunciation variations, but it also enlarges the among-syllable confusion. It is obvious that we cannot judge the original canonical IF given only the observed GIF without a language model or GIF level context information even if the GIF 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 and has relatively more sufficient observation data. In this section, we will theoretically analyze the confusion extent of the MPL related to different weighting schemes [3].

The 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% at the IF/GIF level for any testing set; and (2) neither character-level nor 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 $P(s, b)= P(s|b)\cdot P(b)$ forming the weighting scheme $W=\{P(s|b), P(b) \mid b\in B, s\in S\}$.

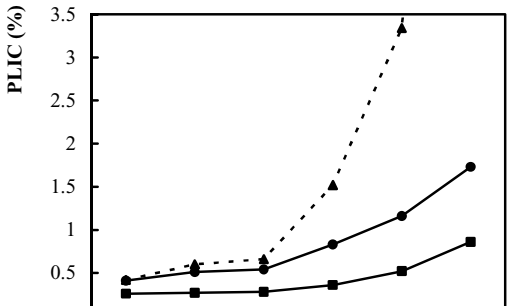
PLIC is designed to reflect the syllable level intrinsic confusion extent for a given L and a given W on L, and is defined as the lower bound of the canonical syllable error rate (SER) under the above two assumptions, as follows.

$$PLIC(L,W)=\sum_{s\in S}P(s)\cdot\left(1-\max_{b\in B}P(b\mid s)\right) \qquad (8)$$

where $P(s)$ is the probability of the syllable observation s, and $P(b|s)$ 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 a b' with a maximum *a posteriori* probability. And,

$$PLIC(L,W)=\sum_{s\in S}\left\{\sum_{b\in B}P(s\mid b)\cdot P(b)-\max_{b\in B}P(s\mid b)\cdot P(b)\right\} \qquad (9)$$

Based on CASS corpus and the choosing of MPL as in [6], the PLIC values for different weighting schemes, EOP, DOP, and CDW-M, are compared and illustrated in Figure 1. (The CDW-P and CDW-Q curves, which are not drawn, are between those of DOP and CDW-M.)



CPPV (%)	75	80	85	90	95	100
EOP	0.42	0.6	0.66	1.52	3.34	12.29
DOP	0.41	0.51	0.54	0.83	1.16	1.73
CDW-M	0.26	0.27	0.28	0.36	0.52	0.86

Figure 1. The PLIC curve as a function of the weighting scheme and the syllable level CPPV.

From Figure 1, we can conclude that PLIC is an increasing function of the coverage percentage of pronunciation variations (CPPV) and hence is 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 a single pronunciation lexicon (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 1 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 weighting schemes. So it is straightforward that CDW-M will achieve the best recognition performance, theoretically.

3. USING DATA WITHOUT PHONETIC TRANSCRIPTIONS

The proposed methods, including the concept of GIF, the refined acoustic modeling (IF-GIF modeling), and the context-dependent weighting, are effective, but they seem much dependent on the phonetically transcribed database [6]. 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.

3.1. Use of Seed Database

Though the phonetic transcription as a seed database costs a lot of time and manpower, it is necessary for the proposed method. It is used to define the GIF set and the syllable-to-GIF MPL and to train the initial RAM and CDW. It need not be too big but should be big enough to cover the most common seen GIFs and to get the relatively accurate RAM and CDW.

3.2. Automatic GIF Transcription for Extra Database

All the speech data without phonetic (GIF) transcriptions form a so-called extra database.

The HTK tools [5] provide the HV_{ITE} command which can be used to compute forced alignments. It can compute a new network for each input Chinese utterance using the canonical Chinese syllable level transcriptions, an MPL, a RAM and a CDW, and output the transcription containing the generalized syllables (and of course GIFs) and their boundaries. Figure 2 gives the procedure. It is to determine the actual or best matching pronunciations used in the utterances used to further refine the HMM system.

3.3. Iterative Forced-Alignment Based Transcription

The purpose of the extra database with syllabic transcription is to refine (1) the acoustic IF-GIF model and (2) the CDW weights. To reach this goal, we propose a data-driven iterative forced-alignment based transcribing (IFABT) method which can be described as follows.

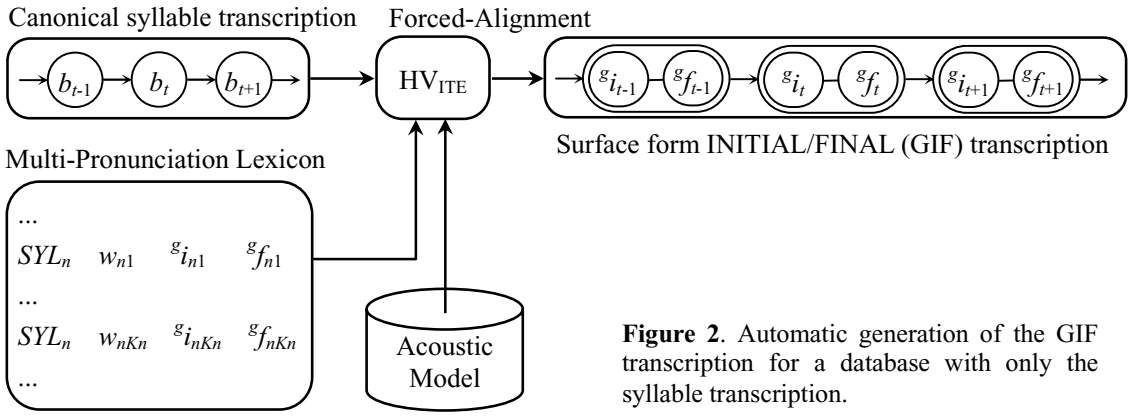


Figure 2. Automatic generation of the GIF transcription for a database with only the syllable transcription.

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

Step 2. Using the modified forced-alignment technique as illustrated in Figure 2 and the MPL to decode both Seed Database and Extra Database so that an IF-GIF transcription can be generated.

Step 3. Using these 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 performance does not achieve a predefined threshold across a supervising set, go back to Step 2, otherwise stop.

4. EXPERIMENTS AND CONCLUSIONS

In this paper, Seed Database is the CASS corpus with 3 hours' speech data while Extra Database contains another 3 hours' data. Extra Database shares the similar recording condition and domain to that of CASS, and hence is called CASS-II which contains canonical syllable transcriptions and no phonetic transcriptions. Totally 86 GIFs, 580 GSs and 609 MPL entries are defined for CASS.

Experiments are done using HTK [5] and the experimental conditions are the same as those in [6]. For EXP1, 15 minutes' data from CASS is the testing set while other 3 hours' data is the training set. For EXP2, the IFABT method is used, the testing test contains additional 15 minutes' data from CASS-II while the training set contains additional 3 hours' data from CASS-II.

Experimental results are given in Figure 3, where 'CI' stands for context-independent, 'CD' context-dependent (tri-IF or tri-GIF), and 'Sharing' decision tree based Gaussian density sharing. The listed numbers are syllable accuracy percentage in form of "EXP1 # EXP2". From the results, we come to the following conclusions:

(1) At the syllable level, the use of GIFs as acoustic models always achieves better results than IFs.

(2) Either the context dependent modeling or the Gaussian density sharing is a good method for pronunciation variation modeling.

(3) The context-dependent weighting is more useful than the Gaussian density sharing for pronunciation modeling, because it can reduce MPL's PLIC value.

(4) The IFABT method is helpful when more data with higher level transcription yet without the phonetic transcription is available.

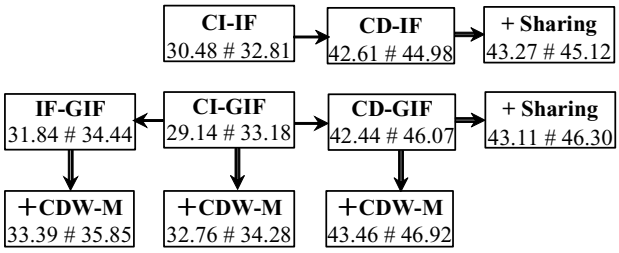


Figure 3. Accuracy comparison.

5. REFERENCES

[1] Fung P, Byrne W, and Zheng F, *et al*. "Pronunciation Modeling of Mandarin Casual Speech," *Final Report of Workshop 2000 on Language and Speech Processing*, http://www.clsp.jhu.edu/ws2000/final_reports/mpm/.

[2] Saraclar M, Nock H, and Khudanpur S. "Pronunciation modeling by sharing Gaussian densities across phonetic models," *EuroSpeech'99*, 1:515-518, 1999.

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

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

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

[6] Zheng F, Song Z-J, Fung P, and Byrne W. "Modeling Pronunciation Variation Using Context-Dependent Weighting and B/S Refined Acoustic Modeling," *EuroSpeech,* 1:57-60, Sept. 3-7, 2001, Aalborg, Denmark