.

Deriving disyllabic word variants from a Chinese conversational speech corpus

Yi-Fen LiuShu-Chuan TsengJyh-Shing Roger JangJFL

Citation: The Journal of the Acoustical Society of America **140**, 308 (2016); doi: 10.1121/1.4954745

View online: http://dx.doi.org/10.1121/1.4954745

View Table of Contents: http://asa.scitation.org/toc/jas/140/1

Published by the Acoustical Society of America

Deriving disyllabic word variants from a Chinese conversational speech corpus

Yi-Fen Liu
Graduate Institute of Information Systems and Applications, National Tsing Hua University, 101, Section 2, Kuang-Fu Road, Hsinchu, 30013, Taiwan

Shu-Chuan Tseng[a]
Institute of Linguistics, Academia Sinica, 128, Section 2, Academia Road, Taipei, 11529, Taiwan

Jyh-Shing Roger Jang
Department of Computer Science and Information Engineering, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei, 10617, Taiwan

(Received 30 October 2015; revised 8 June 2016; accepted 8 June 2016; published online 15 July 2016)

Motivated by the quasi-categorical reduced forms of disyllabic words produced in Chinese conversational speech, a frequency-based selection procedure of typical pronunciation by disyllabic word type and reduction degree is proposed in this paper. This variant-selection algorithm utilizes techniques of free phone recognition and phonetic similarity score calculation to account for Chinese syllable structure. Four reduction types are suggested by considering the presence of a within-word syllable boundary: Citation form-like reduction, marginal segment deletion, nuclei merger, and syllable merger. The results show that the most frequent reduction types for disyllabic words in Chinese conversation are citation form-like reduction and syllable merger. In particular, high-frequency disyllabic words preferentially take the extreme syllable-merger form. As shown in the analysis, segmental reduction in Chinese disyllabic words is morphology-dependent. It is also related to the prosodic position at which a disyllabic word is produced as well as the temporal quality of the word. Finally, in the automatic speech recognition experiments, the performance was improved by adding a small number of variants selected by the algorithm to the pronunciation dictionary of the system.
[http://dx.doi.org/10.1121/1.4954745]

[JFL] Pages: 308–321

I. INTRODUCTION

In realistic speech communication, words are articulated in sequence with a wide range of phonetic variability. Regardless of slightly or extremely reduced pronunciation, seemingly diverse word variants seldom cause problems in understanding casual speech for humans. However, recognizing reduced speech remains a challenging task for automatic speech recognition (ASR) systems. Analyzing phonetic variation in reduced speech not only is relevant in human speech communication research (e.g., production form and mental representation of spoken words), but also strengthens the performance of ASR systems. Although numerous statistical approaches (Akita and Kawahara, 2010; Hofmann *et al.*, 2010; Jyothi *et al.*, 2013; Karanasou *et al.*, 2013; McGraw *et al.*, 2013) have been adopted to handle problems in pronunciation modeling, highly diverse word variants that deviate from their citation form often lead to severe performance deterioration. Knowledge on the cognitive representation and processing of reduced spoken words is necessary for both engineering applications and linguistic research. Most importantly, to build an automatic system that functions as humans do, understanding why and how humans can connect diverse reduced word forms with their canonical meaning so rapidly in casual speech is imperative. Speech information is delivered more than merely by acoustic signals of speech; accordingly, this information is possibly decoded by the specific language system that functions in the users.

Specific reduced word forms are likely preferential, leading to such preferred forms being more closely connected with the meaning than the other word forms. Pierrehumbert (1994) similarly suggested that "linguistic competence" consists of underlying principles that enable humans to use language in a quasi-categorical manner. This notion of linguistic competence implies that semantic meaning is connected with "categorical" phonetic forms of spoken words. More concretely in the exemplar model, the surface forms of spoken words are stored as exemplars (Pierrehumbert, 1994). Usage-based production frequency might be related to specific categorical phonetic forms that can be restored from the acoustic correlates of the phonetic forms. In other words, word production frequency and acoustic properties are likely the two decisive factors enabling humans to construct a type of system with representative word variants (Dilley and Pitt, 2007; Pitt *et al.*, 2011). In the process of decoding acoustic information to interpret semantic meaning in a given

[a]Electronic mail: tsengsc@gate.sinica.edu.tw

language, the linguistic system that depends on individual languages also plays a role in processing reduced word forms such as in the morphological and phonological structure of words.

In recent decades, efforts have been focused on strengthening ASR performance by embedding speech variability in pronunciation modeling modules. The pronunciation dictionary in an ASR system, which connects the acoustic decoding and linguistic interpretation of spoken words, is expanded by including word variants in addition to the citation form. Often, surface transcription obtained using automatic free phone recognition is used for deriving pronunciation variants, which are elicited by learning rules and deriving parameters through various approaches such as decision trees (Fosler-Lussier, 1999; Liu and Fung, 2004a), artificial neural networks (Fukada *et al.*, 1999; Chen and Hasegawa-Johnson, 2004), and the conditional random field model (Karanasou *et al.*, 2013). More oriented toward linguistic systems, word variants can also be derived from phonological rules (Schuppler *et al.*, 2011).

Concerning Chinese ASR systems, improvements have been achieved by utilizing pure lexicon enhancement such as using the pruning method (Tsai *et al.*, 2007), acoustic models allowing more tolerance (Liu and Fung, 2004b), or both (Byrne *et al.*, 2001; Liu and Fung, 2004a). Collectively, various degrees of improvement have been achieved. However, the extracted surface forms are not necessarily a faithful reflection of the phonetic forms that speakers prefer in speech communication. Another limitation caused by implicit pronunciation modeling approaches is the possibility of connecting word variants with the prosodic position and context in which words are spoken, though these two factors are closely related to the pronunciation of words (Torreira and Ernestus, 2011; Hanique *et al.*, 2013).

The central concern of this paper is identifying a formal means of deriving typical word variants that, to a certain degree, model common usages by humans and simultaneously assist system developers in selecting word variants them appropriately. Thus, we propose a derivation algorithm by employing an automatic usage-based approach with considerations of language-dependent syllable structure and reduction degree. Accordingly, we account for the three aforementioned factors: acoustic properties, language system, and usage. This paper also reports the results of implementing our proposed algorithm on disyllabic words that comprise most modern Chinese usages in both written and spoken forms (Tseng, 2013a).

II. CHINESE DISYLLABIC WORD VARIANTS

Disyllabic words account for approximately 40% of the overall word tokens and 60% of word types in the 42-h Chinese Conversational Corpus (Tseng, 2013a). Only monosyllabic words outnumber disyllabic words in tokens, because singular pronouns and many frequently used function words in Chinese, such as the structural particle 的*de* and past tense particle 了*le*, are monosyllabic. Disyllabic words not only have high coverage in conversational use but also play a critical role in the tradition of Chinese phonology. Because of the lack of phonetic transcription systems in ancient China, a character is phonetically transcribed by two already existing characters. In concrete terms, the to-be-transcribed character inherits the onset from the first character syllable and the rhyme from the second character syllable. This manner of syllable merger not only functions as a means of phonetic transcription, but also has some impact on word morphology and the writing system. New phonetic forms and characters can be invented by applying this two-syllable merging rule. For instance, the phrase “rise up” *ki* (rise) *lai* (hither) in Southern Min has a shortened form *kiai* in colloquial speech. In Mandarin, the function word 諸 *zhū* is the merger of the two function words 之*zhī* and 乎*hū*, which is an example of syllable mergers influencing the writing system.

In a similar notion, Chinese phonologists have observed this merging rule in many Chinese dialects, called the Edge-in Theory (Chung, 1997; Hsu, 2003). Syllable mergers may be a representative spoken word form, which is also referred to as the coalescence of syllables. However, syllable mergers are in no case the only form of a disyllabic word, because spoken word reduction resembles a spectrum with varying degrees. When reduction is viewed from the presence of the word-internal syllable boundary, two syllables of a spoken disyllabic word can be uncontracted, contracted, or merged, forming a type of categorical phonetic representation of reduced spoken words. Our approach involved adopting the categorical phonetic forms of spoken words, in principle conforming to the notions of the exemplar model (Pierrehumbert, 1994) and the results of syllable contraction (Tseng *et al.*, 2013b). In addition to reduction degree and word frequency, we also considered the syllable structure of words as a key feature affecting the final surface forms of spoken words.

A. Disyllabic words in conversational speech

The present study analyzed disyllabic word variants in the 8-h Mandarin Conversational Dialogue Corpus (MCDC8), which is distributed by the Association for Computational Linguistics and Chinese Language Processing (2016). Figure 1 shows a summary of the number of word tokens and types as well as their proportion percentage in the MCDC8. The word production frequency in the MCDC8 is similar to that reported in a larger corpus (Tseng, 2013a). Thus, we directly used the MCDC8 word frequency to represent the general use of words in realistic speech communication.

In the MCDC8, the duration of ordinary syllables ranges from 15 to 1110 ms (mean, 173 ms) and is equal to 5.78 syllables per second, which is faster than the articulation rate of news reporters in a Chinese Broadcast News corpus (Chien and Huang, 2003). We ran the variant selection algorithm on all disyllabic words in the MCDC8. For subsequent ASR experiments, we divided the MCDC8 into three subsets. Eighty percent of the speech data were randomly selected as the training set (75 104 word tokens). The remaining 20% of speech data were used as development (9046 word tokens) and evaluation sets (9383 word tokens). That is, 90% of the MCDC8 data were used for constructing a word-bigram language model. The evaluation set was mainly used to assess

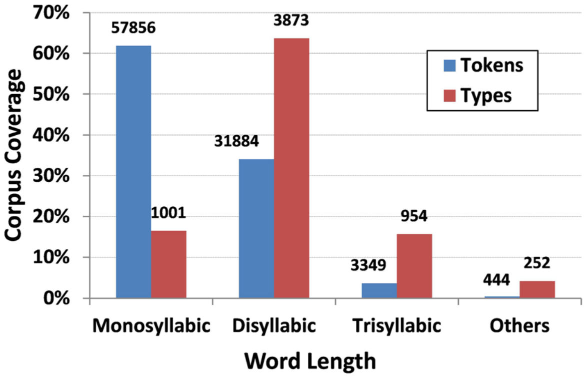

FIG. 1. (Color online) Word distribution in the MCDC8 Corpus.

the impact on ASR performance of adding variants selected using our proposed method.

B. Reduced disyllabic words

Concerning the reduction degree of disyllabic words, two factors are considered: the presence of an identifiable within-word syllable boundary and segment deletion across the syllable boundary (Tseng *et al.*, 2013b). Phonologically speaking, a Chinese syllable has the CGVN structure: an optional onset consonant, an optional prevocalic glide (/j, w/), a nucleus, and an optional nasal coda (/n, ŋ/) with no consonant clusters. The phonological tree of a disyllabic Chinese word, *xiàn zài* /ɕ j e n ts ai/ (meaning "now"), is shown in Fig. 2. Concerning the notation, the terms "INITIAL" and "FINAL" are conventionally used in Chinese phonology, which are equivalent to the onset and rhyme (including the prevocalic glide, called the Medial).

When the word-internal syllable boundary is present and none of the consonant segments across the syllable boundary (i.e., the nasal coda of the first syllable and the onset of the second syllable) are deleted, the word is classified into the category of Canonical Form (CAN). For instance, /ɕ j e n ts^{h} ai/ is regarded as a case of CAN. Note that substituting consonant segments across the syllable boundary is permitted in this category. Second, when a disyllabic word has a clear within-word syllable boundary but some (not all) of the consonant segments across the syllable boundary are omitted, it is classified as a case of Marginal Segment Deletion (MSD; e.g., /ɕ j e n ai/). Furthermore, when the nuclei across the syllable boundary are somehow merged and the syllable boundary is blurred, it is a case of a Nucleus Merger (NUM; e.g., /ɕ j e ai/). This type of phonetic erosion is more severe than MSD because the nuclei of the two syllables are merged. Finally, the extreme case is a Syllable Merger (SYM), in which two syllables are merged into one (e.g., /ɕ j ai/). In Sec. III, we describe the formulation of our algorithm, which automatically derives these four reduction types (RT) from disyllabic word tokens according to the acoustic properties and syllable structure of the words.

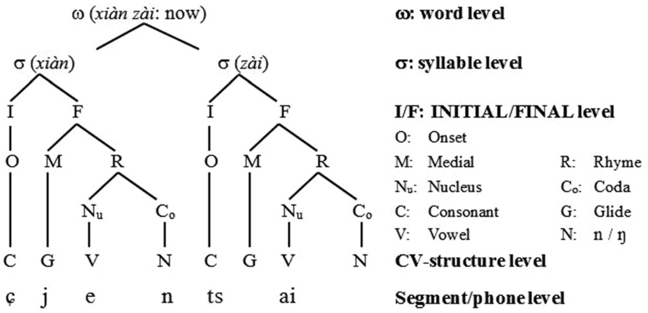

FIG. 2. Phonological tree structure of a Chinese disyllabic word.

III. DERIVING WORD VARIANTS

Our first step in deriving word variants was to categorize the surface forms of disyllabic words into the four reduction types defined in Sec. II. To do this, we conducted automatic generation of word-level pronunciations by first training a free phone recognizer, as shown in Fig. 3. Reduction types were then categorized by comparing the phone sequences of the surface and canonical forms, whereas the canonical form was generated from the lexical information. Typical variants of spoken disyllabic words were selected from the most frequent reduction types to be later added into a dictionary for evaluating ASR performance.

A. Surface form generation

For data pre-processing, long speech stretches in the MCDC8 were first segmented into inter-pause units (IPU) according to their silent pauses and diverse types of paralinguistic sounds such as laughter and inhalation (Liu *et al.*, 2014). For free phone recognition, word segmentation was required because Chinese is written consecutively in characters with no word boundary marks, which are normally available in alphabetic languages. Moreover, Chinese word segmentation is controversial because distinct morphological theories can yield different results. Among the many fine-grained word segmentizers that have been developed for processing Chinese, we adopted the word segmentation system developed by the team of Chinese Knowledge and Information Processing (CKIP) at Academia Sinica to process the transcripts (Ma and Chen, 2004). The segmented transcripts and sounds were forced-aligned to obtain initial word boundaries, which were verified by professional phonetic labelers and used for automatically generating word-level surface forms.

The Hidden Markov Model Toolkit (HTK) and the SRI Language Modeling Toolkit (SRILM) (Stolcke, 2002; Young *et al.*, 2006) were employed to train the acoustic and language models and perform the free phone recognition and ASR experiments. We used 52 gender- and context-independent monophone Hidden Markov Models (HMM) comprising 39 phones in ordinary syllables and 13 speech-

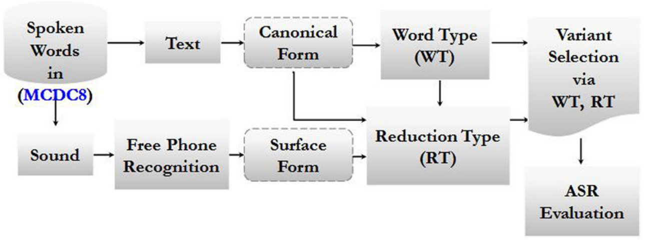

FIG. 3. (Color online) Flowchart of variant selection.

related phenomena that are common in conversational speech, such as fillers, particles, word fragments, and paralinguistic sounds (Liu *et al.*, 2014). For fillers, we used three acoustic models to model the nasality and length; two for monosyllabic instances, one with and one without a nasal coda, and one for multisyllabic fillers, irrespective of the presentence of a final coda. Four acoustic HMM models were trained for discourse particles originating from Mandarin Chinese and Southern Min, a Chinese dialect predominantly spoken in Taiwan. Each HMM comprises three left-to-right states, each with only one Gaussian mixture. The acoustic features were 12 Mel-Frequency Cepstral Coefficients (MFCCs) plus their energy, as well as their delta and acceleration for 15-ms frames with a 5-ms frame shift. The subset of 39 acoustic phone models was used for deriving the surface forms from the training data, but the whole set of 52 acoustic models was used for the ASR experiment.

Finally, the surface form of all disyllabic words in the MCDC8 was aligned with the citation form through dynamic programming, in which phonetic similarity was used as the principal score for generating a word-level pronunciation table containing the freely recognized phone sequences of words with the paths of deletion, substitution, and insertion of citation phones.

B. Word type definition

In disyllabic words, the presence of the word-internal syllable boundary is closely connected with the surface form. Thus, whether any consonant segment exists across the word-internal syllable boundary is critical. To account for this factor, we defined three word types on the basis of syllable structure. Starting from syllable types, the coda position in a Chinese syllable can be occupied only by nasal consonants, leading to eight distinct syllable types: V, GV, VN, GVN, CV, CGV, CVN, and CGVN. As mentioned, the presence of an onset (On) and a coda (Co) is critical in formally presenting these eight syllable types, as shown in Table I. To clarify our notation, we take *xiàn zài* (meaning “now”) as an example. The first syllable (σ_1) *xiàn* has both an onset and a coda, whereas the second syllable (σ_2) *zài* has an onset but no coda. Adopting the notation in Table I, the two syllables are formally represented as $O_nC_o\#O_n\varnothing$, where # denotes the syllable boundary and $C_o\#O_n$ means that both the σ_1 coda and σ_2 onset are present. We then regroup the 8×8 syllable pairs in disyllabic words into three word types by considering the presence of C_o and O_n across the syllable boundary (#).

1. *Word type I ($C_o\#O_n$)*

As shown in Fig. 4(a), both the σ_1 coda and σ_2 onset are present, leading to 16 syllable-type combinations, ${}_{\sigma1}\{\varnothing C_o, O_nC_o\} \times {}_{\sigma2}\{O_n\varnothing, O_nC_o\}$. The previous example *xiàn zài* belongs to Word type I.

TABLE I. Classification of Chinese syllables.

		Coda	
		$\varnothing$	C_o
Onset	$\varnothing$	V, GV	VN, GVN
	O_n	CV, CGV	CVN, CGVN

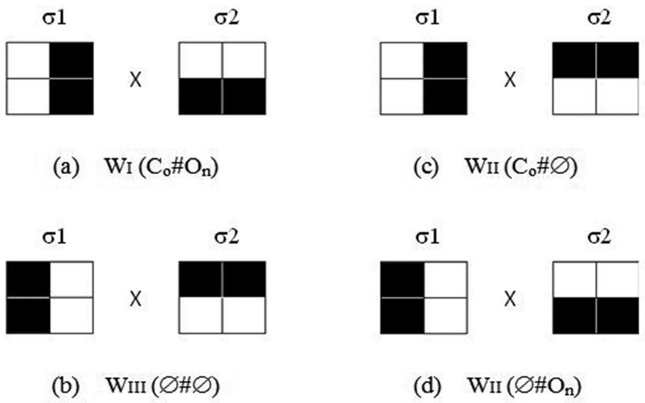

FIG. 4. Syllable-type combinations of three WTs.

2. *Word type II: ($C_o\#\varnothing$ and $\varnothing\#O_n$)*

Either the σ_1 coda or σ_2 onset is present, leading to 16 syllable-type combinations in cases of $C_o\#\varnothing$: ${}_{\sigma1}\{\varnothing C_o, O_nC_o\} \times {}_{\sigma2}\{\varnothing\varnothing, \varnothing C_o\}$ and 16 combinations in cases of $\varnothing\#O_n$, ${}_{\sigma1}\{\varnothing\varnothing, O_n\varnothing\} \times {}_{\sigma2}\{O_n\varnothing, O_nC_o\}$. The combination pairs are summarized in Figs. 4(c) and 4(d), respectively.

3. *Word type III: ($\varnothing\#\varnothing$)*

As denoted in Fig. 4(b), both the coda of σ_1 and onset of σ_2 are empty, leading to 16 combinations ${}_{\sigma1}\{\varnothing\varnothing, O_n\varnothing\} \times {}_{\sigma2}\{\varnothing\varnothing, \varnothing C_o\}$ in this case.

C. Phonetic similarity score

When the automatic phoneme alignment was conducted and the word-level surface forms were generated, phonetic similarity scores were used to select the optimally matched phone sequences. These scores were also used in cases of equal frequency. Because an effectively designed similarity schema rewards good matches and penalizes poor matches to achieve meaningful lengths in alignment, we implemented the phonetic alignment approach employed by Kondrak (2003) with a set of operating functions (i.e., insertion, deletion, and substitution) for our dynamic alignment between the citation form and surface form.

As defined in Table II, the scoring functions C_{skip} and C_{sub} are the maximum scores for insertion, deletion, and substitutions with default values $C_{skip} = -10$, $C_{sub} = 35$, and

TABLE II. Scoring functions.

$$\sigma_{skip}(p) = C_{skip}$$

$$\sigma_{sub}(p,q) = C_{sub} - \delta(p,q) - V(p) - V(q)$$

where

$$V(p) = \begin{cases} 0 & \text{if } p \text{ is a consonant} \\ C_{vwl} & \text{otherwise} \end{cases}$$

otherwise

$$\delta(p,q) = \textstyle\sum_{f \in R} \text{diff}(p,q,f) \times \text{salience}(f)$$

where

$$R = \begin{cases} R_C & \text{if } p \text{ or } q \text{ is a consonant} \\ R_V & \text{otherwise} \end{cases}$$

$C_{vwl} = 10$. Similar to a distance table, the similarity table was established using the σ scoring functions defined in Table II and was employed to retrieve the optimal alignments. Furthermore, the phonetic segments used in our system for Mandarin Chinese were encoded as the vectors of feature values in floating-point numbers in the range [0, 1]. The function $diff(p, q, f)$ returns the difference between segments p and q for a given feature vector f in 12 dimensions.

For determining similarity scores, the numerical values, which range from 0.0 to 1.0, convey four principal features on the Place of Articulation (bilabial = 1.0, labiodental = 0.95, dental = 0.9, alveolar = 0.85, retroflex = 0.8, palate-alveolar = 0.75, palatal = 0.7, velar = 0.6, uvular = 0.5, pharyngeal = 0.3, glottal = 0.1), Manner of Articulation (stop = 1.0, affricate = 0.9, fricative = 0.8, approximant = 0.6, high vowel = 0.4, mid vowel = 0.2, low vowel = 0.0), Vocalic Property-Highness (high = 1.0, mid = 0.5, low = 0.0), and Vocalic Property-Backness (back = 1.0, central = 0.5, front = 0.0). This quantification process was adapted from Connolly (1997) to advocate the phonetically based multivalued feature system proposed in Ladefoged (2006) with a supplementary weight on salience (Kondrak, 2003). The salience settings for feature sets concerning the vocalic and consonantal properties R_V and R_C are detailed in Table III.

Please note that some language-dependent changes were made for the variant of Mandarin Chinese spoken in Taiwan. In principle, vowel length feature was excluded because no such distinction exists in Chinese. The salience settings were suppressed for Lateral and Retroflex, but raised for Aspirated.

D. Reduction type categorization

This section presents our approach to categorizing reduction types of Chinese disyllabic words by using information from acoustic properties (surface form), linguistic structure (word type), and usage (reduction degree). For notation, W_I, W_{II}, and W_{III} were adopted to represent sets containing words belonging to Word types I, II, and III, respectively. Moreover, we used x to denote any disyllabic word and y to denote the freely recognized phone sequence for x. The reduction type was implemented in a straightforward equation as follows. As shown in Eq. (1), if y recognized for a word x has a similar phone sequence as that of the citation form (i.e., no consonant across the syllable boundary omitted), then y is classified as CAN. Concerning the length constraint, the number of segments should exceed that of the shortest segment sequence of the word type to which x belongs. As shown in Eq. (2), if $x \in W_I$ and the decoded y has a phone sequence with at least one consonant segment (deleted) and one consonant segment (original or substituted), then y is categorized as MSD. If $x \in W_{II}$, and y has a glide substituting the consonantal segment, then it is also classified as MSD. For Word type III with empty consonantal segments across the syllable boundary, no MSD is defined. The length constraint guarantees that the phone number of y does not exceed that of the citation form with a slightly reduced segment number. As shown in Eq. (3), the CV-structure constraint on y for all three word types is that y must have two consecutive Vs The length constraint limits the length of y to have at least two deleted segments. In contrast to CAN and SYM, the word-internal syllable boundary is more blurred in NUM than in MSD. As shown in Eq. (4), y contains at most one V, and the length of y should not exceed the length of one syllable.

TABLE III. Features and salience settings for Mandarin Chinese.

Feature	Salience	R_C	R_V	Feature	Salience	R_C	R_V
Syllabic	5	+	+	Place	40	+	–
Voicing	10	+	–	Nasal	10	+	+
Lateral	5	+	–	Aspirated	10	+	–
High	5	–	+	Back	5	–	+
Manner	50	+	–	Retroflex	5	+	+
Diphthong	5	–	+	Round	5	–	+

$$\begin{aligned}
&\mathrm{R_{CAN}}: \mathit{For}\,\forall x,\ \text{and its } y \text{ constrained by conditions (a) and (b)}\\
&\quad (a)\ \text{if } x \in W_{\mathrm{I}},\ \text{and its } y = {}^{\wedge}\mathrm{C}*[\mathrm{GV}]+\mathrm{CC}+[\mathrm{GV}]+\mathrm{C}*\ \$,\\
&\quad\quad \text{if } x \in W_{\mathrm{II}},\ \text{and its } y = {}^{\wedge}\mathrm{C}*[\mathrm{GV}]+\mathrm{C}+[\mathrm{GV}]+\mathrm{C}*\ \$,\\
&\quad\quad \text{if } x \in W_{\mathrm{III}},\ \text{and its } y = {}^{\wedge}\mathrm{C}*\mathrm{G?V}+\mathrm{C}*[\mathrm{GV}]+\mathrm{C}*\ \$\\
&\quad (b)\,\mathit{length}(y) \ge \varepsilon,\\
&\quad\quad \text{where } \varepsilon = \text{minimum segment length of the}\\
&\quad\quad \text{word type to which } x \text{ belongs,}
\end{aligned} \tag{1}$$

$$\begin{aligned}
&\mathrm{R_{MSD}}: \mathit{For}\,\forall x,\ \text{and its } y \text{ constrained by conditions (a) and (b)}\\
&\quad (a)\ \text{if } x \in W_{\mathrm{I}},\ \text{and its } y = {}^{\wedge}\mathrm{C}*[\mathrm{GV}]+[\mathrm{CG}]\{1\}+[\mathrm{GV}]+\mathrm{C}*\ \$,\\
&\quad\quad \text{if } x \in W_{\mathrm{II}},\ \text{and its } y = {}^{\wedge}\mathrm{C}*[\mathrm{GV}]+G[\mathrm{GV}]+\mathrm{C}*\ \$\\
&\quad (b)\ \mathit{length}(x)+1 \ge \mathit{length}(y) \ge \varepsilon-1,\\
&\quad\quad \text{where } \varepsilon = \text{minimum segment length of the}\\
&\quad\quad \text{word type to which } x \text{ belongs,}
\end{aligned} \tag{2}$$

$$\begin{array}{l}
\mathrm{R}_{\mathrm{NUM}} : \mathit{For}\ \forall x,\ \text{and its } y \text{ constrained by conditions (a) and (b)}\\
\quad (a)\ \text{if } x \in W_{\mathrm{I}} \cup W_{\mathrm{II}} \cup W_{\mathrm{III}},\ \text{and its } y = {}^{\wedge}\mathrm{C} * \mathrm{G?VV} + \mathrm{C} * \$\\
\quad (b)\ \mathit{length}(x) \geq \mathit{length}(y) \geq \varepsilon - 2,\\
\qquad \text{where } \varepsilon = \text{minimum segment length of the}\\
\qquad \text{word type to which } x \text{ belongs,}
\end{array} \qquad (3)$$

$$\begin{array}{l}
\mathrm{R}_{\mathrm{SYM}} : \mathit{For}\ \forall x,\ \text{and its } y \text{ constrained by conditions (a) and (b)}\\
\quad (a)\ \text{if } x \in W_{\mathrm{I}} \cup W_{\mathrm{II}} \cup W_{\mathrm{III}},\ \text{and its } y = {}^{\wedge}\mathrm{C} * \mathrm{G?V?G?C} *\ \$\\
\quad (b)\ \mathit{length}(y) \leq \varepsilon,\\
\qquad \text{where } \varepsilon = 5,\ \text{the maximum segment length of a Chinese}\\
\qquad \text{syllable with a pre–nucleus G and a post–nucleus G.}
\end{array} \qquad (4)$$

IV. ANALYSIS OF SELECTED VARIANTS

We ran the variant selection procedure on all disyllabic words in the MCDC8, except for 10% of the words in the evaluation set, which were later regarded as unseen words in the ASR experiment. As a result, only one disyllabic word was included in the unseen word set. Thus, 3872 disyllabic words from the MCDC8, totaling 31878 tokens, were processed through free phone recognition, surface form generation, word type classification, and reduction type categorization. Subsequently, the most frequent phone sequence from the top-ranked reduction type was selected as the typical variant. This section analyzes various linguistic aspects of the variants selected using our proposed method.

A. Frequent words prefer syllable merger

According to the Exemplar Theory (Pierrehumbert, 1994) and the Magnet Theory (Kuhl *et al.*, 2008), clusters of phonetic forms are formed by production frequency and phonetic similarities to the citation forms. Solid clusters may, to a certain degree, represent typical word variants other than the citation form in the mental lexicon, which are also closely linked with the word meaning. However, few studies have discussed word variants in spontaneous speech because of a lack of automatic tools for processing large-scale natural speech data. Applying our algorithm yielded insightful observations, as shown in Fig. 5.

First, the more frequent a disyllabic word is produced, the more likely it is that the extreme merger form (SYM) is selected as the representative RT. Figure 5 depicts the tendencies of the four RTs for disyllabic words that appear more than ten times and are ranked in order of production frequency. In the figure, the darker the color is, the higher the percentage of the selected RT is. For the 100 most frequent disyllabic words, the most representative RT is SYM. For words that are produced less frequently, CAN, indicating the least deviation from the citation form, is selected. The two types in between, MSD and NUM, are clearly less frequently used. The tendency is obvious: the more often a word is produced, the more reduced it is in spontaneous speech. In addition, the observed RT preference may lead us to reconsider the observation of Kuhl *et al.* (2008) that reduced speech may serve as the connection that forges a learning map between the produced form and the perceived sound in the developmental nature of language acquisition.

Concerning word length in terms of duration, the kernel density plots in Fig. 6 support the notion that the more reduced a spoken word is, the shorter it tends to be. Note that we used a scaled duration for normalization, $DUR_{\mathrm{norm}} = (x_i - \bar{x}) / (x_{\max} - x_{\min})$, where x_i is the word duration and $\bar{x}$ denotes the average duration from all the tokens of a given disyllabic words spoken by a given speaker. The nominator term is the difference of the maximum and minimum durations observed from the sample in order to scale the original duration to a fixed range between 0 and 1 (Lobanov, 1971). As the confidence intervals show, the mean duration of words for which SYM is selected as the typical RT is 0.315, which is considerably shorter than that of those words categorized as NUM, MSD, or CAN. Moreover, the RT coverage in the case of SYM is 51.92%, indicating that native speakers prefer producing SYM in conversational speech,

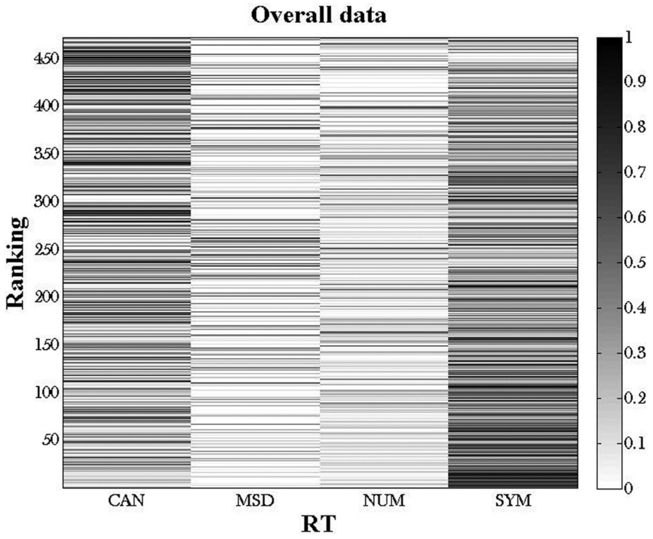

FIG. 5. Results of RT classification in terms of word frequency ranking and the RT coverage percentage (the darker the line boxes, the higher the RT percentage).

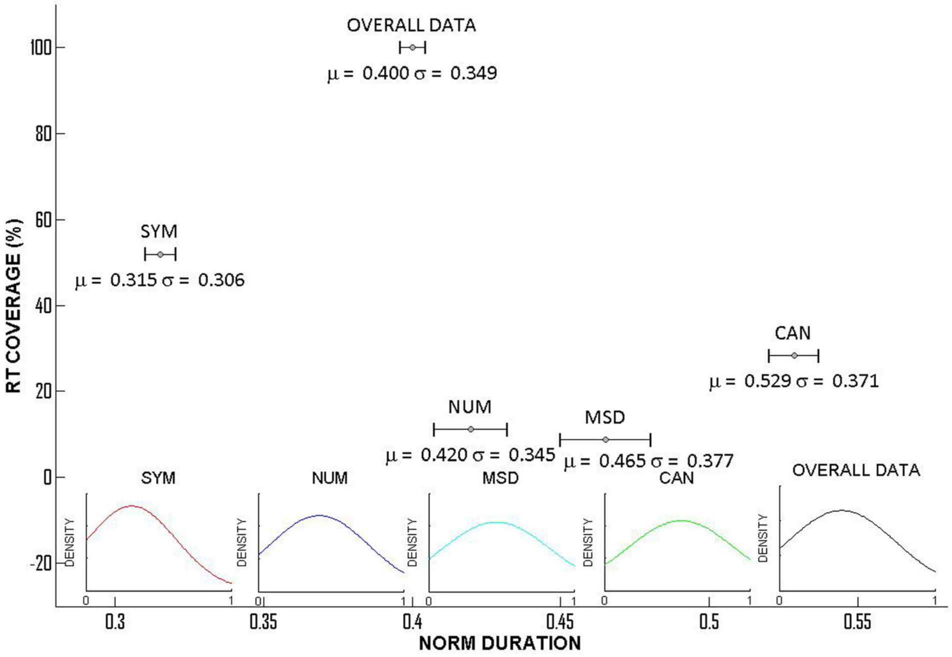

FIG. 6. (Color online) RT coverage and duration pattern for overall data and RTs.

followed by CAN with a coverage rate of 28.29%. As for MSD and NUM, their sum coverage is approximately 20%.

B. Prosodic position results in differing variants

The prosodic position in which a word is produced affects the prosodic as well as segmental properties of the word (Torreira and Ernestus, 2011; Hanique *et al.*, 2013). To observe whether our algorithm would choose different RTs in pools of words occurring in different prosodic positions, we divided our disyllabic words into four groups according to their position in the IPU: initial, medial, final, and isolated IPU (an IPU formed by a single disyllabic word). Figure 7 shows the results. For IPU-initial disyllabic words, more SYM are selected with a higher coverage (the darker column). The figure also shows that the more frequently a word

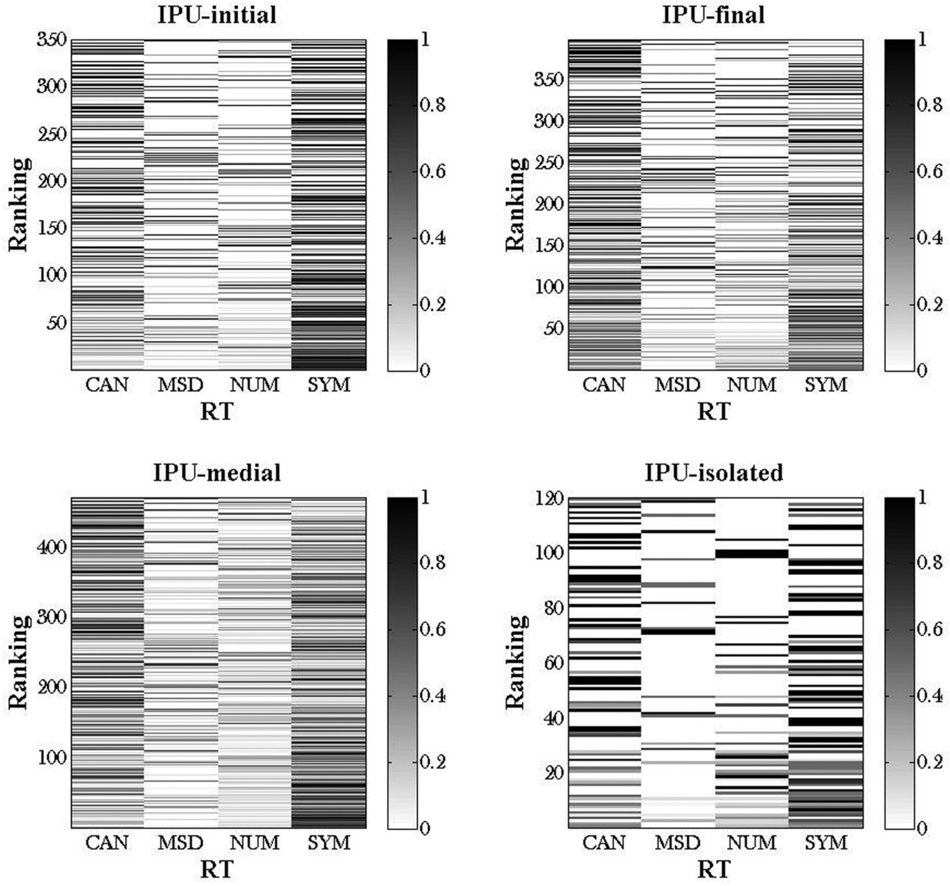

FIG. 7. Results of RT classification, separated for four prosodic positions.

is produced, the higher the chance of it being categorized as SYM is. By contrast, for IPU-final words, more CANs are selected.

Concerning Chinese, the final lengthening and initial shortening of words are mentioned in Tseng *et al.* (2013b) from the perspective of duration. Our results suggest that prosodic position is not only related to temporal properties, as shown in Fig. 8, but also reflected in the surface form. Observing the peaks of the normalized duration of words in the IPU-initial position, SYM words are skewed to the left whereas the other RTs are more centric. Initial shortening is clearly observable in SYM words, whereas final lengthening is observable in CAN, MSD, and NUM words. This is evidence that initial shortening may be caused by words of which the syllable merger form is preferred, which are mostly high-frequency words. Moreover, the peaks in IPU-final and -isolated tokens for SYM words remain centric, whereas the CAN, MSD, and NUM words skew more to the right because they have a longer duration. For example, for the frequently used word *yīn wèi* (meaning "because"), SYM is the predominant RT in the overall data and in all prosodic positions. However, word variants selected from different pools of data are distinct. It is /ʐ ei/ if the pools are from IPU-final and -isolated tokens of *yīn wèi*. However, the variants selected from the pools of IPU-initial and -medial tokens are /ʐ/ and /i/, respectively. In other words, to provide a full account of pronunciation modeling, contextual information such as prosodic position is necessary.

To statistically test the aforementioned effects, linear regression models were built. Because word duration was normalized relative to the word and speaker, we modeled the probability of a highly reduced token with a generalized mixed effect model with the normalized duration (NormDUR) as the observations, and the nominal RT and IPU-position groups as the fixed predictors. Compared with the null model, a main effect of the RT group exists [$F(3, 26708) = 692.82$, $p < 0.001$]. The differences between automatically derived RT groups are all statistically significant (all $p < 0.001$). Furthermore, we included predictors for the IPU-position group and the interactions between the IPU-position and RT groups by means of gradual addition. As suggested by Hanique *et al.* (2013), if models with more predictors or interactions added have a lower absolute Akaike Information Criterion (AIC) value (Akaike, 1973) than the same model without the particular predictor or interaction, the effects and across-group interactions may statistically be more effective. Comparing two models by using the simulated likelihood ratio test with 1000 replications revealed that a complex model with added predictors on the IPU-position group has a lower AIC value ($p < 0.001$). The statistically significant difference between the simpler and complex regression models strongly suggest that both the RT and IPU-position groups are crucial for predicting word duration.

Moreover, a complex model built with the interaction effect across the RT and IPU-position groups was significantly more accurate than the aforementioned regression models ($p < 0.05$). However, further analysis revealed nonsignificant differences between all combined, across RT and IPU-position group interactions and the specific added predictor IPU-medial group ($p > 0.1$). Our statistical results show that prosodic position is a critical factor related to the temporal property of spoken words in spontaneous speech, among which utterance-medial disyllabic words seem to have relatively unstable durations. Nevertheless, our results on the reduction types confirmed that the RTs classified according to the acoustic properties and phoneme sequence comparison reflected the differences in word duration.

C. Variants correlate with word types and phonetic similarity

According to the presented results, the typical variants might not always be phonetically similar to the citation form. The surface form is determined by reduction degree, production frequency, and word position in utterance (i.e., prosodic position). Compared with MSD and CAN words, SYM and NUM words with severe segment omission are phonetically more dissimilar to the citation form, as shown in Fig. 9(a). The density distribution of phonetic similarity in SYM and NUM words are similar, suggesting that speakers retain a similar degree of phonetic similarity to preserve a certain degree of intelligibility in the variant forms to distinguish these forms from other lexical items (Lindblom, 1990; Schuppler *et al.*, 2012).

Concerning the relationship between word type and phonetic similarity, there is always one particular word type that performs differently from the others, except for SYM words, as shown in Fig. 9(b). This means of examining reduced word forms facilitates identifying particular preferences of speech reduction from the perspective of word structure. For WI ($C_o\#O_n$), fewer segment substitutions are preferred for CAN words. Regarding the preboundary nasals

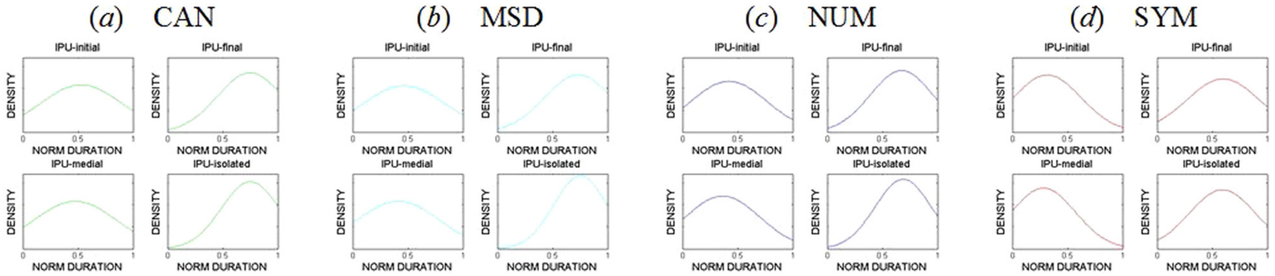

FIG. 8. (Color online) Duration patterns of RTs in four prosodic positions.

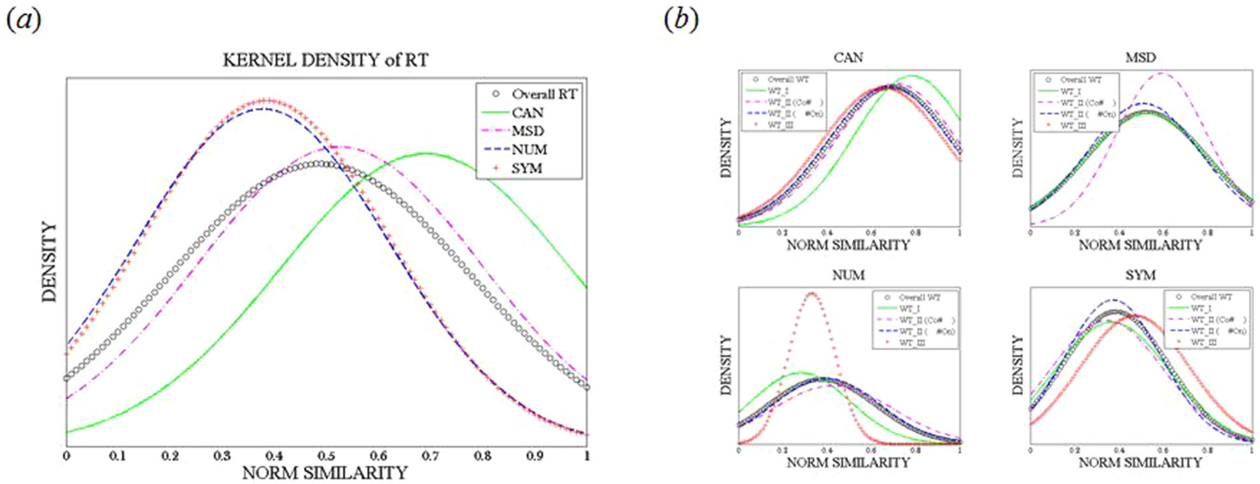

FIG. 9. (Color online) Kernel density of normalized similarity scores for (a) all RTs and (b) individual RT separated for different word types.

/n, ŋ/ for WII (C_o#∅), MSD words tend to be preserved in the form of glides rather than entirely omitted as those in W_{II} (∅#O_n) and W_I (C_o# O_n) are. In NUM words, the centric distribution in W_{III} (∅#∅) indicates that the variants are not as diverse as they are in the other word types.

To statistically test the above observations, we took the word type group as fixed predictors. Regression models for CAN, NUM, and SYM words were well represented with statistical significance ($p < 0.001$), except for the MSD words ($p > 0.05$), in the relationship between the reduced word forms and their distance of similarity from citation forms. Notably, for the CAN words, W_I (C_o#O_n) was statistically significant in the phonetic similarity scores ($p < 0.001$), whereas the other word types were nonsignificant (all $p > 0.01$). For the NUM words, W_{III} (∅#∅) was less explanatory ($p = 0.192$), whereas the other word types differed significantly in their similarity scores (all $p < 0.001$). Finally, the similarity scores of W_{II} (C_o#∅) and W_{III} (∅#∅) in the SYM words differed significantly, but seemingly shared a similar variation derivation pattern (all $p < 0.001$). The difference in the similarity scores of the other two word types in the SYM words was nonsignificant (both $p > 0.001$, but $p < 0.05$) and may suggest that the patterns in phonetic similarity were somehow more diverse.

D. Likely more than one typical variant

Most RT selected for disyllabic words might not always be significantly representative in terms of percentage. As shown in the Appendix, the coverage percentage ranges from 40% to 90%, implying that for certain words, multiple variants deviating from each other in the temporal and spectral domains (Hämäläinen *et al.*, 2009) are possible. For *méi yǒu* (negation), for example, the top two RTs are SYM and CAN, with 66.9% and 31.35% of coverage, respectively. The form /*m ə*/ is selected from SYM, and /*m e ou*/ is selected from CAN (Fig. 10). We observed that both variants are used very frequently in colloquial speech. Typically, the pronunciation dictionary in an ASR system contains only the citation form /*m ei j ou*/. In the ASR experiment, we added the most representative variant to the pronunciation dictionary; in this case, /*m ə*/. Typical variants may come from the same RT or belong to different RTs, seemingly depending on the RT coverage and prosodic position. More concretely, negation can be produced in a heavily reduced form, where the negation meaning is self-apparent from the context. Negation *méi yǒu* can also be prolonged when produced in a prosodically final position, accompanied with hesitation. A similar notion is proposed in the hypothesis of Hyper- and Hypo-speech

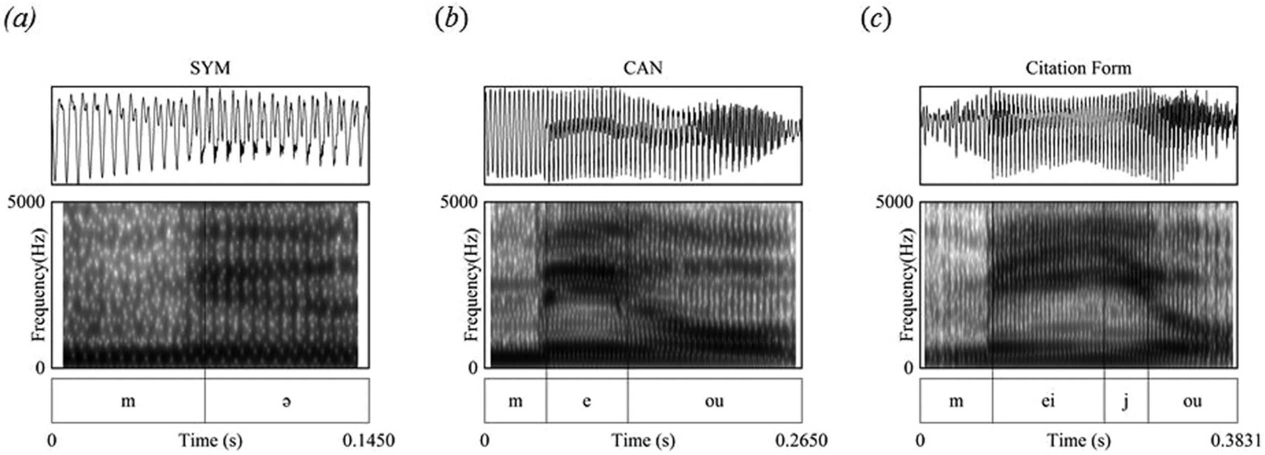

FIG. 10. Examples of typical variants selected for *méi yǒu* (negation): (a) SYM /*m ə*/; (b) CAN /*m e ou*/; and (c) the citation form /*m ei j ou*/.

(H&H) Theory (Lindblom, 1990) that speakers might use two distinct strategies to produce words in two contrary forces to maintain minimal self-articulatory effort and to satisfy a sufficient degree of discriminability for listener intelligibility of the uttered words (Schuppler *et al.*, 2012). The diverse means of articulation may result in varying degrees of reduced word forms in different prosodic positions. Thus, for some words, more than one typical variant form is possible. Currently, our algorithm selects only one typical variant for each word. However, in later refinement, we will establish a threshold to allow for selecting multiple variants.

E. Variant selection of realistic speech data

The MCDC8 was recorded in a lab environment. To test whether our variant selection algorithm was also robust for speech data in a realistic noise environment, we ran the variant selection procedure using a corpus of street interviews (Tseng, 2016) for the words listed in the Appendix. The same acoustic models used for the MCDC8 were also used for the Sociophonetic Interview Corpus (Tseng, 2016), which comprises 1402 street interviews recoded in 12 areas in Taiwan. It consists of speech recorded from 605 male and 797 female interviewees. The most frequently selected RTs of the 42 disyllabic words listed in the Appendix for the MCDC8 and Sociophonetic Interview Corpus are nearly the same, with only six words differing. However, the variants selected for each of the words differ. This has to do with discrepancies of speaker numbers and the number of word tokens produced by each speaker in the two corpora. Nevertheless, our selection algorithm through RT reflects intrinsic speech reduction processes related to linguistic systems. Considerations of knowledge-based linguistic structure and usage-based empirical information seem to provide a favorable solution for identifying important spoken word forms in casual speech data, which may also be relevant in the mental representation of spoken words. For instance, nine of the 42 words shown in the Appendix have identical variants selected for both corpora. These variants are all heavily reduced (SYM), such as *wǒ men* /ŋ/ (we) and *nǐ men* /n/ (plural of you). Neither can be directly derived using the phonological rules of the Edge-in Theory.

V. INCLUSION OF VARIANTS INTO ASR SYSTEM

We conducted two ASR experiments to assess the impact of variants selected using our method on a Chinese ASR system. In the first ASR experiment, we added variants for disyllabic words that occurred more than 20 times in the MCDC8, totaling 216 variants. In cases of CAN, variants were not included if they were identical to canonical forms. In the second experiment, we used the same datasets and parameter settings, but applied a data-driven statistical algorithm to include pronunciation variants into the ASR dictionary (i.e., the pruning method).

A. Speech data

After discarding laughter, silent pause, speech-like background noises, paralinguistic sounds, and non-Chinese words, 366 min of speech data were used to train the acoustic models. Word boundaries were manually verified. The evaluation set contained 41 min of speech data. As a result, the training and development sets comprised 14 630 IPUs, equivalent to 84 150 word tokens. The evaluation set contained 1630 IPUs, totaling 9383 word tokens. With 31 unseen words in the evaluation set, the number of words actually used in the dictionary of our ASR experiment was 6049, covering 99% of the words of the MCDC8.

B. Experiment setup

An ASR system normally models at least two probability distributions: (1) the probability of the acoustics (A) matching the particular hypothesized utterance (W), noted as $P(A|W)$, and (2) the prior probability of the hypothesized utterances $P(W)$. However, a commonly used method of considering variations in pronunciation and duration is to introduce an intermediate pronunciation model $P(V|W)$, thereby bridging acoustics models $P(A|V)$ and words $P(W)$. As shown in Eq. (5), the goal of an ASR system is to identify the string of words W and the corresponding variant strings V that maximize this objective function. The utterances used in our ASR experiment were the development (D_{EV}) and evaluation ($\mathrm{E}_{\mathrm{VAL}}$) sets, with the utterances in the D_{EV} but not in $\mathrm{E}_{\mathrm{VAL}}$ seen and trained in the language model. The recognition process operated on both sets was performed through an exhaustive search with a pronunciation model weight of 5 (α_R), as well as the empirically tuned word insertion penalty and weight on the language model (β).The latter two tuned parameters, namely, insertion penalty and weight on the language model, used for deriving the optimal recognition result from both evaluated utterance sets were 20 and 16, respectively. We fixed the pronunciation model weight to 5 in the ASR experiments because our task was not to optimize the weight on the pronunciation model but to test the performance of our variant inclusion in the pronunciation dictionary.

$$\hat{W} = \arg\max_{W}\max_{V}\left[P(A|V)P(V|W)^{\alpha R}P(W)^{\beta}\right]. \tag{5}$$

C. The pruning method

For comparison, we implemented a previously introduced method through the pronunciation pruning of our data (Tsai *et al.*, 2007). In contrast to the free phone recognition used in our variant selection, the experiment with the pruning method used "phone-level" forced recognition with a phone dictionary with prior probabilities to generate the surface form. Analogous to the *tf–idf* score for indexing terms in information retrieval, the *pf–iwf* score is used for pronunciation pruning. If a pronunciation $\mathrm{v_j}$ occurs more frequently for the word w_i, the *pf–iwf* score δ_{ij} is generally higher. It is lower if $\mathrm{v_j}$ appears more frequently in other words. With pronunciation variants ranked by *pf–iwf* score δ_{ij}, an adjustable threshold μ_S is used to select pronunciations to be included in the dictionary. In the following, we briefly introduce the terms used in the pruning method.

1. Pronunciation frequency (pf)

In terms of indexing pronunciations for words, the eligibility of pronunciation v_j for word w_i is measured by pf,

$$pf_{ij} = \frac{c_{ij}}{\sum_{allj} C_{ij}} = P\left(v_j|w_i\right), \tag{6}$$

where C_{ij} is the count of word w_i being pronounced as v_j, and $P(v_j \mid w_i)$ is the prior probability that w_i is pronounced as v_j. Typically, a pronunciation being observed more frequently for a word stipulates a higher correlation with the word.

2. Inverse word frequency (iwf)

While the pf presents the scores of varying pronunciations within a word, the iwf scores a particular pronunciation across different words. A pronunciation shared by many words leads to intense confusion for ASR systems; thus, such a pronunciation should be less eligible for inclusion in the dictionary. The iwf for a pronunciation is defined as

$$iwf_j = \log \frac{|\Omega|}{|\Omega_j|}, \tag{7}$$

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 the set. In Eq. (7), all of the words with pronunciation v_j are treated equally, but the resulting confusion depends on both of the frequencies of the words causing confusion and the probabilities that those words are pronounced as v_j. Therefore, the inverse word frequency (iwf) is redefined as follows:

$$iwf_j = \frac{1}{\sum_{w_k \in \{\Omega_j\}} P\left(v_j|w_k\right)P\left(w_k\right)} = \frac{1}{P\left(v_j\right)}, \tag{8}$$

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 observed 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.

$$\delta_{ij} = pf_{ij} \cdot (iwf_j)^{\gamma}, \tag{9}$$

γ is the adjustable weight for tuning the iwf score. When it is set to zero, it is reduced to the original pronunciation frequency, pf. In the case of equal unity, it is the mutual information between pronunciation v_j and word w_i.

D. Experimental results

For reporting the results, the measures of intrinsic and added confusability of a baseline lexicon and a newly enhanced pronunciation dictionary are used. The confusability of a dictionary is calculated according to the ratio of the number of words with confusing pronunciation over the number of words in the dictionary. Added confusability was defined as the ratio of the number of added variants shared by at least two distinct words over the number of words in the dictionary. ASR performance is presented in terms of character error rate (CER). Table IV summarizes the experimental results. The definitions of the "confusability of a dictionary" (*Conf. of Dic.*) and the "added confusability for added pronunciations" (*Added Confusion*) conform to those used in Tsai *et al.* (2007). The baseline performance in CER with optimized weights on word insertion penalty, pronunciation model, and language model for the $\mathrm{D_{EV}}$ and $\mathrm{E_{VAL}}$ sets is 74.27% and 65.68%, respectively. The difference is due to the training of language model using the IPU content in the $\mathrm{D_{EV}}$ set. To evaluate the impact of variants selected using our algorithm and the pruning method, the settings in Table IV (B)–(D) were used. In our proposed selection algorithm (B), the number of added variants is 216. Improvement in ASR performance was achieved, with the CER reduced by 1.54% and 1.21% for the $\mathrm{D_{EV}}$ and $\mathrm{E_{VAL}}$ sets, respectively. The results presented in Table IV (C) are the most favorable results obtained after tuning the weight γ and putting a single threshold on pruning (such as $\delta_{ij} > \mu_S$) for our data. Controlling the pruning threshold on pf–iwf score δ_{ij} for a similar level of added confusability as that in the case of our method, the ASR performance improved for the $\mathrm{D_{EV}}$ data but not for the $\mathrm{E_{VAL}}$ data. Because a large number of variants (12 667) were added to the pronunciation dictionary, the high confusion may have affected the correct word chosen for the unseen IPU content. Thus, we lowered the pruning threshold on pf–iwf score δ_{ij} to select only the 216 best variants, as done in our approach (B). In (D), the dictionary

TABLE IV. Character Error Rate (CER) with disyllabic word variants selected by our method and those by pruning method.

	Surface Form Generation	Variant Selection	Dev. CER (%)	Eval. CER (%)	Conf. of Dict. (%)	Added Conf. (%)	Added Variants (#)	Added Variants that Conf. (#)
(A)	Baseline: Canonical dictionary		65.68	74.27	22.19	—	—	—
(B)	Free phone recognition	Proposed method	64.14	73.06	23.57	0.96	216	58
(C)	Phone-level forced recog.	pf-iwf ($\gamma = 0.8$, $\mu_S = 0.88$)	64.30	77.01	23.62	0.94	12,667	57
(D)	Phone-level forced recog.	pf-iwf (Top N added variants)	66.30	74.93	22.28	0.05	216	3

confusion increased slightly, and the performance did not improve. However, the noise of word recognition on the unseen data (E_{VAL} set) was clearly reduced by controlling the number of added variants. Table IV shows that after the variants were included, the confusability increased, but the ASR performance was improved, suggesting that the variants selected by our algorithm may to some degree represent typical spoken word forms in reduced, spontaneous speech, thus improving ASR performance.

E. Discussion on added variants in ASR system

Because poorly selected variants may be included in the recognition processes or in data-driven decision trees, numerous approaches have been proposed to reject variants that are highly confusable by using phoneme confusability matrices (Wester, 2003; Tsai *et al.*, 2007). Thus, confusability is used as a key feature in advanced processing on the *pf–iwf* (Tsai *et al.*, 2007), or phonological rule constraints are used to exclude improper variants by adopting measures such as logarithmic likelihood-based criterion (Amdal *et al.*, 2000), entropy (Yang *et al.*, 2002), and absolute or relative frequency (Kessens *et al.*, 2003). Alternatively, variants can be implicitly handled in embedded, discriminatively learned probabilistic models (Jyothi *et al.*, 2013; Karanasou *et al.*, 2013) and pronunciation models in a weighted finite-state transducer for a speech recognizer (Jyothi *et al.*, 2013; Karanasou *et al.*, 2013; McGraw *et al.*, 2013). For resolving word variant problems, manual phonetic transcription (Oostdijk, 2002; Maekawa, 2003; Pitt *et al.*, 2005; Van Bael *et al.*, 2007) has been shown to facilitate shortlisting the most frequently produced phonetic forms. However, conducting projects of this scale to guarantee the reliability of data annotation is expensive. Our approach is based on the acoustic properties of word tokens with consideration of language-dependent knowledge on reduction degree and word type. We attempted to allocate a linguistic unit (word) that preserves a certain degree of invariance between the meaning and phonetic form, which differs from the strategies used in previous studies for deriving variants with pruning measures or with a discriminant or generative weight-learning on pronunciation modeling, which focus mainly on the phonemic level for pronunciation modeling.

VI. CONCLUSION

We observed an improvement in ASR performance after adding only a small number of disyllabic word variants to the pronunciation dictionary. Although we evaluated only those variants selected by our algorithm on a relatively small spontaneous speech data set, the results revealed that the categorical RTs we proposed did improve ASR system performance. CAN and SYM are preferred for Chinese disyllabic words, comprising over 75% of all RTs. For the next step of research focus, we may reconsider reduction in spontaneous speech in a more complex manner, encompassing production frequency, prosodic position, reduction degree, and additional language-dependent properties (Bell *et al.*, 2003). For future works, we will extend the scope from disyllabic words to any disyllables to study whether the proposed RT classification is also applicable to nondisyllabic words. Viewing word variants from the cognitive perspective, we will also test whether the variants selected using our method have a certain degree of psychological reality by conducting perceptual experiments.

ACKNOWLEDGMENTS

The authors sincerely thank the National Digital Archives Project and the National Science Council, under Grant No. 103-2410-H-001-063-MY2 for financial support.

APPENDIX

For information about the reduction types and variants selected by the proposed algorithm for 42 most frequent disyllabic words in the MCDC8, please see Table V.

TABLE V. Information about word count, WT, RT and variant form of 42 disyllabic words that occur more than 100 times in the MCDC8 are listed for the MCDC8 and the Socio-phonetic Interview Corpus (Socio).

Word	Gloss	MCDC8 Count	WT	Citation	MCDC8RT	MCDC8 RT (%)	MCDC8 Variant	Socio Count	Socio RT	Socio RT (%)	Socio Variant
jiù shì	that is	1039	WII	/ʨ j ou ʂ ʉ/	SYM	50.82	/tɕ ʉ/	3185	SYM	47.66	/ɨ/
wǒ men	we	878	WII	/w o m ə n/	SYM	84.97	/ŋ/	579	SYM	73.40	/ŋ/
rán hòu	then	733	WI	/ʐ a n x ou/	SYM	79.81	/t au/	2252	SYM	77.13	/l au/
jué dé	to feel	676	WII	/ʨ ye t ə/	SYM	64.64	/tɕ ye/	794	SYM	62.09	/tɕ ye/
yīn wèi	because	651	WII	/i n w ei/	SYM	67.28	/ʐ ei/	1024	SYM	73.14	/ʐ /
méi yǒu	not/have not	571	WIII	/m ei j ou/	SYM	66.90	/m ə/	3820	SYM	68.43	/ʐ au/
suǒ yǐ	so	479	WIII	/s w o i/	SYM	90.61	/ts ei/	413	SYM	84.75	/ts ei/
kě shì	but	466	WII	/k^h ə ʂ ʉ/	SYM	61.16	/k^h ʉ/	421	SYM	58.91	/ɨ/
qí shí	in fact	431	WII	/ʨ^h i ʂ ʉ/	SYM	54.06	/tɕ^h y/	198	SYM	60.10	/ɨ/
shí hòu	time	420	WII	/ʂ ʉ x ou/	SYM	83.10	/ʂ ou/	707	SYM	79.92	/ts ou/
tā men	they	415	WII	/t^h a m ə n/	SYM	83.37	/t^h a ŋ/	516	SYM	69.77	/ɨ/
xiàn zài	now	397	WI	/ɕ j e n ts ai/	SYM	69.02	/ɕ j ai/	746	SYM	64.34	/ʂ j ai/
bǐ jiào	more	371	WII	/p i ʨ j au/	SYM	63.34	/tɕ j au/	1902	SYM	68.19	/p j au/
shé me	what	346	WII	/ʂ ə m ə/	SYM	90.75	/s ŋ/	681	SYM	88.11	/p/
zhè yàng	in this way	339	WIII	/tʂ ə j a ŋ/	SYM	87.02	/tɕ j a/	885	SYM	94.58	/p/

TABLE V. *Continued*

Word	Gloss	MCDC8 Count	WT	Citation	MCDC8RT	MCDC8 RT (%)	MCDC8 Variant	Socio Count	Socio RT	Socio RT (%)	Socio Variant
zhēn de	really	281	WI	/tʂ ə n t ə/	SYM	64.77	/tʂ ə/	58	SYM	62.07	/tʂ ə/
nà biān	there	269	WII	/n a p j e n/	SYM	46.84	/n/	246	SYM	44.72	/ɨ/
kě yǐ	can	260	WIII	/kʰ ə i/	SYM	85.00	/kʰ y/	251	SYM	73.71	/ɨ/
hǎo xiàng	it seems	259	WII	/x au ɕ j a ŋ/	SYM	57.53	/x a/	392	SYM	64.03	/ŋ/
zhī dào	to know	234	WII	/tʂ i t au/	SYM	51.28	/ts au/	513	SYM	59.26	/ɨ/
hěn duō	a lot	232	WI	/x ə n t w o/	SYM	43.53	/ŋ t ou/	185	SYM	51.35	/p w/
kě néng	possibly	223	WII	/kʰ ə n ə ŋ/	SYM	84.75	/kʰ ə/	788	SYM	79.06	/kʰ ə/
hái shì	still/or	218	WII	/x ai ʂ ʉ/	CAN	47.71	/x ai ʂ ɨ/	510	SYM	42.55	/ɨ/
yì xiē	Some	202	WII	/i ɕ j e/	CAN	62.87	/i ɕ j e/	1244	CAN	61.82	/i ɕ ye/
bú huì	will not	199	WII	/p u x w ei/	SYM	64.32	/p w ei/	2418	SYM	59.76	/p/
ér qiě	however	194	WII	/ɚ tɕʰ j e/	CAN	59.28	/ɚ ɕ ye/	24	CAN	79.17	/ɚ tɕ y/
xué xiào	school	172	WII	/ɕ ye ɕ j au/	SYM	55.81	/ɕ j au/	117	CAN	61.54	/tɕʰ ye ɕ j au/
yīng gāi	should	168	WI	/i ŋ k ai/	SYM	41.07	/n ə/	1528	SYM	64.99	/ɨ/
zì jǐ	-self	168	WII	/ts ɨ tɕ i/	SYM	62.50	/tʂ i/	547	SYM	61.79	/tʂ i/
dōng xī	thing	149	WI	/t o ŋ ɕ i/	MSD	41.61	/w ɕ i/	307	SYM	47.56	/t ŋ s ɨ/
yǐ qián	before	147	WII	/i tɕʰ j e n/	CAN	78.91	/i ɕ ye n/	144	CAN	69.44	/i tɕ ye/
rú guǒ	if	139	WII	/ʐ u k w o/	SYM	81.29	/w o/	203	SYM	67.49	/u/
nǐ men	you (plural)	138	WII	/n i m ə n/	SYM	85.51	/n/	36	SYM	88.89	/n/
dàn shì	but	137	WI	/t a n ʂ ʉ/	SYM	44.53	/t ə/	321	MSD	43.61	/ɚ ʂ ɨ/
lǐ miàn	inside	126	WII	/l i m j e n/	CAN	60.32	/i m e/	70	SYM	64.29	/n e/
zěn me	how	124	WI	/ts ə n m ə/	SYM	91.13	/ts ŋ/	52	SYM	84,62	/ts n/
gōng sī	company	121	WI	/k o ŋ s ɨ/	MSD	46.28	/k u ʂ ɨ/	135	MSD	42.22	/k ou ʂ ɨ/
yǐ jīng	already	117	WII	/i tɕ i ŋ/	SYM	49.57	/i/	127	SYM	55,91	/ɨ/
yí yàng	the same	115	WIII	/i j a ŋ/	CAN	54.78	/i j a/	440	SYM	45.23	/ɨ/
rén jiā	people	112	WI	/ʐ ə n tɕ j a/	SYM	51.79	/n j ai/	199	SYM	60.30	/n ə/
dà gài	probably	108	WII	/t a k ai/	SYM	62.96	/t ai/	835	SYM	55.21	/t ai/
dà jiā	everyone	100	WII	/t a tɕ j a/	SYM	72.00	/t a/	38	SYM	73.68	/t a/

Akaike, H. (**1973**). "Information theory and an extension of the maximum likelihood principle," in *Proceedings of ISIT-1973*, Budapest, Hungary, pp. 267–281.

Akita, Y., and Kawahara, T. (**2010**). "Statistical transformation of language and pronunciation models for spontaneous speech recognition," IEEE Trans. Audio Speech Lang. Process. **18**, 1539–1549.

Amdal, I., Korkmazskiy, F., and Surendran, A. C. (**2000**). "Joint pronunciation modelling of non-native speakers using data-driven methods," in *Proceedings of ICSLP-2000*, Beijing, China, pp. 622–625.

Association for Computational Linguistics and Chinese Language Processing (**2016**). http://www.aclclp.org.tw/ (Last viewed June 17, 2016).

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (**2003**). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," J. Acoust. Soc. Am. **113**, 1001–1024.

Byrne, W., Venkataramani, V., Kamm, T., Zheng, T. F., Song, Z., Fung, P., Liu, Y., and Ruhi, U. (**2001**). "Automatic generation of pronunciation lexicons for Mandarin spontaneous speech," in *Proceedings of ICASSP-2001*, Salt Lake City, Utah, pp. 569–572.

Chen, K., and Hasegawa-Johnson, M. (**2004**). "Modeling pronunciation variation using artificial neural networks for English spontaneous speech," in *Proceedings of Interspeech-2004*, Jeju Island, South Korea, pp. 400–403.

Chien, J.-T., and Huang, C.-H. (**2003**). "Bayesian learning of speech duration models," IEEE Trans. Speech Audio Process. **11**, 558–567.

Chung, R.-F. (**1997**). "Syllable contraction in Chinese," in *Chinese Languages and Linguistics III: Morphology and Lexicon*, edited by F.-F. Tsao and H. Samuel Wang (Academia Sinica, Taipei, Taiwan), pp. 199–235.

Connolly, J. H. (**1997**). "Quantifying target-realization differences: Part I: Segments," Clin. Linguist. Phonetics **11**, 267–287.

Dilley, L., and Pitt, M. (**2007**). "A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition," J. Acoust. Soc. Am. **122**, 2340–2353.

Fosler-Lussier, E. (**1999**). "Dynamic pronunciation models for automatic speech recognition," Ph.D. dissertation, International Computer Science Institute, University of California, Berkeley, CA, pp. 84–88.

Fukada, T., Yoshimura, T., and Sagisaka, Y. (**1999**). "Automatic generation of multiple pronunciations based on neural networks," Speech Commun. **27**, 63–73.

Hämäläinen, A., Gubian, M., ten Bosch, L., and Boves, L. (**2009**). "Analysis of acoustic reduction using spectral similarity measures," J. Acoust. Soc. Am. **126**, 3227–3235.

Hanique, I., Ernestus, M., and Schuppler, B. (**2013**). "Informal speech processes can be categorical in nature, even if they affect many different words," J. Acoust. Soc. Am. **133**, 1644–1655.

Hofmann, H., Sakti, S., Isotani, R., Kawai, H., Nakamura, S., and Minker, W. (**2010**). "Improving spontaneous English ASR using a joint-sequence pronunciation model," in *Proceedings of IUCS-2010*, Beijing, China, pp. 58–61.

Hsu, H.-C. (**2003**). "A sonority model of syllable contraction in Taiwanese Southern Min," J. East Asian Linguist. **12**, 349–377.

Jyothi, P., Fosler-Lussier, E., and Livescu, K. (**2013**). "Discriminative training of WFST factors with application to pronunciation modeling," in *Proceedings of Interspeech-2013*, Lyon, France, pp. 1961–1965.

Karanasou, P., Yvon, F., Lavergne, T., and Lamel, L. (**2013**). "Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR," in *Proceedings of Interspeech-2013*, Lyon, France, pp. 1966–1970.

Kessens, J. M., Cucchiarini, C., and Strik, H. (**2003**). "A data-driven method for modeling pronunciation variation," Speech Commun. **40**, 517–534.

Kondrak, G. (**2003**). "Phonetic alignment and similarity," Comput. Hum. **37**, 273–291.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (**2008**). "Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e)," J. Philos. Trans. R. Soc. B **363**, 979–1000.

Ladefoged, P. (**2006**). *A Course in Phonetics* (Thomson Wadsworth, Boston, MA), Chap. 6–9, pp. 133–236.

Lindblom, B. (**1990**). “Explaining phonetic variation: A sketch of the H&H theory,” in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Springer, New York), pp. 403–439.

Liu, Y., and Fung, P. (**2004a**). “Pronunciation modeling for spontaneous Mandarin speech recognition,” Int. J. Speech Technol. **7**, 155–172.

Liu, Y., and Fung, P. (**2004b**). “State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition,” IEEE Trans. Speech Audio Process. **12**, 351–364.

Liu, Y.-F., Tseng, S.-C., and Jang, R. J.-S. (**2014**). “Phone boundary annotation in conversational speech,” in *Proceedings of LREC-2014*, Reykjavik, Iceland, pp. 848–853.

Lobanov, B. M. (**1971**). “Classification of Russian vowels spoken by different speakers,” J. Acoust. Soc. Am. **49**, 606–608.

Ma, W.-Y., and Chen, K.-J. (**2004**). “Design of CKIP Chinese word segmentation system,” Int. J. Asian Lang. Process. **14**, 235–249.

Maekawa, K. (**2003**). “Corpus of spontaneous Japanese: Its design and evaluation,” in *Proceedings of SSPR-2003*, Tokyo, Japan, pp. 7–12.

McGraw, I., Badr, I., and Glass, J. R. (**2013**). “Learning lexicons from speech using a pronunciation mixture model,” IEEE Trans. Audio Speech Lang. Process. **21**, 357–366.

Oostdijk, N. (**2002**). “The design of Spoken Dutch Corpus,” in *New Frontiers of Corpus Research*, edited by P. Peters, P. Collins, and A. Smith (Rodopi, Amsterdam, the Netherlands), pp. 105–112.

Pierrehumbert, J. (**1994**). “Knowledge of variation,” in *CLS 30 Vol. 2: Papers From The Parasession on Variation*, edited by K. Beals, J. Denton, R. Knippen, L. Melnar, H. Suzuki, and E. Zeinfeld (Chicago, IL), pp. 232–256.

Pitt, M. A., Dilley, L., and Tat, M. (**2011**). “Exploring the role of exposure frequency in recognizing pronunciation variants,” J. Phonetics **39**, 304–311.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (**2005**). “The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability,” Speech Commun. **45**, 89–95.

Schuppler, B., Ernestus, M., Scharenborg, O., and Boves, L. (**2011**). “Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions,” J. Phonetics **39**, 96–109.

Schuppler, B., van Dommelen, W. A., Koreman, J., and Ernestus, M. (**2012**). “How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level,” J. Phonetics **40**, 595–607.

Stolcke, A. (**2002**). “SRILM-an extensible language modeling toolkit,” in *Proceedings of ICSLP-2002*, Denver, CO, pp. 901–904.

Torreira, F., and Ernestus, M. (**2011**). “Vowel elision in casual French: The case of vowel /e/ in the word *c'était*,” J. Phonetics **39**, 50–58.

Tsai, M.-Y., Chou, F.-C., and Lee, L.-S. (**2007**). “Pronunciation modeling with reduced confusion for Mandarin Chinese using a three-stage framework,” IEEE Trans. Audio Speech Lang. Process. **15**, 661–675.

Tseng, S.-C. (**2013a**). “Lexical coverage in Taiwan Mandarin conversation,” Int. J. Comput. Linguist. Chinese Lang. Process. **18**, 1–18.

Tseng, S.-C. (**2016**). “/kwo/ and / y/ in Taiwan Mandarin: Social factors and phonetic variation,” Lang. Linguist. **17**, 383–405.

Tseng, S.-C., Soemer, A., and Lee, T.-L. (**2013b**). “Tones of reduced T1-T4 Mandarin disyllables,” Int. J. Comput. Linguist. Chinese Lang. Process. **18**, 81–106.

Van Bael, C., Baayen, H., and Strik, H. (**2007**). “Segment deletion in spontaneous speech: A corpus study using Mixed Effects Models with crossed random effects,” in *Proceedings of Interspeech-2007*, Antwerp, Belgium, pp. 2741–2744.

Wester, M. (**2003**). “Pronunciation modeling for ASR – knowledge-based and data-derived methods,” Comput. Speech Lang. **17**, 69–85.

Yang, Q., Martens, J.-P., Ghesquiere, P.-J., and Van Compernolle, D. (**2002**). “Pronunciation variation modeling for ASR: Large improvements are possible but small ones are likely to achieve,” in *Proceedings of PMLA-2002*, Estes Park, CO, pp. 123–128.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2006**). *The HTK Book 3.4* (Cambridge University Press, London, UK).