Pronunciation Modeling With Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework

Ming-Yi Tsai, Fu-Chiang Chou, and Lin-Shan Lee, *Fellow, IEEE*

***Abstract*—Multiple-pronunciation dictionaries have been found to be useful in pronunciation modeling for speech recognition. However, the extra pronunciation variants added in the dictionary inevitably increase the confusion among different words during recognition, and consequently limit the achievable improvements in the recognition performance. This paper proposes a three-stage framework for Mandarin Chinese to construct automatically the multiple-pronunciation dictionary while reducing the possible confusion caused. The proposed framework includes pronunciation generation (Stage 1), ranking (Stage 2) and pruning (Stage 3). New measures of confusability for multiple-pronunciation dictionaries were developed and shown to have a very strong correlation with recognition performance. With the proposed framework, it was shown that the confusability as measured can be reduced and recognition performance improved stage by stage. All of the above findings were verified by a series of experiments performed on both planned (LDC HUB-4NE) and spontaneous (LDC CALLHOME) Mandarin Chinese speech corpora.**

***Index Terms*—Confusability, confusion, multiple-pronunciation dictionary, pronunciation modeling, pronunciation variation, speech recognition.**

I. Introduction

It IS well known that the pronunciation variation that is present in natural speech is one of the major sources of errors in automatic speech recognition (ASR). To capture such variation, many ASR systems have employed multiple-pronunciation dictionaries that include pronunciation variants in addition to the canonical pronunciations of words. The construction of a good multiple-pronunciation dictionary is thus critical to favorable ASR performance.

A variety of methods have been utilized to obtain the multiple-pronunciation dictionaries, and such approaches have been comprehensively surveyed [1]. Most of these methods acquire the pronunciation variation information by automatically transcribing the surface forms from speech corpora with speech recognizers [2]–[16] rather than by manual transcription for better efficiency and consistency with the later recognition processes. However, automatically extracting reliable surface forms that faithfully reflect what has actually been pronounced is very challenging because recognition errors are inevitable. Such transcription errors may increase the confusion of the constructed dictionary since "incorrect" variants may be added and shared by different words. In order to reduce the transcription errors and the confusion, phone-bigram recognizers [2], [3], [10], [11], [15], [16] or decision trees [2], [7], [9]–[11], [16] have been adopted for the transcription. Some experimental investigations have shown that the dictionaries obtained using phone-bigram recognizers significantly improves recognition performance, while decision trees offer less, or even insignificant improvement over the baseline dictionary [2]. However, phone-bigram recognizers still produce many recognition errors and introduce extra confusion into the constructed dictionary.

Once obtained, the pronunciation variation information can be incorporated into dictionaries either explicitly or implicitly. In an explicit approach, the pronunciations of words observed in surface transcriptions are enumerated in a table, which are later selected in the dictionary [2], [7]–[12], [14], [15]. In an implicit approach, phonological rules [5], [6], [13], [16]–[19], decision trees [2], [7], [9]–[11], artificial neural networks [20], or a phone confusion table [4] are applied to baseforms of words to generate their surface pronunciations. Although having the potential to generate surface pronunciations of unseen words, such implicit approaches may under- or over-generate the pronunciations of words that share the same phonotactic context, and so have been found to result in comparable [7], [9] or sometimes worse [16] recognition performance than explicit approaches. Therefore, in this paper, the multiple-pronunciation dictionaries were explicitly obtained.

Regardless of whether pronunciation variants in a dictionary are obtained implicitly or explicitly, the added variants inevitably introduce extra lexical confusion in the dictionary by increasing the number of words that share identical or similar pronunciations. Such increased confusion seriously limits the achievable improvement in recognition performance when multiple pronunciations are considered. Many methods have been proposed to avoid such confusion. One straightforward scheme is to prune the variants based on their frequencies [2], [9], [13]–[15], [21]. Phonological rules for generating the variants can also be pruned based on acoustic likelihood [22] or the pronunciation entropy of phones [23], [24]. Word frequencies [10], word pronunciation entropy [14], and the ratio of pronunciation probabilities [17] have also been used to select the variants. However, none of these approaches actually measured or decreased the confusability directly. Direct methods, such as inverted finite-state transducers, have been

Manuscript received March 14, 2005; revised January 17, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

M.-Y. Tsai was with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan, R.O.C. She is now with MediaTek, Inc., Hisinchu 300, Taiwan, R.O.C. (e-mail: pancho.tsai@gmail.com).

F.-C. Chou is with the Graduate Institute of Computer and Communication Engineering, Ming Chuan University, Taipei 111, Taiwan, R.O.C. (e-mail: fuchiang@mcu.edu.tw).

L.-S. Lee is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan, R.O.C. (e-mail: lslee@gate.sinica.edu.tw).

Digital Object Identifier 10.1109/TASL.2006.876769

1558-7916/$25.00 © 2006 IEEE

adopted to estimate explicitly the lexical confusion, but they have not yet practically been integrated into pronunciation modeling process to prevent confusion [25]. Another approach is to measure directly the confusability between a pair of words by calculating the similarity between their pronunciations using a phone confusion table. Such measured confusability has been used as a criterion to reject pronunciation variants for a specific word when they are phonetically similar to variants of other words [4]. However, this criterion does not include pronunciation frequencies or word frequencies in the consideration of the confusability. The relationship between the estimated confusability and the recognition performance is also not clear. Yet another approach is to estimate the confusability of individual variants by matching them with time-aligned phone sequences of utterances obtained from forced recognition. This confusability metric has also been used to reject confusing variants. Experiments have shown that the confusability of a dictionary, thus evaluated, is not very closely correlated with recognition performance [16].

In this paper, two different measures of the confusability of a dictionary were explicitly defined based on pronunciations that are shared by at least two distinct words in the dictionary. When pronunciation variants are introduced into a multiple-pronunciation dictionary, the confusability as measured by the proposed metric are inevitably increased. Therefore, a major purpose of this paper is to reduce such extra confusability when constructing a multiple-pronunciation dictionary. Our experiments show that the confusability measured by either of the proposed metrics is very strongly correlated with recognition performance, at least for Mandarin Chinese. The eligibility of a pronunciation variant to be included for a given word was thus estimated by both the prior probability of the variant for the word and the potential of the variant to increase the confusability in the dictionary. Accordingly, the eligibility is less under- or over-estimated than it would be when only pronunciation frequencies or confusability was considered alone. The potential of variants to increase the confusability of a dictionary was further analyzed in various aspects.

Based on the aforementioned strategies, a three-stage framework for constructing a multiple-pronunciation dictionary for Mandarin Chinese was proposed, with a focus on reducing the increased confusability that is caused by adding pronunciation variants. In Stage 1, *pronunciation generation*, an automatic procedure was proposed to generate surface transcriptions with less confusion. In Stage 2, *pronunciation ranking*, the pronunciation variants of each word observed in the surface transcriptions are ranked by their eligibility to be included for the word, considering not only the pronunciation frequencies of the variants, but also their potential to increase confusability. In Stage 3, *pronunciation pruning*, the less eligible (or lower-ranked) pronunciations are pruned by some measure based on the estimated eligibility. The confusability of the constructed dictionaries can be maximally reduced by considering various aspects of the extra confusion added in the three stages. The approaches were tested in large-vocabulary continuous speech recognition (LVCSR) experiments on both planned (LDC HUB-4NE) and spontaneous (LDC CALLHOME) speech in Mandarin Chinese. During the recognition, the pronunciation probability weight used in the process was also tuned to observe its effect on confusability. The interactions among the pronunciation, acoustic and language models were also analyzed with additional experiments. Although the proposed approaches were verified to be useful for Mandarin Chinese, it is not known at all whether they can be used for other languages, because no such tests were ever performed. In particular, very possibly these approaches may not be useful for western alphabetic languages due to the significant difference between the linguistic and phonetic structures of Chinese and western alphabetic languages. However, these concerns are out of the scope of this paper.

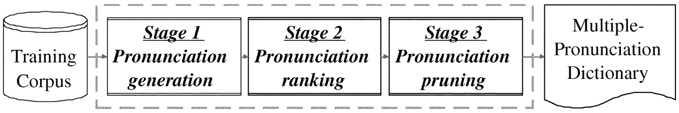


Fig. 1. Three-stage framework for constructing multiple-pronunciation dictionaries.

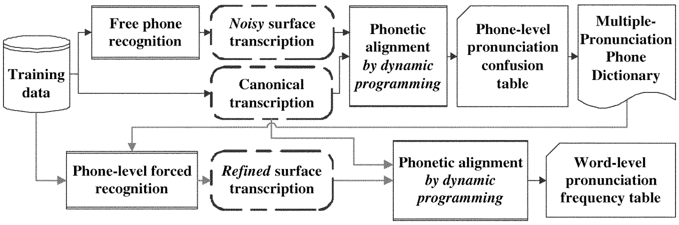

Free phone recognition
Noisy surface transcription
Phonetic alignment by dynamic programming
Phone-level pronunciation confusion table
Multiple-Pronunciation Phone Dictionary
Training data
Canonical transcription
Phone-level forced recognition
Refined surface transcription
Phonetic alignment by dynamic programming
Word-level pronunciation frequency table

Fig. 2. Proposed automatic surface form generation procedure.

The rest of this paper is organized as follows. Section II introduces in detail the three-stage framework. Section III then describes the corpora and experimental configurations. Next, Section IV defines the two proposed confusability measures, and presents, analyzes, and discusses the experimental results to verify the concepts mentioned here. A conclusion is finally drawn in Section V.

II. Three-Stage Framework for Constructing Multiple-Pronunciation Dictionaries

This section presents the three-stage framework for automatically constructing multiple-pronunciation dictionaries for Mandarin Chinese. Fig. 1 presents the framework. In the first stage (*pronunciation generation*), surface transcriptions of words from a training corpus were generated; these generated pronunciations were then ranked for each word by their eligibility for inclusion in the dictionary in the second stage (*pronunciation ranking*). Finally, the higher-ranked pronunciations were selected in the dictionary in the *pronunciation pruning* stage.

A. Stage 1—Pronunciation Generation

An *automatic surface form generation procedure* for Mandarin Chinese [14], [15] was proposed in this stage to generate surface forms. This procedure utilizes phone-level forced recognition following a free phone recognizer, so as to introduce less lexical confusion into the compiled dictionary.

The upper arm in Fig. 2 demonstrates that this procedure firstly utilizes a free phone recognizer (without any constraint of word dictionaries or language models) to produce a "noisy" surface transcription from the corpus. This transcription is then aligned with the canonical transcription of the same corpus by dynamic programming. Phonetic similarity is taken into account as an additional score in the dynamic programming. A phone-level pronunciation confusion table is obtained from the alignment and converted into a multiple-pronunciation phone dictionary in the upper right-hand corner of this figure. It consists of pronunciation variants, with prior probabilities, observed for each phone. This phone dictionary accommodates both phone-level substitutions and deletions. The deleted phones are modeled using a special hidden Markov model with a single state which took only one speech frame, and then the next frame is transited automatically to the next model with a probability of unity. The insertions, however, are simply removed from this phone dictionary for the following reasons. First, most inserted phones observed are not actually produced in the acoustic signal, but are very often simply incorrectly recognized. Second, the number of such inserted phones is relatively very small, as compared to the number of phone deletion or substitution. Third, modeling such inserted phones accurately in a multiple-pronunciation phone dictionary is very difficult.

The phone dictionary obtained above is then used in the following phone-level forced recognition on the same corpus as in the lower arm of Fig. 2. This forced recognition chooses for each individual canonical phone, the pronunciation that best matched the acoustic signal among the variants provided by the dictionary. The "phone-level," instead of "word-level," forced recognition is used here because pronunciation variation can be accommodated much more efficiently at phone-level representation, given the fact that such variation is commonly described by the symbolic change of "phones." For instance, considering a word with five phones in its canonical transcription, four variants for each of the five phones (a total of 20 entries in phone-level representation) may need 1024 variants (or entries) to present the same information in word-level representation.

Additionally, a pronunciation probability weight (denoted as α_F) is used during the phone-level forced recognition, and can be tuned to control how conservatively the surface forms will be transcribed. In other words, the best phone sequence $\hat{S}$ obtained during the forced recognition is the one that maximized the following probability:

$$\hat{S} = \arg\max_{S} \left[P(A|S) P(S|S_c)^{\alpha_{\mathrm{F}}} \right] \tag{1}$$

where S_c is a canonical phone sequence, S is any possible corresponding pronunciation sequence provided in the dictionary, and A is the acoustic signal. Clearly, the larger the α_F is, the more conservatively the surface forms are transcribed, because in this case only those pronunciations with relatively higher prior probabilities (the canonical pronunciations in most cases) would be considered in the forced recognition. Hence, infrequent pronunciation variants are very often rejected, and only those speech segments that are pronounced very frequently and differently from the canonical pronunciations would be transcribed. The resulting "refined" transcription is then aligned with the canonical transcription to obtain the pronunciation variants of words. These variants, excluding those with occurrence frequencies of less than a predetermined threshold, are finally explicitly enumerated in a word-level pronunciation frequency table obtained in the lower right-hand corner of Fig. 2. They are ranked in the following stage.

Notably, in this paper, the variants were obtained explicitly from the transcription alignment rather than implicitly by expanding the baseforms of words, because, as mentioned previously, implicit approaches have been shown to offer comparable [7], [9] or worse [16] recognition performance than explicit approaches. Clearly, the explicit approach is not able to generate pronunciations of unseen words, and should be adopted only when a sufficiently large training corpus is available. As will become clear later on, more than 97% of the word tokens in the test data in this paper appeared in the training set, so choosing an explicit approach to generate the pronunciations herein was reasonable.

Both the free phone recognition and the phone-level forced recognition mentioned previously were performed with a set of intrasyllable right-context-dependent acoustic models which has been found to be very useful for Mandarin Chinese, and the same set of models was also used in the recognition performance tests that are reported below. Although the literature [8] has suggested that monophone models with fewer Gaussian mixtures can be used for the automatic transcription of surface forms for ensuring that the models do not become "overly exposed" to the canonical transcriptions, it was also reported [7] that using the same set of acoustic models in both the automatic transcription of surface forms and the recognition performance tests improved the consistency. Certainly, it is also a good choice to use more sophisticated acoustic models, such as the cross-word triphone models, in the final recognition performance tests. In this case, the use of a separate set of monophone models to obtain pronunciation variants and then expand them into context-dependent triphones for later recognition may be preferred to avoid the explosion of the search space during the forced recognition. On the other hand, the simple use of the same set of acoustic models in the forced recognition with a relatively smaller search beam width may also be considered for better consistency.

In this paper, the quality of the surface transcriptions and the pronunciation variants generated by the approach proposed above, including the use of different pronunciation probability weights α_F in (1) in the forced recognition, was investigated. They were compared with those generated by a free phone recognizer [4] or a phone bigram [2] in terms of the confusability of the constructed dictionaries and the resulting recognition performance.

B. Stage 2—Pronunciation Ranking

In this stage, the pronunciation variants generated in Stage 1 are ranked by their eligibility for inclusion in the dictionary. This paper proposes to use a $pf{-}iwf$ score which not only considers how frequently the pronunciation is realized for the word [the pronunciation frequency (pf)], but also how much extra confusability would be introduced by including this pronunciation in the dictionary [the inverse word frequency (iwf)] [14]. The number of higher-ranked pronunciations to be included in the dictionary is then determined by the pruning criteria in the next stage. The proposed $pf{-}iwf$ score, integrating the pf and the iwf, is analogous to the term frequency (tf) and the inverse document frequency (idf), $tf{-}idf$ scores, for indexing terms used in information retrieval [26], as explained next.

1) Pronunciation Frequency (pf)*:* Conventionally, the eligibility of pronunciation v_j for word w_i is measured by its pf

$$pf_{ij} = \frac{c_{ij}}{\sum_{allj} c_{ij}} = P(v_j|w_i) \tag{2}$$

where c_{ij} is the count of word w_i being pronounced as v_j, and $P(v_j|w_i)$ is the prior probability that w_i is pronounced as v_j. This pf is analogous to the tf in information retrieval, for which an indexing term (or a pronunciation herein) that is observed more frequently in a document (or a word herein) typically implies a higher correlation with the document (or the word).

2) Inverse Word Frequency (iwf)*:* While the pf described above concerns various pronunciations within a word, the iwf proposed herein concerns a pronunciation across different words. A pronunciation that is frequently realized for many different words may introduce extra confusion during recognition, so such a pronunciation is less eligible for inclusion in the dictionary. The iwf for a pronunciation v_j can therefore be defined as

$$iwf_j = \log \frac{|\Omega|}{|\Omega_j|} \tag{3}$$

where Ω is the vocabulary of words, Ω_j is the set of words whose pronunciation variants include v_j, and $| \bullet |$ is the number of elements in a set. This definition is almost identical to that of the inverse document frequency (idf) in information retrieval, in which an indexing term (or a pronunciation herein) that is frequently observed in many different documents (or words herein) typically implies relatively low discriminability in identifying relevant documents (or words). Hence, the importance of an indexing term (or pronunciation) is related to the inverse of its frequency of appearance in different documents (or words), which is the idf (or iwf).

Based on the definition in (3), all of the words with pronunciation v_j are treated equally, but the confusion caused undoubtedly also depends on both the frequencies of the confused words and the probabilities that those words are pronounced as v_j. Therefore, the inverse word frequency $(iwf)$is redefined as follows:

$$iwf_j = \frac{1}{\sum_{\omega_k \in \{\Omega_j\}} P(v_j|w_k)P(w_k)} = \frac{1}{P(v_j)} \tag{4}$$

where $P(w_k)$ and $P(v_j)$ are the prior probabilities of the word w_k and the pronunciation v_j in the corpus. The inverse word frequency for pronunciation v_j, iwf_j is thus higher when v_j is more frequently realized for commonly used words.

3) Pronunciation Frequency and Inverse Word Frequency $(pf{-}iwf)$*:* The $pf{-}iwf$ score, obtained by integrating the pronunciation frequency and inverse word frequency as defined above, is proposed to evaluate the eligibility of a pronunciation v_j to be included for a word w_i in the dictionary

$$\delta_{ij} = pf_{ij} \cdot (iwf_j)^{\gamma} \tag{5}$$

TABLE I
PRONUNCIATION VARIANTS WITH OCCURRENCE FREQUENCIES OF THE WORD "有" ("*have*," WITH CANONICAL PRONUNCIATION */#+i iou/*) RANKED BY pf OR $pf{-}iwf$ SCORES

by *pf* score			by *pf-iwf* score		
(1)	/#+i iou/	670	(1)	/#+i iou/	670
(2)	/ iou/	137	(2)	/ iou/	137
(3)	/#+U iau/	66	(3)	/#+U iou/	20
(4)	/j+i iou/	64	(4)	/#+i ou/	25
(5)	/#+i ou/	25	(5)	/n+u iou/	4
(6)	/#+i iau/	24	(6)	/#+i uo/	11

where γ is an adjustable weight parameter for the iwf score. When γ is set to zero, δ_{ij} is reduced to the original conventionally used pronunciation frequency pf_{ij}. When γ equals unity, δ_{ij} turns out to be the mutual information between pronunciation v_j and word w_i

$$\delta_{ij} = \frac{P(v_j|w_i)}{P(v_j)} = \frac{P(v_j, w_i)}{P(v_j)P(w_i)} = \frac{P(w_i|v_j)}{P(w_i)} \tag{6}$$

which expresses how much the likelihood is increased once the pronunciation v_j is known. In this case, the pronunciations having higher mutual information with a particular word tend to have higher $pf{-}iwf$ scores. The weight γ can also be an arbitrary number other than zero or unity. The $pf{-}iwf$ score δ_{ij} is generally higher if v_j occurs more frequently for the word w_i, and lower if v_j appears more frequently in many other commonly used words.

Table I presents an example of the pronunciation variants (together with the respective frequencies in the training corpus) ranked either by the conventional pf score alone or by the $pf{-}iwf$ score proposed herein (with $\gamma = 0.8$) for a commonly used Chinese word "有" ("have," with canonical pronunciation /#+i iou/, where #+i is an Mandarin Initial1 with right context of phone i and iou is a Mandarin Final[1] that consists of three phones i, o, and u). The left column of this table ranks the variants by the pf alone, and two of the six top-ranked variants are shared by other frequently used words—item (4), /j+i iou/ (the frequently used function word "就", occurring 798 times in the training data) and item (6), /#+i iau/ (another frequently used function word "要" occurring 709 times). However, among the variants ranked by the $pf{-}iwf$ scores in the right column of the table, these two confusing variants are ranked below six and so are not shown in the table. Moreover, even if item (3) in the left column, /#+U iau/, is not identical to any canonical pronunciation of other words, it is ranked below six by the $pf{-}iwf$ score and so is not shown in this column, because some other frequently used words (such as "要" /#+i iau/) have this pronunciation as a noncanonical variant, and so greatly reduced the $pf{-}iwf$ score. In other words, a pronunciation

[1] Conventionally, a Mandarin syllable is decomposed into an Initial and a Final. The Initial is the way in which a syllable begins, normally with a consonant. However, a small number of syllables do not begin with a consonant and are referred to as beginning with a zero Initial. The Final of a syllable is the syllable minus the Initial. The longest Final consists of three parts—an optional medial, or a semivowel; a main vowel, or a head vowel, and an optional ending. Mandarin has a total of 21 Initials and 36 Finals.

is given a lower $pf{-}iwf$ score if it is often realized for some other frequently used words. Furthermore, reductions occur often in frequently used words (with some phones missing). Such partial pronunciations of frequently used words are also given lower $pf{-}iwf$ scores and thus are more likely to be removed from the dictionary.

C. Stage 3—Pronunciation Pruning

With the pronunciation variants as ranked above, the pronunciation pruning stage determines the number $N(w_i)$ of top-ranked variants of each word w_i to be included in the dictionary. Four pruning approaches are presented and will be tested next.

1) Fixed Pruning: Simply keeping a fixed number N_0 of top-ranked variants for each word, or keeping all of a word's variants if the number of available variants is fewer than N_0.

2) Count-Based Pruning: Frequently spoken words typically have widely spread pronunciation variation, so the number of pronunciations to be included for each word $w_i$has been suggested to be determined based on the occurrence count of the word [10]

$$N_C(w_i) = \lfloor \mu_C \cdot \log c_i \rfloor \tag{7}$$

where "$\lfloor \bullet \rfloor$" is a flooring function, or the largest integer less than or equal to the argument, c_i is the occurrence count of the word w_i in the corpus, and μ_C is an empirically tuned parameter.

3) Entropy-Based Pruning: The pronunciation entropy for a word w_i has been shown to be a good measure of the spread or variability of the pronunciations of the word [23], [24]

$$H_i = -\sum_j P(v_j|w_i) \log P(v_j|w_i) \tag{8}$$

so more pronunciations are reasonably included for a word with a higher pronunciation entropy [14]

$$N_E(w_i) = \lfloor \mu_E \cdot H_i \rfloor \tag{9}$$

where μ_E is an empirically tuned parameter.

4) Score-Based Pruning: It has also been proposed that a pronunciation v_j should be included for a word w_i only if the prior probability $P(v_j|w_i)$ is large enough compared to that of the most probable pronunciation of the word [17]

$$P(v_j|w_i) \geq \mu_S \cdot \max_k P(v_k|w_i) \tag{10}$$

where μ_S is an empirically tuned parameter.

This approach is further generalized herein by replacing the probability $P(v_k|w_i)$ in (10) with the $pf{-}iwf$ score δ_{ij} defined in (5), referred to as *score*-based pruning: a pronunciation v_j is included for a word w_i only if the $pf{-}iwf$ score is large enough compared to that of the top-ranked pronunciation of the word

$$\delta_{ij} \geq \mu_S \cdot \max_k \delta_{ik}. \tag{11}$$

Notably, the value of the $pf{-}iwf$ score δ_{ij} of the top-ranked pronunciation for a word w_i differs considerably for words with different pronunciation spreads. Inequality (11) therefore makes better sense than using a single threshold (such as $\delta_{ij} \geq \mu_S$). This finding was also verified in preliminary tests. This generalized form in (11) is reduced to the original form in (10) when the weight parameter γ in (5) is set to zero.

III. Speech Corpora and Experimental Setup

A. Speech Corpora

Two styles of Mandarin Chinese speech, planned and spontaneous, were used to investigate the various approaches presented in this paper.

1) Planned Speech: The LDC 1997 Mandarin Broadcast News corpus (HUB-4NE), comprising 41 hours of recordings, was used as the training set in the planned speech task. Approximately 700K word tokens from the corpus were used to train the word-bigram language model. The task vocabulary comprises 23 779 words, covering all of the words appearing in the training set. The single canonical pronunciations of these words are included in the canonical dictionary. In the canonical dictionary, the 23 779 words are described with 15 334 distinct pronunciations. Of the 23 779 words, 10 455 words are confusable, or have pronunciations that are shared by at lest two words. Therefore, the "intrinsic confusability" of the canonical dictionary is 10 455/23 779 = 44.0%.[2] These 10 455 confusable words are described with 2010 distinct pronunciations.

After discarding those parts of laughters, filled pauses, corruptive background and channel noise, and words in other languages, 28 h of speech data were actually used to train the acoustic models; 1.5 h was used as the development set, and one hour of speech from the 1997 HUB-4 Broadcast News Evaluation Non English Test Material was used as the evaluation data. The acoustic training data covers 58 speakers, 34, 445 utterances, 271 174 word tokens or 452 777 character tokens. The evaluation data covers 28 speakers, 1077 utterances, 8405 word tokens or 13 987 character tokens. The 23 779 words in the canonical dictionary cover 97.8% of the evaluation data. Only pronunciations that occur at least three times in the training data were considered in the approaches proposed, and words with at least one such pronunciation (canonical or not) that occur at least three times cover 84.0% of the evaluation data.

2) Spontaneous Speech: The LDC CALLHOME Mandarin Chinese Speech corpus plus the LDC CALLFRIEND Mandarin Chinese-Mainland Dialect corpus, comprising about 30 h of telephone conversations without prespecified topics, were used as the training set in the spontaneous speech task. Approximately 400K word tokens from the training set were used to train the word-bigram language model. The task vocabulary comprises 14 590 distinct words, covering all of the words in the training data. The single canonical pronunciations of these

[2] All the confusability measures mentioned in this paper consider only *phonetic confusability*, disregarding variants that are differentiated by prosodic features, such as tones in Mandarin Chinese. Clearly, in Mandarin, tones also carry lexical meanings. If tonal information is considered, some of the 10 455 words that share the same pronunciations are differentiated by tones, and the "intrinsic confusability" of the canonical dictionary is reduced from 44.0% to 30.9%.

words are included in the canonical dictionary. In the canonical dictionary, the 14 590 words are described with 7596 distinct pronunciations. Of the 14 590 words, 7947 words are confusable, or have pronunciations that are shared by at lest two words. Therefore, the "intrinsic confusability" is 7947/14 590 $=$ 54.5%.[3] These 7947 confusable words are described with 953 distinct pronunciations.

After discarding those parts of laughters, filled pauses, corruptive background and channel noise, and words in other languages, 18 h of speech data were actually used to train the acoustic models. One hour of data from the CALLHOME development set was used as development data, and 2 h of data from the CALLHOME Evaluation Set were used as the evaluation data. Therefore, this task is referred to as CALLHOME throughout this paper, even though CALLFRIEND was also used in training. The acoustic training data covers 120 speakers, 38 374 utterances, 241 137 word tokens, or 320 550 character tokens. The evaluation data covers 22 speakers, 2672 utterances, 16 771 word tokens, or 23 487 character tokens. The 14 590 words in the canonical dictionary cover 99.0% of the evaluation data. The words with at least one pronunciation that occur at least three times in the acoustic training cover 89.8% of the evaluation data.

B. Experimental Setup

All of the experiments reported in this paper were performed using HTK tools [27]. The HTK toolkit was used to train acoustic and language models and to perform recognition tests or forced recognition. The acoustic models consist of 150 gender-dependent, intrasyllable right-context-dependent HMMs for Mandarin Initial/Finals, or right-context-dependent Initials and context-independent Finals, together with one silence and one short-pause HMM. Each Initial HMM consists of three left-to-right states and each Final HMM consists of four states, each state with 24 Gaussian mixtures. The acoustic features were 12 MFCCs plus energy, their delta and acceleration for 32-ms frames with a 10-ms frame shift. The Initial/Finals play the role of phones in all of the experiments reported below. The same set of acoustic models was used in both acquiring the surface pronunciations from the training data and in performing the recognition performance evaluation. All recognition processes were performed by exhaustive search with an insertion penalty of zero and a language model weight of seven, both of which were empirically tuned on the development set.

In recognition performance evaluation for Chinese, character error rates[4] are commonly used instead of the word error rates usually used for other languages. These character error rates were adopted in this paper, because the segmentation of a Chinese sentence into words is usually not unique due to the lack of word boundary marks (such as the blanks in alphabetic languages) in Chinese.

IV. Experimental Results and Discussion

The planned speech corpus (HUB-4NE) was first employed for primary analysis. Then, the key experiments were repeated and analyzed using the spontaneous speech corpus (CALLHOME).

A. Stage 1—Pronunciation Generation

In this part, surface transcriptions were automatically generated from the training corpus using the following approaches—a free phone[5] recognizer [4] (*Free*), a phone recognizer constrained by a phone[4] bigram [2] (*Bigram*) and the proposed automatic surface form generation procedure, as shown in Fig. 2, and using different values of the pronunciation probability weight, α_F, in the forced recognition as in (1) (*Proposed*).

The degree of phonetic discrepancy (actually Initial/Final discrepancy ratio) of the surface transcriptions generated with the respective approaches mentioned above to the canonical transcription is depicted as the solid curve on the first right scale of Fig. 3. This curve shows that the transcription generated by free phone recognition (*Free*) differs most from the canonical transcription, but using a phone bigram for transcription (*Bigram*) can greatly reduce this discrepancy. The *Proposed* transcription approach, on the other hand, not only further reduces this discrepancy but can also properly adjust the discrepancy by tuning the value of α_F, which actually specifies the degree of the constraints imposed in the phone-level forced recognition process. Using a larger value of α_F implies that the transcription is generated more conservatively, and resulting in a lower discrepancy ratio.

Although the discrepancy ratio reveals how conservatively the transcriptions were generated, it does not necessarily indicate the quality of the transcriptions for recognition purposes. To analyze the latter, an additional set of experiments was conducted with the multiple-pronunciation dictionaries that had been constructed from these transcriptions. Each of these dictionaries had been compiled in exactly the same way, using the conventionally used pronunciation frequency (pf) score (or $\gamma = 0$) in Stage 2 to rank the pronunciation variants obtained from these surface transcriptions, and using *score*-based pruning criterion in Stage 3 to determine the number of pronunciations included for each word in the dictionary.[6] For a fair comparison, the parameter μ_S in inequality (11) was empirically adjusted, resulting in an empirical average of 1.14 pronunciations per word, or about 3300 pronunciation variants were added to each dictionary.

The recognition performance of the compiled dictionaries in terms of *Character Error Rate* (CER) is illustrated as the vertical bars on the left scale in Fig. 3, together with the baseline character error rate of the canonical dictionary (37.02%) for comparison. These bars show that the dictionaries obtained from the *Proposed* transcriptions results in lower recognition

[3] The "intrinsic confusability" is reduced to 46.1% when the tonal information is considered.

[4] A Chinese word is composed of one to several characters. Most of characters are morphemes (or monocharacter words). All Chinese characters are always pronounced as monosyllables.

[5] The notation *phone* is kept for the generality of presentation, although the Mandarin Initial/Finals were actually used in all of the experiments playing the role of phones.

[6] Other values of γ will be considered below when Stage 2 is analyzed. The *score*-based criterion is chosen here for Stage 3 because it outperforms the other criteria, as will also be showed later on.

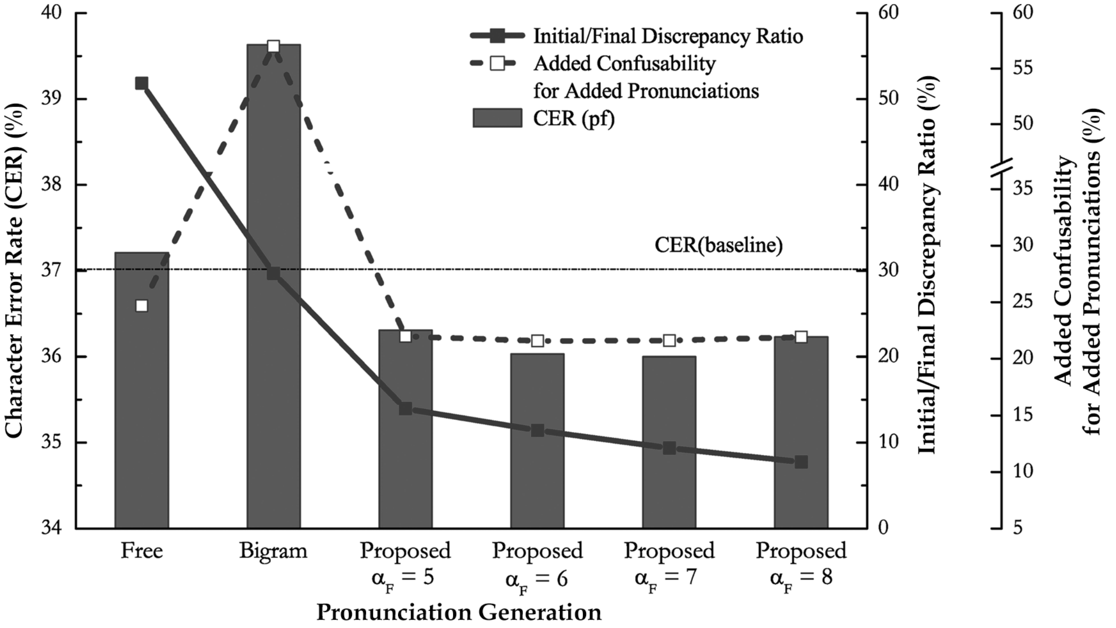


Fig. 3. Initial/final discrepancy ratio (solid curve) between the surface and canonical transcriptions on the first right scale, character error rates (CERs), (vertical bars) on the left scale, and the added confusability for added pronunciations (dotted curve) on the second right scale obtained with the different versions of transcriptions.

error rates than those from the *Free* and the *Bigram* transcriptions and the baseline (p-value[7] <0.001). Notable, the *Free* and *Bigram* transcriptions could not improve the recognition performance from the baseline. Another interesting finding is that the *Bigram* transcription gives much poorer recognition performance than the *Free* transcription, despite the lower Initial/Final discrepancy ratio of the *Bigram*. This observation is inconsistent with that reported for tasks in English [2], probably because of the special characteristics of the Chinese language. Mandarin Chinese has only about 400 phonologically allowed syllables disregarding tones, which restrains the canonical combinations of Initials and Finals. Therefore, using an Initial/Final bigram trained with canonical transcriptions to generate surface forms tends to confine most of the surface transcriptions to the very limited number of allowed phonological patterns. Consequently, adding these canonical combinations of Initial/Finals, which are very often shared by other words, introduces a great lexical confusion in the dictionary, leading to poor recognition performance. This assertion is confirmed below by explicitly evaluating the confusability of the dictionaries.

As the conclusion of this subsection, the *Proposed* approach does generate better dictionaries giving relatively lower character error rates. However, although the phonetic discrepancy between the generated surface transcriptions and the canonical transcription indicates how conservatively the transcription is generated, this ratio is not necessarily correlated with the recognition performance. Better measures of the confusability of the dictionaries are thus needed, as will be discussed in the next part.

B. Confusability Measures for the Dictionaries

Fig. 3 indicates that the recognition performance of the compiled dictionaries apparently is not closely correlated with the phonetic discrepancy of the corresponding surface transcriptions to the canonical transcription. In order to find better measures for the dictionary quality that are more correlated to the recognition performance, five sets of statistical measures each for the different dictionaries discussed above are listed in the five rows in Table II. These are: (1) the percentage of words that retain their canonical pronunciations; (2) the percentage of words that include noncanonical pronunciations; (3) the percentage of words that have at least two pronunciations; (4) the percentage of words that have confusing pronunciations (shared by at least two distinct words); and (5) the percentage of added pronunciations that are confusing (shared by at least two distinct words). For comparison, the last row in this table also lists the recognition performance (CER) of the dictionaries as shown in Fig. 3. Table II shows that the *Proposed* dictionaries ($\alpha_F = 5, 6$, 7, and 8) have more words that retain their canonical pronunciations [row (1)] than the *Free* and the *Bigram*, apparently because of the lower phonetic discrepancies with the canonical transcription. On the other hand, the *Free* dictionary has the smallest percentages of words that have non-canonical [row (2)] or at least two pronunciations [row (3)], since for the *Free* transcription, which has many transcription errors, only those words that appear frequently enough would have surface pronunciations with sufficient frequencies and thus could be well transcribed.

A more important observation, however, is that both of the two measures in rows (4) and (5) in Table II are found to be strongly correlated with the recognition performance in row (6) (with correlation coefficients[8] $R = 0.952$ and 0.977, respec-

[7] The level of significance, the p-value, is based on hypothesis testing and defined as the risk or probability of rejecting the null hypothesis when it is in fact true. Differences are considered significant if the p-value is less than 0.05. All p-values reported in this paper were based on the standard Wilcoxon Signed-Rank Test [28]

[8] The correlation coefficient is also known as the product-moment coefficient of correlation or Pearson's correlation.

TABLE II
STATISTICAL MEASURES OF THE DICIONARIES COMPILED WITH pf SCORE

		Can. Dic.	*Free*	*Bigram*	*Proposed*			
					α_F=5	α_F=6	α_F=7	α_F=8
(1)	percentages of words retaining canonical pronunciations (%)	100.0	98.9	98.3	99.6	99.7	99.7	99.8
(2)	percentages of words with non-canonical pronunciations (%)	0.0	7.3	9.2	8.9	8.9	8.8	8.7
(3)	percentages of words with at least two pronunciations (%)	0.0	6.7	8.3	8.6	8.7	8.6	8.5
(4)	percentages of words with confusing pronunciations (%) – "confusability of a dictionary"	44.0	47.8	54.4	47.8	47.7	47.7	47.8
(5)	percentages of added pronunciations which are confusing (%) – "added confusability for added pronunciations"	–	24.7	57.0	22.0	21.6	21.6	21.9
(6)	*Character Error Rate* (%)	37.02	37.21	39.63	36.31	36.03	36.00	36.23

tively). The measure in row (4) is thus referred to as the "confusability of a dictionary." It has the advantage that it is actually reduced to the "intrinsic confusability" of a canonical dictionary, as mentioned in Section III. However, this measure counts the percentage of words that are confusing, and so it may underestimate the "confusability of a dictionary" if serious confusion occurs among a few very frequently used words with many shared variants. This is actually the case for the *Free* dictionary here. The measure in row (5), however, counts the percentage of added pronunciations that are confusing, and thus avoids the above problem. This fact may explain why the correlation coefficient with the recognition performance of the measure in row (5) (0.977) is higher than that in row (4) (0.952). The measure in row (5) is thus referred to as the "added confusability for added pronunciations." Fig. 3 plots the "added confusability for added pronunciations" as the dotted curve on the second right scale. This figure clearly shows the very high correlation of this measure with the character error rate (vertical bars on the left scale). On the other hand, rows (4) and (5) in Table II indicate that the *Bigram* dictionary has much higher confusability than the others dictionaries, confirming the previous explanation that the bigram constraints of Mandarin Initial/Finals introduce more confusion and hence degrade the recognition performance. The dictionary with $\alpha_F = 8$ has a slightly higher confusability and character error rate than that with $\alpha_F = 6$ or 7 in rows (4)–(6) (p-value $= 0.04$). The higher confusability of the dictionary with $\alpha_F = 8$ may follow from the fact that its transcription was generated so conservatively that many less frequently used words are simply transcribed in their canonical forms. Consequently, most of the added pronunciations are variants of frequently used words including those ranked lower, and therefore are more likely to increase confusability.

As the conclusion of this section, the two confusability measures defined properly indicate the quality of a dictionary and are shown to be strongly correlated with the recognition performance of a dictionary.

C. Stage 2—Pronunciation Ranking

The multiple-pronunciation dictionaries investigated above were all compiled using the pf score alone [or $\gamma = 0$ in inequality (5)] to rank the pronunciation variants obtained from the surface transcriptions. Another set of dictionaries was compiled in exactly the same way, except in that the pronunciation variants were ranked by the $pf - iwf$ score, as the weight γ for the iwf score in inequality (5) is varied from 0 to 0.9. Fig. 4 presents the resulting recognition performance of these dictionaries in terms of character error rates. The left-most bars labeled "$\gamma = 0$" correspond to the vertical bars in Fig. 3, for the dictionaries compiled using the pf score alone, while the other bars are for $\gamma = 0.5$ up to $\gamma = 0.9$. A trend can be observed from Fig. 4 that γ values that exceed zero generally improve recognition performance by reducing the confusion. However, the performance may also be degraded when the γ values are too large, such that the pronunciation frequency (pf) is excessively de-emphasized. In most cases, γ values from 0.6 to 0.8 give the best recognition performance. Notably, the differences among the character error rates in the many cases in Fig. 4 may not be statistically significant and so it may not make much sense to consider the recognition performance for this specific given task alone. However, the relationships and the trend implied by these results are actually clear, and believed to be valuable as references in considering other recognition tasks. For simplicity, $\gamma = 0.8$ will be used in compiling all of the dictionaries in the following discussions. Of course, in practice, the value of γ should be tuned on a proper development set, as will be done later in the final recognition tests.

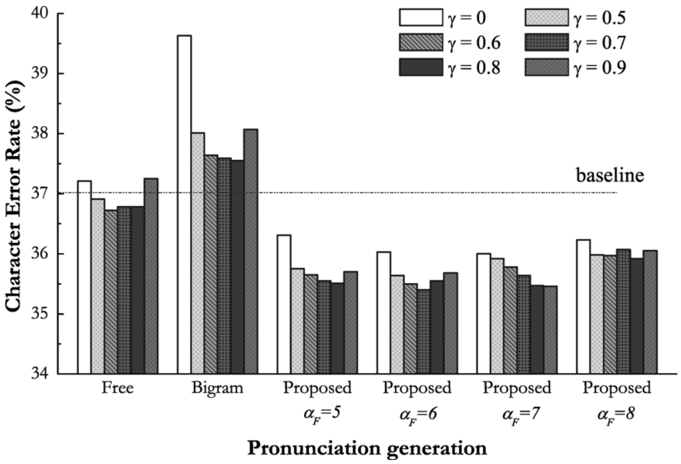


Fig. 4. CERs for the dictionaries obtained from different versions of transcriptions and, for each version of the transcriptions, using the $pf - iwf$ score with different values of γ ($\gamma = 0$ equivalent to the pf score alone).

TABLE III
STATISTICAL MEASURES OF THE DICIONARIES COMPILED WITH $pf{-}iwf$ ($\gamma = 0.8$) SCORE

		Can. Dic.	*Free*	*Bigram*	*Proposed*			
					α_F=5	α_F=6	α_F=7	α_F=8
(1)	percentages of words retaining canonical pronunciations (%)	100.0	99.2	98.1	99.2	99.4	99.5	99.6
(2)	percentages of words with non-canonical pronunciations (%)	0.0	5.3	8.5	8.4	8.4	8.2	8.1
(3)	percentages of words with at least two pronunciations (%)	0.0	4.7	7.0	7.8	7.9	7.8	7.8
(4)	percentages of words with confusing pronunciations (%) – "confusability of a dictionary"	44.0	45.7	48.5	44.8	45.0	44.9	45.2
(5)	percentages of added pronunciations which are confusing (%) – "added confusability for added pronunciations"	–	11.3	22.5	6.8	7.3	6.5	7.4
(6)	*Character Error Rate* (%)	37.02	36.78	37.55	35.51	35.55	35.47	35.92

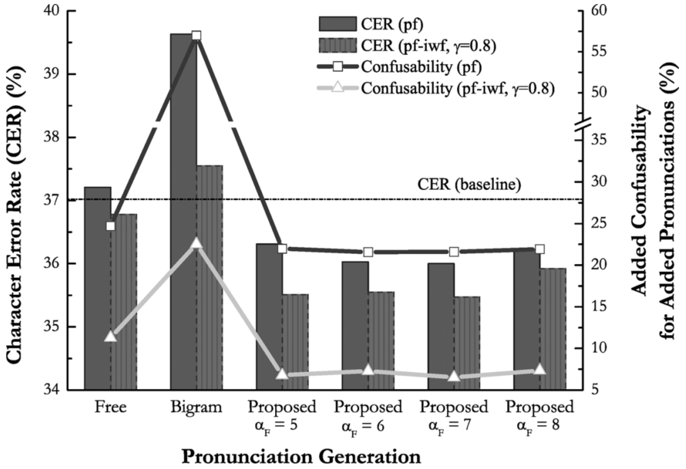


Fig. 5. Confusability (curves) on the right scale and CERs (bars) on the left scale for the dicitonaries obtained with pf (upper curve and left bars) and with $pf{-}iwf$ (lower curve and right bars) scores, respectively.

Table III presents exactly the same information of dictionaries, as is presented in Table II, for the newly compiled dictionaries with $\gamma = 0.8$. From this table, again a strong correlation between the recognition performance in row (6) and the two confusability measures in rows (4) and (5) can be found (with correlation coefficient $\mathrm{R} = 0.922$ and 0.931, respectively). In particular the "added confusability for added pronunciations" in row (5) is very strongly correlated with the recognition performance. The confusability measure in row (5) and the recognition performance in row (6) of Table III are therefore plotted as the lower curve on the right scale and the right bars on the left scale in Fig. 5. These are compared to the upper curve and left bars for $\gamma = 0$ (labeled with "pf"), as discussed above and copied from Fig. 3. Fig. 5 shows that the correlation between the bars (recognition performance) and the curves (the confusability) is evident for both sets of dictionaries compiled using pf $(\gamma = 0)$ and the $pf{-}iwf$ scores $(\gamma = 0.8)$. Note that although the "confusability" measured may substantially affect the recognition results, it is certainly not the only dominator for the recognition performance. The results in Fig. 5 also show that both the confusability and the character error rates are significantly reduced by using the $pf{-}iwf$ score as compared to using the pf score alone (such as for $\alpha_F = 8$, p-value $= 0.005$). The improvements provided by the $pf{-}iwf$ score are most remarkable in the case of the *Bigram* transcription, which has the highest confusability. Furthermore, the lowest confusability and character error rates are obtained when the proposed pronunciation generation procedure was used in Stage 1 and the $pf{-}iwf$ was used in Stage 2.

The proposed approach using $pf{-}iwf$ score with $\gamma = 0.8$, as discussed above, was also compared to an approach based on another criterion, previously proposed [4], to reduce the confusion directly. In the latter approach, the variants for a specific word are rejected when they are phonetically similar to the variants of any other words, without considering the corresponding pronunciation frequencies and word frequencies. Based on exactly the same surface transcription (*Proposed* with $\alpha_F = 5$), the parameters used in compiling both dictionaries were empirically tuned to have the same average of 1.14 pronunciations per word. The experimental results show that the $pf{-}iwf$ score yields both lower "added confusability for added pronunciations" (6.8% versus 10.4%) and lower character error rates (35.5% versus 36.1%, with p-value <0.001).

As the conclusion of this subsection, the $pf{-}iwf$ score ranking with an optimized value of the parameter γ in Stage 2 can both reduce the confusability as measured and improve the recognition performance. Moreover, the two proposed confusability measures again show a very strong correlation with recognition performance.

D. Effect of Weighting the Pronunciation Probabilities in Recognition

During the recognition process, the *a posteriori* probability for a word sequence W is maximized

$$\hat{W} = \arg\max_{W} \max_{V} \left[P(A|V)P(V|W)^{\alpha_R}P(W)^{\beta}\right] \qquad (12)$$

where A is the acoustic signal and V is the pronunciation sequence, and α_R and β are the weight parameters for the pronunciation probabilities and the language model scores, respectively. In all of the experiments reported previously, $\alpha_R = 1$ was used or the pronunciation probabilities were not specially

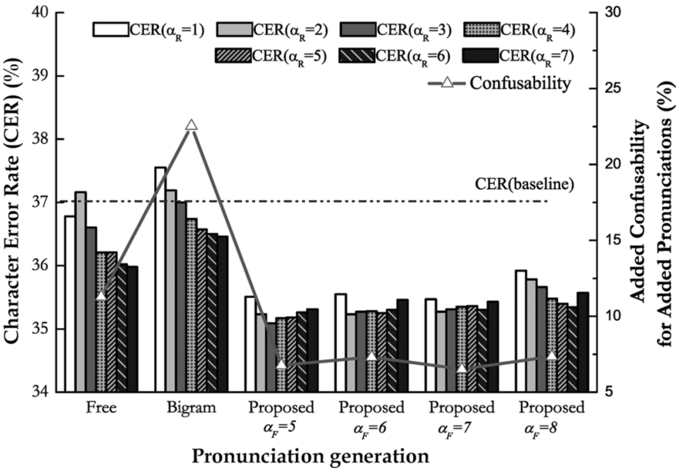


Fig. 6. Confusability (curve) on the right scale for the different versions of dictionaries and, for each version of the dictionaries, the CER (bars) on the left scale obtained with different values of pronunciation probability weight (α_R from 1 to 7) in the recognition process.

weighted. However, in fact, this weight parameter can be adjusted depending on the importance of the pronunciation probabilities considered in the recognition. Therefore, this subsection investigates the influence of different values of α_R, and its interactive relationship with the confusability of the dictionaries.

The investigation employs exactly the same set of dictionaries compiled using the $pf{-}iwf$ score ($\gamma = 0.8$) in Fig. 5. The vertical bars on the left scale in Fig. 6 plot the character error rates of these dictionaries, but with the value of the weight parameter α_R from 1 to 7. The left-most bars for $\alpha_R = 1$ in Fig. 6 are exactly those labeled as $\mathrm{CER}(pf{-}iwf, \gamma = 0.8)$ in Fig. 5. The "added confusability for added pronunciations" of these dictionaries, represented by the lower curve in Fig. 5, is also plotted in this figure for reference. Fig. 6 reveals a trend that $\alpha_R > 1$ generally offers better recognition performance than $\alpha_R = 1$ (in almost all cases). This is reasonable. The less reliable pronunciation variants usually possess lower probabilities in the dictionary, so increasing the weight can help differentiate them from the reliable variants. This effect is therefore more significant for dictionaries with higher "added confusability for added pronunciations," such as those obtained from the "*Free*," the "*Bigram*" or the *Proposed* $(\alpha_F = 8)$ transcriptions. However, when the value of α_R is too large, those variants associated with lower probabilities may be excessively suppressed and the recognition performance may be degraded. Therefore, each dictionary in Fig. 6 has a favorable value of α_R. However, the achieved recognition performance of dictionaries with lower "added confusability for added pronunciations" (*Proposed* with $\alpha_F = 5, 6$, and 7) are less influenced by the value of α_R. Again, the differences among the character error rates in the many cases of the *Proposed* approach with $\alpha_F = 5$, 6, 7, and 8 in Fig. 6 may not be statistically significant, and so it may not make much sense to consider recognition performance for this specific given task alone. However, the relationships and the trend implied by these results are actually clear, and believed to be valuable as references for other recognition tasks.

As the conclusion of this section, an optimized value of the pronunciation probability weight α_R in the recognition process can improve the recognition performance by properly reducing the confusion during recognition. Of course, in practice, the value of α_R should be tuned on a proper development set, as will be done later in the final recognition tests.

E. Stage 3—Pronunciation Pruning

All of the dictionaries mentioned above were compiled using the *score*-based method in Stage 3 of pronunciation pruning. In this section, all four of the pruning criteria discussed in Section II-C—the fixed, count-, entropy-, and score-based, were compared. This comparison was made with four dictionaries that were compiled using the *Proposed* pronunciation generation approach with $\alpha_F = 5$ in Stage 1, the $pf{-}iwf$ score with $\gamma = 0.8$ in Stage 2, and one of the four pruning criteria for pruning pronunciation variants in Stage 3.

Table IV lists exactly the same set of information about dictionaries, as those in Table II and III, for the four dictionaries obtained here. In Table IV, comparison with the *score*-based method shows that the dictionaries compiled with the other three criteria not only has fewer words with non-canonical pronunciations [row (2)] or with more than one pronunciation [row (3)], but also exhibits much higher confusability as measured in rows (4) and (5) and much worse recognition performance (with $\alpha_R = 1$) in row (6).

Properly adjusting the weight parameter α_R for the pronunciation probability in the recognition process may help in differentiating the confusing pronunciations, as was mentioned above. Therefore, the pruning criteria discussed are reasonably believed to have to do with the choice of the value of α_R. Another set of experiments was thus performed on the four dictionaries, obtained with different pruning criteria as discussed above, with α_R ranging from 1 to 9 for recognition. The resulting recognition performance is illustrated by the four solid curves in Fig. 7. This figure shows that the score-based method outperforms the other three, at lower values of α_R from 1 to 3 ($\mathrm{p{-}value} < 0.001$), particularly at $\alpha_R = 1$. However, the relatively poor recognition performance by the other three criteria can be alleviated to a great extent using an appropriate value of α_R during the recognition process to differentiate the confusing pronunciations. In fact, higher α_R values would make the various dictionaries more similar. Therefore, when sufficiently large values of α_R are employed in the recognition (such as $\alpha_R = 7$ or 9), the performance obtained using the other three criteria become more comparable to each other and closer to the score-based performance. However, Fig. 7 also shows that the best performance is only achieved using the score-based criterion with $\alpha_R = 3$ or 5. Therefore, a good pruning criterion is still important, even if the choice of α_R is helpful. Moreover, the score-based criterion is the least influenced by the values of α_R. Also, simply increasing the value of α_R is not always helpful, as shown in Fig. 7. The extreme case of using an excessively high value of α_R corresponds to using a single-pronunciation dictionary that contains only the top-ranked variant (with the highest $pf{-}iwf$ score) of each word for the recognition. An additional experiment was conducted with such a specially designed single-pronunciation dictionary, which included 380 words with noncanonical pronunciations, 42 of which are confusing, being shared by at least

TABLE IV
STATISTICAL MEASURES OF THE DICTIONARIES COMPILED WITH DIFFERENT PRUNING CRITERIA

		Can.	*Proposed,* $\alpha_F = 5$			
		Dic.	*Fixed*	*Count*	*Entropy*	*Score*
(1)	percentages of words retaining canonical pronunciations (%)	100	99.3	99.3	99.3	99.2
(2)	percentages of words with non-canonical pronunciations (%)	0.0	7.8	7.8	7.8	8.4
(3)	percentages of words with at least two pronunciations (%)	0.0	7.2	7.2	7.2	7.8
(4)	percentages of words with confusing pronunciations (%) – "confusability of a dictionary"	44.0	48.5	48.3	47.9	44.8
(5)	percentages of added pronunciations which are confusing (%) – "added confusability for added pronunciations"	–	26.2	26.4	22.1	6.8
(6)	*Character Error Rate* (%) with $\alpha_R = 1$	37.02	37.96	38.01	37.10	35.51

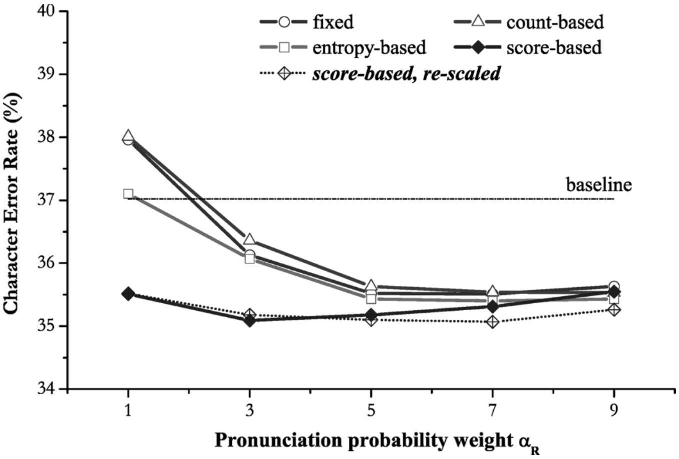


Fig. 7. Recognition performance with different pronunciation pruning criteria in Stage 3 and with re-scaled pronunciation probabilities, using different pronunciation probability weights α_R.

two distinct words. This dictionary did not result in significant change in the character error rate (37.37% versus 37.02% for the baseline canonical dictionary, p-value $= 0.08$). This result indicates that the pronunciations further to the top-ranked pronunciations are important in improving the recognition performance over that of the canonical dictionary.

As the conclusion of this section, the *score*-based pruning criterion outperforms the other three criteria in both reducing the confusability as measured and improving the recognition performance. It is also the least influenced by the pronunciation probability weight α_R. Using an optimized value of α_R in recognition improves recognition performance and makes these criteria more comparable to each other.

F. Scaling the Pronunciation Probabilities

Summing the likelihoods associated with multiple pronunciations during the recognition is difficult, so a maximum approximation is usually employed in the Viterbi search. It has been pointed out [2], [9] that such an approximation may penalize a word that has multiple pronunciations by splitting its pronunciation probabilities among different pronunciations during the recognition, in case the pronunciation probabilities in the dictionary are normalized to have a sum of unity for each word. It has been proposed that the pronunciation probabilities of each word can be re-scaled so that the top-ranked pronunciation has a probability of unity to solve this problem [2], [9]. This rescaling scheme was applied in the dictionary that had been compiled using the score-based criterion, labeled "*score-based*" in Fig. 7. The recognition performance of this re-scaled dictionary is plotted as the dotted curve labeled "*score-based, re-scaled*" in Fig. 7. Fig. 7 shows that this re-scaling approach causes slight improvements when larger α_R values are used in recognition (for example, $\mathrm{p{-}value} = 0.03$ for $\alpha_R = 5$), apparently because a larger value of α_R emphasizes the penalty of splitting the probabilities among different variants.

TABLE V
LOG LIKELIHOODS AND CERS TO ANALYZE THE INTERACTIONS AMONG ACOUSTIC, PRONUNCIATION, AND LANGUAGE MODELS

Acoustic Models (AM)				AM_c		AM_r
Dictionaries				*CanD*	*MPD*	*MPD*
Forced Rec.	training data	(a)	LL_{AM}	-65.44	-65.35	-65.28
	test data	(b)	LL_{AM}	-67.15	-67.06	-67.02
Rec.	test data	(c)	LL_{AM}	-66.89	-66.86	-66.81
		(d)	LL_{LM}	-44.64	-44.15	-44.11
		(e)	CER(%)	37.02	35.10	35.46

G. Interactions Among Acoustic, Pronunciation, and Language Models

Pronunciation models can be considered as the interface between acoustic and language models, so recognition performance undoubtedly depends on the interactions among these models. To analyze the influence of such interactions on the recognition performance, Table V presents averaged log likelihood (decimal logarithms) of the acoustic models (LL_{AM}) or the word-bigram language model (LL_{LM}) in forced recognition $(\mathrm{Forced\ Rec.})$ on the training and test data [rows (a) and (b)] and in the recognition $(\mathrm{Rec.})$ on the test data [rows (c) and (d)]. The CER in row (e) is used as the recognition performance indicator. The forced recognition or recognition employed either the canonical acoustic models (AM_c) that were trained with canonical transcription (used in all the experiments above) or the retrained acoustic models (AM_r) trained with the surface transcription of the same training data. Two dictionaries were used in this analysis: the canonical dictionary ($CanD$, used above in the baseline experiment), and the multiple-pronunciation dictionary (MPD) that was compiled previously using the *Proposed* procedure in Stage 1, the $pf-iwf$ score in Stage 2, and score-based pruning in Stage 3.

The results in the first two rows (a) and (b) in Table V show that, in forced recognition on the training and test data, MPD increases the acoustic likelihood LL_{AM} over that of the $CanD$ (from -65.44 to -65.35 and from -67.15 to -67.06). LL_{AM} is further increased by the retrained acoustic models AM_r (from -65.35 to -65.28 and from -67.06 to 67.02). These findings imply that both the pronunciation variants offered by the multiple-pronunciation dictionary and the retrained acoustic models better match the acoustic signals than the canonical pronunciations and models. In contrast, the LL_{AM} increment made by the MPD and the AM_r in the recognition on test data in row (c) is relatively limited (from -66.89 to -66.86 and to -66.81). This result may be explained as follows. Unlike in forced recognition, in which only the pronunciations of a particular word can be chosen for matching a particular acoustic segment, the recognizer in recognition can chose among the pronunciations of different words that are permitted by the language model. In other words, the recognizer can find some other word hypotheses whose pronunciations give a higher acoustic likelihood than the reference word in this case, even with a canonical dictionary.

Unlike the acoustic likelihood LL_{AM}, the language model likelihood LL_{LM} in recognition, as shown in row (d), is considerably increased by the MPD (from -44.64 to -44.15), since the pronunciation variants enable the language model to retrieve the words that are more likely. Therefore, the character error rate, as shown in row (e), could be significantly improved (from 37.02% to 35.10%). On the other hand, the retrained acoustic models (AM_r) only slightly improve both the LL_{AM} and the LL_{LM} (from -66.86 to -66.81 and from -44.15 to -44.11, respectively, in the last column). The retrained models AM_r even degrade the recognition performance from that of the canonical acoustic models AM_c (from 35.10% to 35.46%, p-value $= 0.04$), which finding is consistent with other works [9], [29] on the very limited or even the negative improvement achievable by retraining the acoustic models.

TABLE VI
CERs (%) With Different Multiple-Pronunciation Dictionaries With Different Numbers of Variants and “Fair” and “Cheating” Language Models

	language models	$CanD$	MPD (1.14)	MPD (1.25)
CER(%) (Rec. on test data)	LM_{fr}	37.02	35.10	35.32
	LM_{ch}	14.24	12.27	11.58

In order to further look into the interaction between the pronunciation model and the language model, another experiment was conducted using a cheating language model LM_{ct} that was trained with the test data and was regarded more adequate in recognition. A comparison was made with the fair language model LM_{fr} that was trained with the training data and was used in all of the above experiments. In addition to using the multiple-pronunciation dictionary $MPD(1.14)$ used above (with an average of 1.14 pronunciations per word), the experiment also employed a different multiple-pronunciation dictionary $MPD(1.25)$ which includes more pronunciations (with an arbitrarily predetermined average of 1.25 pronunciations per word), constructed using exactly the same approaches as was $MPD(1.14)$. Table VI presents the recognition performance of the different language models and dictionaries in terms of character error rates. The first row in this table shows that including more pronunciations in the $MPD(1.25)$ dictionary results in insignificant change in the recognition performance, as compared to $MPD(1.14)$, when the fair language model LM_{fr} is used (from 35.10% to 35.32% with p-value $= 0.5$). However, when the cheating language model LM_{ch} is used, the $MPD(1.25)$ dictionary significantly decreases the character error rate (from 12.27% to 11.58%, p-value < 0.001), apparently because the LM_{ct} better distinguishes among the confusing words caused by the extra pronunciation variants in the test data. In other words, the additional variants offer more opportunities to better match the acoustic signal with correct as well as with incorrect words, depending on the strength of the language model. A language model that can tell the linguistic context more precisely can more effectively prevent the lexical confusion introduced by the variants that appear in different linguistic contexts.

The above analysis shows that both the pronunciation variants included in the dictionary and the retrained acoustic models match the acoustic signals better than do the canonical pronunciations and models. The pronunciation variants offer more opportunities for the recognizer to retrieve both correct and incorrect words during the recognition, depending on the strength of the language model. A strong language model that can identify the linguistic context more precisely may be able to prevent the introduced confusion or incorrect words.

H. Summary With Further Analysis

This section summaries the key experimental results discussed above for the planned speech task (HUB-4NE).

Table VII lists the results for the planned speech recognition task performed with the dictionaries that were compiled using the key approaches discussed previously. The recognition performance is represented by the CERs in this table. The “confusability of a dictionary (*Conf. of Dic.*)” and the “added confusability for added pronunciations (*Added Conf.*)” in rows (4) and (5) of the previous tables are also listed here. Moreover, the last three columns of the table list the numbers of characters that have previously been incorrectly recognized with the canonical dictionary but are correctly recognized with the compiled dictionary (*No. Char. Corrected*), the numbers of characters that have previously been recognized correctly with the canonical dictionary but incorrectly recognized with the compiled dictionary (*No. Char. Incorr.*), and the number of extra character insertion errors over those obtained using the canonical dictionary (*No. Extra Char. Ins.*), respectively.

Row (a) in Table VII shows the results obtained using the canonical dictionary baseline. Row (b) presents those obtained using the dictionary that was constructed using free phone recognition (*Free*) for Stage 1 and the pf score for Stage 2, and by imposing $\alpha_R = 1$ in the recognition. All experiments

TABLE VII
RECOGNITION RESULTS FOR THE PLANNED SPEECH TASK (HUB-4NE)

	Stage 1 :	*Stage 2 :*			*Conf.*	*Added*	*No.*	*No.*	*No.*
	Pron.	*Pron.*	*Weight*	*CER*	*of Dic.*	*Conf.*	*Char.*	*Char.*	*Extra*
	Gen.	*Rank.*	*in Rec.*	*(%)*	*(%)*	*(%)*	*Corrected*	*Incorr.*	*Char. Ins.*
(a)	*Baseline*: canonical dictionary		$\alpha_R = 1$	37.02	44.0	–	0	0	0
(b)	*Free*	*pf*	$\alpha_R = 1$	37.21	47.8	24.7	553	457	123
(c)	*Proposed*($\alpha_F = 5$)	*pf*	$\alpha_R = 1$	36.21	47.8	22.0	504	410	20
(d)	*Proposed*($\alpha_F = 5$)	*pf-iwf*	$\alpha_R = 1$	35.51	44.8	6.8	553	360	4
(e)	*Proposed*($\alpha_F = 5$)	*pf-iwf*	$\alpha_R = 5$	35.18	44.8	6.8	464	221	-8

*Score-based pruning was used in Stage 3 in rows (b)-(e).

TABLE VIII
RECOGNITION RESULTS FOR THE SPONTANEOUS SPEECH TASK (CALLHOME)

	Stage 1 :	*Stage 2 :*			*Conf.*	*Added*	*No.*	*No.*	*No.*
	Pron.	*Pron.*	*Weight*	*CER*	*of Dic.*	*Conf.*	*Char.*	*Char.*	*Extra*
	Gen.	*Rank.*	*in Rec.*	*(%)*	*(%)*	*(%)*	*Corrected*	*Incorr.*	*Char. Ins.*
(a)	*Baseline*: canonical dictionary		$\alpha_R = 1$	60.88	54.5	–	0	0	0
(b)	*Free*	*pf*	$\alpha_R = 1$	64.65	62.8	50.8	2080	1896	1068
(c)	*Proposed*($\alpha_F = 6$)	*pf*	$\alpha_R = 1$	61.41	60.8	35.8	1748	1330	542
(d)	*Proposed*($\alpha_F = 6$)	*pf-iwf*	$\alpha_R = 1$	60.07	55.9	9.3	1813	1321	296
(e)	*Proposed*($\alpha_F = 6$)	*pf-iwf*	$\alpha_R = 5$	58.96	55.9	9.3	1339	853	61

*Score-based pruning was used in Stage 3 in rows (b)-(e).

summarized used score-based pruning for Stage 3. In row (b), 553 incorrectly recognized characters are corrected with reference to the canonical dictionary baseline, but 457 previously correctly recognized characters are recognized incorrectly with the compiled dictionary. The overall character error rate is slightly increased (from 37.02% to 37.21%, p-value $= 0.05$) because of an additional 123 insertion errors, mostly caused by the reduced forms that are taken from the transcription errors as the pronunciation variants. Then, the *Proposed* procedure with $\alpha_F = 5$ was used in Stage 1 in row (c), the $pf{-}iwf$ score with $\gamma = 0.8$ was used in Stage 2 in row (d), and $\alpha_R = 5$ was used in the recognition in row (e). Unlike in the experiments reported previously, these parameters were tuned empirically using the development set, but the values of these parameters obtained with the development set are almost identical to those found previously from the test set. The character error rates monotonically decrease from 37.02% in row (a) to 35.18% in row (e), while the "the confusability of a dictionary (*Conf. of Dic.*)" and the "added confusability for added pronunciations (*Added Conf.*)" of the compiled dictionaries also decrease from 47.8% and 24.7% in row (b) to 44.8% and 6.8% in row (d). The data in row (e) are obtained using exactly the same dictionary that is used in row (d), except in that a different value of α_R is used in recognition. Additionally, the number of characters previously correctly recognized but incorrectly recognized by these compiled dictionaries monotonically decreases from 457 in row (b) to 221 in row (e), and the number of additional characters inserted also decreases from 123 to almost zero. These improvements were achieved using the approaches discussed above.

I. Parallel Results for Spontaneous Speech

A set of parallel experiments, like those summarized in Table VII, for the spontaneous speech task (CALLHOME) was conducted with dictionaries that were compiled in exactly the same way, in which the development set mentioned in Section III was used to determine $\alpha_F = 6$ for Stage 1, $\gamma = 0.8$ for Stage 2 and the average of 1.16 pronunciations per word in each dictionary. The results presented in Table VIII demonstrate very similar trends as those in Table VII for planned speech: recognition performance is monotonically improved and the "confusability of a dictionary" and the "added confusability for added pronunciations" is monotonically reduced from row (b) to row (e) and so on. However, the dictionary compiled from the *Free* transcription here results in much poorer recognition performance than dose the canonical baseline [64.65% in row (b) versus 60.88% in row (a)] and much higher "confusability of a dictionary" than does the canonical baseline (62.8% versus 54.5%). This result may follow from the fact that the surface transcription was generated by much poorer baseline acoustic models (telephone speech, with channel noise and a lower sampling rate as compared to the broadcast news data) and hence contains many more transcription errors. Therefore, the character error rate in row (c) is significantly lower than that in row (b) (61.41% versus 64.45%), and it is still slightly higher than that of the baseline in row (a) (60.88%). Comparing Table VIII with Table VII shows that the compiled dictionaries used in the spontaneous speech task has much higher "confusability of a dictionary" and "added confusability for added pronunciations" than do those used in the planned speech task. Besides, the spontaneous speech task also has much higher "intrinsic confusability" or "Confusability of a dictionary" of the canonical dictionary than that of the planned speech task (54.5% versus 44.0%).

Similar to that for planned speech, an experiment on the spontaneous speech was conducted using a cheating language model LM_{ct}, trained with the test data and regarded as more adequate for the recognition task than the fair language model LM_{fr} that was trained with the training data and was used above. An additional dictionary, $MPD(1.24)$, which includes more pronunciations (with an arbitrarily predetermined average of 1.24 pronunciations per word) was compared to the dictionary $MPD(1.16)$

TABLE IX
CERs (%) With Different Multiple-Pronunciation Dictionaries With Different Numbers of Variants and "Fair" and "Cheating" Language Models for the Spontaneous Speech Task (Callhome)

	language models	*CanD*	*MPD* (1.16)	*MPD* (1.24)
CER(%)	LM_{fr}	60.88	58.96	58.78
(Rec. on test data)	LM_{ch}	42.81	39.04	38.52

used above. Table IX presents the results in terms of character error rates. The table shows a very similar trend to that revealed by Table VI. Including more pronunciations in the $MPD(1.24)$ dictionary results in insignificant change in the recognition performance from that obtained using $MPD(1.16)$, when the fair language model LM_{fr} is used (from 58.96% to 58.78% with p-value $= 0.28$). When the cheating language model LM_{ch} is used, the $MPD(1.24)$ dictionary reasonably improves recognition performance (from 39.04% to 38.52%, p-value < 0.001), apparently because LM_{ct} can better distinguish among the confusing words caused by the extra pronunciation variants.

Tables VIII and VII show that the compiled multiple-pronunciation dictionaries yield more character error reduction in the planned speech task (4.97%, from 37.02% to 35.18%) than in the spontaneous speech task (3.15%, from 60.88% to 58.96%). This observation seems to be contrary to the expectation that multiple-pronunciation dictionaries should be more helpful in the spontaneous speech task than in the planned speech task. Possible reasons for the relatively less improvement in the spontaneous speech task may include the relatively poor baseline acoustic models used to acquire pronunciation variants (as indicated by the much poorer CER of 64.65% in row (b) in Table VIII) and the higher "intrinsic confusability" of the canonical dictionary (54.5% versus 44.0%).

V. Conclusion

This paper studies pronunciation modeling for Mandarin Chinese. New measures of the confusability of multiple-pronunciation dictionaries were proposed and were shown to be strongly correlated with recognition performance. A three-stage framework for constructing automatically the multiple-pronunciation dictionary was proposed. Experiments with both planned and spontaneous speech tasks show that the confusability as measured by the proposed metrics can be reduced stage by stage, and the recognition performance is improved accordingly.

In Stage 1, pronunciation generation, the proposed surface form generation procedure with an optimized value of the pronunciation probability weight α_F in the forced recognition significantly reduces the extra confusion introduced by the surface transcriptions. In Stage 2, pronunciation ranking, the proposed $pf{-}iwf$ score both reduces the confusability as measured and improves the recognition performance as compared to either the conventionally used pf score or a previously proposed criterion [4]. In Stage 3, pronunciation pruning, the *score*-based criterion outperforms the others, although all four criteria perform similarly when an optimized value of the pronunciation probability weight α_R is also applied in the recognition process to reduce the confusion. In all three stages, the reduction in the measured confusability was shown to be in parallel with the improvements in recognition performance.

The proposed approach may be further improved by integrating some implicit approaches, such as the use of phonological rules or decision trees to generate pronunciation variants for only those words that are less frequently used or absent from the training data, while explicitly enumerating the variants for relatively frequently used words, as proposed in this paper. Accordingly, the under- or over-generated pronunciation variants are fewer and the pronunciation variation of most of the words can still be better modeled with reduced confusion, indicating a possible direction for future work.

Acknowledgment

The authors would like to thank the reviewers for their extensive and valuable comments. They would also like to thank Dr. H. Strik, Dr. C. Cucchiarini, Dr. F. Wet, and Dr. O. Scharenborg in the Department of Language and Speech, Radboud University Nijmegen, The Netherlands, for very helpful discussions.

References

[1] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Commun.*, vol. 29, pp. 225–246, 1999.

[2] M. Weintraub, E. Fosler, C. Galles, Y.-H. Kao, S. Khudanpur, M. Saraclar, and S. Wegmann, "WS96 project report: Automatic learning of word pronunciation from data," presented at the *JHU Workshop Pronunciation Group*, 1996.

[3] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Int. Conf. Spoken Lang. Process.*, 1996, pp. 2328–2331.

[4] D. Torre, L. Villarrubia, L. Hernandez, and L. Elvira, "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," in *Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 1463–1466.

[5] M. Finke and A. Waibel, "Flexible transcription alignment," in *Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 34–40.

[6] ——, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Eur. Conf. Speech Commun. Technol.*, 1997, pp. 2379–2382.

[7] B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling for conversational speech recognition: A status report from WS97," in *IEEE Workshop Speech Recognition Understanding*, 1997, pp. 26–33.

[8] W. Byrne, V. Venkataramani, T. Kamm, T. F. Zheng, Z. Song, P. Fung, Y. Liu, and U. Ruhi, "Automatic generation of pronunciation lexicons for Mandarin spontaneous speech," in *Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 569–572.

[9] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Commun.*, vol. 29, pp. 209–224, 1999.

[10] E. Fosler-Lussier and G. Williams, "Not just what, but also when: Guided automatic pronunciation modeling for broadcast news," in *DARPA Broadcast News Workshop*, 1999, pp. 171–174.

[11] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," in *Eur. Conf. Speech Commun. Technol.*, 1999, pp. 463–466.

[12] T. Holter and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," *Speech Commun.*, vol. 29, pp. 177–191, 1999.

[13] N. Cremelie and J.-P. Martens, "In search of better pronunciation models for speech recognition," *Speech Commun.*, vol. 29, pp. 115–136, 1999.

[14] M.-Y. Tsai, F.-C. Chou, and L.-S. Lee, “Improved pronunciation modelling by inverse word frequency and pronunciation entropy,” in *Proc. Automatic Speech Recognition and Understanding Workshop*, 2001, pp. 53–56.

[15] ——, “Improved pronunciation modeling by properly integrating better approaches for baseform generation, ranking and pruning,” in *Proc. ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, 2002, pp. 77–82.

[16] M. Wester, “Pronunciation modeling for ASR—Knowledge-based and data-derived methods,” *Comput. Speech Lang.*, pp. 69–85, 2003.

[17] G. Tajchman, E. Fosler, and D. Jurafsky, “Building multiple pronunciation models for novel words using exploratory computational phonology,” in *Proc. Eur. Conf. Speech Commun. Technol.*, 1995, pp. 2247–2250.

[18] J. M. Kessens, M. Wester, and H. Strik, “Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciaiton variation,” *Speech Commun.*, vol. 29, pp. 193–207, 1999.

[19] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, “Pronunciation modeling using a finite-state transducer representation,” in *Proc. ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, 2002.

[20] T. Fukada, T. Yoshimura, and Y. Sagisaka, “Automatic generation of multiple pronunciations based on neural networks,” *Speech Communicaiton*, vol. 27, pp. 63–73, 1999.

[21] M.-Y. Tsai, F.-C. Chou, and L.-S. Lee, “Pronunciation variation analysis with respect to various linguistic levels and contextual conditions for Mandarin Chinese,” in *Proc. Eur. Conf. Speech Commun. Technol.*, 2001, pp. 1445–1448.

[22] I. Amdal, F. Korkmazskiy, and A. C. Surendran, “Joint pronunciation modelling of non-native speakers using data-driven methods,” in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, pp. 622–625.

[23] Q. Yang and J.-P. Martens, “Data-driven lexical modeling of pronunciation variations for ASR,” in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, pp. 417–420.

[24] Q. Yang, J.-P. Martens, P.-J. Ghesquiere, and D. V. Compernolle, “Pronunciation variation modeling for ASR: Large improvements are possible but small ones are likely to achieve,” in *Proc. ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, 2002, pp. 123–128.

[25] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, “On the road to improved lexical confusability metrics,” in *Proc. ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, 2002, pp. 53–58.

[26] R. Baeza and B. Ribeiro*, Modern Information Retrieval*. New York: ACM Press, 1999.

[27] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland*, The HTK Book Version 3.0*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[28] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics*, vol. 1, pp. 80–83, 1945.

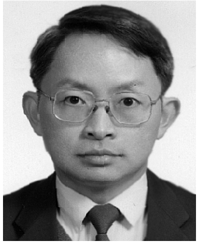
[29] M. Saraclar, H. Nock, and S. Khudanpur, “Pronunciation modeling by sharing Gaussian densities across phonetic models,” *Comput. Speech Lang.*, vol. 14, pp. 136–160, 2000.

Ming-Yi Tsai was born in 1976. She received the B.S. degree in communication engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1998. She was admitted for direct-track program from M.S. to Ph.D. in 1999 and received the Ph.D. degree in communication engineering from National Taiwan University, Taipei, in 2006.

She was a Part-Time Researcher in Applied Speech Technologies, Taipei, from 1999 to 2001. She was a Visiting Researcher in the Department of Language and Speech, Nijmegen University, Nijmegen, The Netherlands, from 2003 to 2004. She is now a Senior Engineer with MediaTek, Inc., Hsinchu, Taiwan. Her research interests include pronunciation variation analysis and modeling for automatic speech recognition, language learning system, and speech quality enhancement.

Fu-Chiang Chou received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1989 and 1999, respectively.

He was a Postdoctoral Fellow at the Institute of Linguistics, Academia Sinica, Taipei, in 1999. He was the Chief Technology Officer of Applied Speech Technologies, Taipei, from 1999 to 2001. Since 2001, he has been with Philips Research East Asia, Taipei, as a Senior Researcher. He is now an Assistant Professor of computer and communication engineering at Ming Chuan University, Taipei. His research interests are in the area of digital speech processing with special interests on text-to-speech and speech recognition systems.



Lin-Shan Lee (F’93) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at National Taiwan University, Taipei, Taiwan, R.O.C., since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech system, natural language analyzer, dictation systems, and voice information retrieval system.

Dr. Lee was Guest Editor of the “Special Issue on Intelligent Signal Processing in Communications” of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION in December 1994 and January 1995. He was Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society. He has been a member of Permanent Council of International Conference on Spoken Language Processing (ICSLP), was the convenor of the International Coordinating Committee of Speech Databases and Assessment (COCOSDA, 2000–2001), and is currently a member of the Board of International Speech Communication Association (ISCA).