Generation and selection of English pronunciation variations using knowledge-based rules and speech recognition techniques*

Tae-Yeoub Jang
(Hankuk University of Foreign Studies)

Jang, Tae-Yeoub. 2006. Generation and selection of English pronunciation variations using knowledge-based rules and speech recognition techniques. *Studies in Phonetics, Phonology and Morphology* 12.2. 361-375. This paper describes a method of designing a pronunciation dictionary which contains an optimal number of pronunciation variants for each word entry. The two key processes, generation and selection of variants, are designed to be performed automatically so that the final version of pronunciation dictionary can be practically used in speech recognition systems. For generation of variants both *prescriptive* and *descriptive* rule extraction methods are employed. First, various optional phonological processes are extracted from literature. Second, phone strings obtained by hand labelling are compared with base-form pronunciation strings and the systematic differences are set as rules. After generation of as many variants as possible, it is determined whether each variant deserves to be kept in the dictionary through relative frequency measure and heuristic decision-making criteria. Through experiments, the variant dictionary is found to play a meaningful role in enhancing performance of automatic speech recognition. **(Hankuk University of Foreign Studies)**

Keywords: pronunciation variation, pronunciation modelling, pronunciation dictionary, phone recognition, auto-labelling, auto-segmentation

1. Introduction

Pronunciation of human speech varies. Even when a single speaker repeats a single word or sentence, it is almost impossible to produce exactly the same acoustic signals. Some variations do not cause any serious trouble in human speech comprehension while others make interlocutors more perceptually attentive to maintain the conversation exchange.

When the target listener of human utterances is not another human but a machine, pronunciation variations have even greater negative influence on decoding performance. To cope with this difficulty, it has been suggested for automatic speech recognition (ASR, henceforth) systems to be equipped with a pronunciation dictionary with pronunciation variants. Many studies have acknowledged that expanding the pronunciation dictionary, used for decoding speech signals into linguistic units in recognition processes, is a fruitful solution to the problem. But their approaches and methods are not uniform. Heine *et al*. (1998) calculate probability for each pronunciation variant and directly use this information in recognition. Bonaventura *et al*. (1998) and Goronzy *et al*. (2004) extend the variation dictionary by including non-native

* This work was supported by Korea Research Foundation Grant (KRF-2003-041-A00213).

pronunciations while Ward *et al*. (2003) focuses on within-language speaker variability. Instead of directly manipulating the word-level pronunciation dictionary, Nock and Young (1998) and Binnenpoorte *et al*. (2005) introduce multi-word units in order to capture cross-word variations, although difficulties in defining the multi-word lexical module and preliminary *N*-gram processing are additional burdens of their method. Yang and Martens (2000), Kessens *et al*. (2003) and Wester (2003) emphasise importance of the selection process proving that simply adding variants to lexicon may fail to increase ASR performance.

Most studies mentioned above adopt data-driven methods for generating and selecting pronunciation variations. Although it is true that the data-driven approach is quite useful for decreasing word error rate of speech recognition, that effect does not normally last when experimental circumstances change. On the other hand, variation modelling in terms of knowledge-based approaches using linguistic information is usually more self-reliant and system independent in that models set in a situation can still be used without any major change in other circumstances. A disadvantage of knowledge-based approaches appears that its effect may not be observable immediately.

In the current study, both knowledge-based methods and data-driven methods are adopted to achieve instant performance improvement and robustness as well. The variant generation process will mainly take advantage of knowledge-based methods while the selection process will be based more on data-driven methodology. These operations will be described in detail followed by a report on speech recognition experiment conducted to check whether such processes are useful.

2. Data

For the most part of the experiment, the speech corpus named *English02* provided by Speech Information TEChnology (SiTEC) & Industry Promotion Center is used. The corpus is composed of 3,678 sentence tokens read in quiet office environments by 300 gender-balanced native speakers of American English. The sentences were recorded and digitised at a sampling rate of 16KHz with 16-bit resolution. The numbers of unique words and sentences are 233 and 124, respectively.[1]

This corpus is suitable for the current study as its sentence tokens are designed to contain many word strings whose edge phones frequently lose the original quality in connected speech. In particular, various words with consonant clusters at both the initial and final position are chosen to constitute sentences so that voice actors/actresses are expected to produce different kinds of phonetic processes regarding inter-word linking.

[1] The original *English02* corpus has been produced by 400 speakers and has more sentence tokens, but I did not obtain the entire database sets that the amount of the data used in the current experiment is not the same as the amount in the original version. For more details of the corpus, visit http://www.sitec.or.kr/English/corpus.asp.

The sentence tokens are subdivided into two sets: SET-1 with 2,942 tokens for training an automatic speech recogniser (see 6.1) and SET-2 with 736 tokens for testing the recognition performance with different-version pronunciation dictionaries which will be constructed.

3. Base dictionary

Prior to comprising a dictionary with pronunciation variations, it is necessary to prepare a base dictionary (DICT_base) first which contains the canonical pronunciation for each word entry. As the current goal of dictionary expansion is not to construct a corpus free pronunciation dictionary for unlimited speech recognition system, but to show whether the automatic method is effective in speech recognition performance, building a base dictionary composed of corpus-internal words and a typical pronunciation for each word item suffices.

A publically available online pronunciation dictionary called *The CMU Pronouncing Dictionary* (v0.6) is used for this work.[2] For each word in the word list consisting of 233 words, its corresponding pronunciation string in the CMU dictionary is copied to compose the corpus-specific base dictionary. Some words in the CMU dictionary have more than one pronunciation strings but some of them are not so much pronunciation variants of a single word as base pronunciations of two different words, also known as heteronyms (*e.g.*, 'lead' [lid] and 'lead' [lɛd]). After copying all the relevant phone strings, necessary character conversion is performed since there is a slight mismatch between the CMU phone units and the phone units in the current study.

In this way, the base dictionary has been formulated as in Table 1.

Table 1. Example of base dictionary (DICT_base)

Word	Baseform Pronunciation
a	ey
about	ə b aw t
appointment	ə p oy n t m ə n t
later	l ey t ə r
yourself	y ɚ s ɛ l f

Note that each base-form pronunciation does not have to be the most frequent pronunciation. For example, the most frequent pronunciation variant of the word 'a' is found to be [ə] rather than [ey] in the phonetically segmented portion of the Switchboard Transcription Corpus (Greenberg 1997). As the base pronunciation dictionary is designed to not reflect dynamicity such as vowel reduction or morphological weakness (as a function word), it has a limited practical role.

[2] Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

4. Generation of variants

Once the base dictionary is created, the next step is generating pronunciation variants for each word. The key to find as many variants as possible is extracting rules first rather than merely finding variants specific to each word. Those rules are applied to the pronunciation strings for each word in the base dictionary to generate a greater number of variants in a short time.

The problem is how to extract these rules. It is not desirable for human researchers and/or native speaker reviewers to pore over each word trying to find all possible variants of it, not only because of length of time it will take but because of the apparent inability of human perception to detect minute but significant variations.

In the current study, two different methods of discovering rules are adopted: (1) *prescriptive* and (2) *descriptive* rule extraction. I will call the rules extracted by those two methods RULES_pre and RULES_des, respectively.

4.1 Extraction of prescriptive rules

The top-down approach is a knowledge-based method using phonological rules or constraints of which the application is optional. These processes can be easily extracted through a variety of theory- or course-books on English phonology and phonetics. The literature used in the current study includes: Kreidler (1989), Kenstovicz (1994), Roach (2004), and Silverman (2006). The obvious advantage of this type of rules is generality. As target sounds as well as their environment(s) are usually described in terms of natural classes instead of individual segments, base-form pronunciation strings can be easily expanded into a number of variants.

Table 2 shows the rules extracted in this way and their examples.

Table 2. Examples of RULES_pre

Name	Example	Orthography
Coalescence	m ɪ s # y u → m ɪ ʃ u	miss you
Consonant deletion (1)	æ k t s → æ k s	acts
Consonant deletion (2)	æ n d → æ n	and
Flapping (across-word level)	f ə r g ɛ t → f ə r g ɛ ɾ	forget
Flapping (within-word level)	l ey t ə r → l ey ɾ ə r	later
Glide deletion	n y u → n u	new
Glide insertion	y u ʒ u ə l → y u ʒ u w ə l	usual
Metathesis	dʒ u ə l r i → dʒ u l ə r i	jewelry
Syllabic consonant formation	ow p ə n → ow p m	open
Vowel deletion	ə s p ɛ ʃ ə l i → ə s p ɛ ʃ l i	especially
Vowel Reduction (1)	b ɪ h ay n d → b ə h ay n d	behind
Vowel Reduction (2)	ey l i ə n → ey l y ə n	alien

4.2 Extraction of descriptive rules

Apart from the RULES_pre, there are a variety of detailed phonetic processes which may not be perceptually distinctive but acoustically effective enough to deteriorate performance of speech processing systems. The bottom-up approach is useful for finding such rules. A portion of data tokens are segmented and labelled manually by nine human labellers.

Seven participants are postgraduate students majoring in phonetics and the other two are PhDs in phonetics. Preceded by practice and tuning sessions for inter-labeller consistency, a total of 504 sentence tokens are labelled to be used in extracting RULES_pre. When phone labels are obtained, phone strings for each word are collected and compared to the corresponding phone string in the DICT_base. Unless that candidate string is judged to be an ignorable pronunciation error (*e. g.*, slips of tongue) or unusual idiosyncracy, and unless it occurs only once, a rule is established based on the phonetic context of the string. For example, there are several cases of hand-labelled string [w ə s], for the word 'was'. As its corresponding string in the DICT_base is [w ə z], a rewrite rule 'w ə z # → w ə s #' can be extracted. Table 3 shows some of such rules.

Table 3. Examples of RULES_des

Name	Example	Orthography
Consonant Devoicing	w ə z → w ə s	was
Despirantisation	ð ey → d ey	(but #) they
Monophthongisation	aw əʶ → a əʶ	our
Spirantisation	dʒ u s → ʒ u s	(orange #) juice

It is not simple to determine whether an observed phone string belongs to a pronunciation error or a systematic variation. A makeshift criterion used in the current study is: the same phenomenon is observed (1) at least three or more times, and (2) in utterances by two or more speakers. As a matter of fact, employing a stricter criterion does not necessarily enhance performance since the rule selection process, described in section 5, is expected to function as a substantial filter of generated variants.

Note some rules are applicable beyond word internal level (*e.g.*, sprantisation). This type of rule makes it possible for a pronunciation dictionary to cover, if not completely, phonological/phonetic processes occurring across word boundaries, helping an automatic speech recogniser better decode connected speech signals into linguistic units.

One characteristic that distinguishes RULE_des from RULE_pre is locality, implying that, in case of RULE_des, environments for rule application are less general than those of RULE_pre. Consequently, a RULE_des will usually not generate as many variants as a RULE_pre.

4.3. Automatic generation

When the list of rules is prepared, the process of variant generation is quite simple. The pronunciation strings in the base dictionary are forced to pass through the generator which contains all the rules extracted. Whenever, each string is affected by a rule, a new pronunciation variant is created.

The result of the operation is a pronunciation dictionary with a maximum number of variants (DICT_maxvar) for each word, as exemplified in Table 4.

Table 4. Examples of maximum variation dictionary (DICT_maxvar)

Word	Variants
a	ey
	ə
appointment	ə p oy n t m ə n t
	ə p oy n t m ə n
	ə p oy n m ə n t
	ə p oy n m ə n
meet	m i t
	m i ɾ
	m i tʃ
yourself	y ɚ s ɛ l f
	y ɔ r s ɛ l f
	y ə r s ɛ l f
	y r s ɛ l f
	y ʊ r s ɛ l f

Note also that if a pronunciation string undergoes two or more rules the number of variants increase exponentially, making longer words produce more variants.

The total number of variants in the DICT_maxvar is 383, resulting in 64.4% increase in size from the base-form dictionary (DICT_base) with 233 variants. 85 words (or 36.5%) are affected by one or more rules producing at least a variant different from the canonical pronunciation.

5. Selection of variants

5.1 Reasons for selection

The final step is to reduce the size of DICT_maxvar by discarding superfluous variants. As mentioned in Kessens *et al*. (2003), the reasons for constraining the number of variants can be summarised as: (1) reducing confusability of the lexicon[3], (2) alleviating computation overhead, and (3) maintaining linguistic plausibility.

[3] Based on Byrne *et al*. (1997:2), the confusability problem can be illustrated as: the more variants of words the dictionary has, the more homophone cases occur (*e.g.*, [kaz] as a variant of both words 'cause' and 'because').

Considering the size of the dictionaries in the current study, (1) and (2) may not be critical issues since the two dictionaries DICT_maxvar (383 entries) and DICT_base (233 entries) do not make a great difference either in confusability or in computation overhead at the time of recognition. However, as the method described in the current study is designed to work regardless of the size or type of the corpus, those challenges are still valid. On the other hand, (3), which is more relevant to the current method, warns that maintaining automatic methods may result in only machine-friendly content concealing the significance of linguistically verified ones. In terms of pronunciation variation, variants might be included in the lexicon on non-plausible artefacts of the speech recognition system instead of being based on genuine pronunciation variation.

5.2 Rules or variants?

There are two ways of constraining variability: one, dismissing rules, the other, discarding variants. When rules are constrained in a way that less applicable rules are abandoned, variants generated by them should be removed as well in a lump from the dictionary. Although this is a powerful and fast way to reduce the size of the dictionary, its disadvantage is lack of elaborateness. Imagine that a rule optionally converts a base-form pronunciation into some other string producing a variant and that only 0.3% of such string tokens are found to undergo conversion for utterances produced by a speaker while the converted pronunciation can appear a lot more frequently by another speaker. Then, it is difficult to decide whether the rule needs to be maintained or not. When dealing with variants, local decisions can be made: *e.g.*, a variant can be discarded or maintained individually.

Consequently, the variant selection, instead of rule selection scheme is adopted in the current study.

5.3 Relative frequency measure

The selection process is based on the relative frequency of each variant of a word. First, the number of cases that each variant is chosen by an automatic phone recogniser is counted, and then its rate is calculated, relative to the total number of variants for the same word. The process can be represented as in (1):

(1) $P(V_i|W_j) = C(V_i) / C(V_{i..N})$
where C() stands for the counting function

Take the word item ‘behind’ for example. It has four variants including the base pronunciation [bɪhaynd], as shown in Table 5. When a set of speech tokens is provided together with the variation dictionary (DICT_maxvar) and orthographic transcription for each token, the ready-made phone

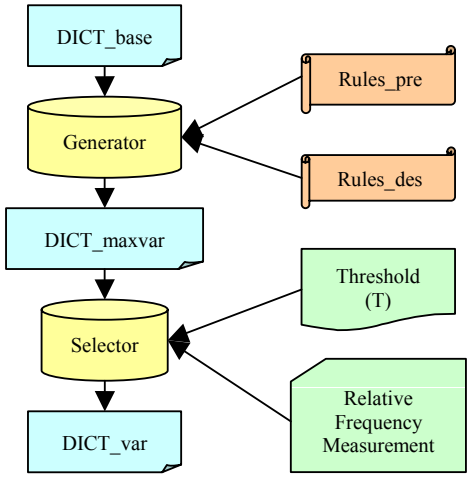
recogniser performs phone-level forced alignment and produces phone strings for each token. Then a simple script counts hits ((A) in Table 5) for each variant of the word 'behind' along with the accumulated hits of all the variants ((B) in Table 5). Later, the relative frequency measure $P(V_i|W=$ 'behind') for *I*'th variant of the word 'behind' is calculated as in the last column in Table 5.

Table 5. Example of relative frequency measure for the word 'behind'

Word	Variants	Hits (A)	Hits Total (B)	P(V*i*_behind) = (A)/(B)
behind	b ɪ h ay n d	20	30	0.667
	b ə h ay n d	5	30	0.167
	b ɪ h ay n	4	30	0.133
	b ə h ay n	1	30	0.333

This process can be schematised as Figure 1.

Figure 1. Schematised generation and selection procedure



5.4 Decision

Once the $P(V_i|W)$ is created as shown in Table 5., the intended final version of the variation dictionary can be formulated with its size depending upon the cut-off threshold T of which the range is $0<T<1$. When T=1, no variant except for the base-form pronunciation will survive and it results in a dictionary that is exactly the same as the DICT_base. When T=0, on the other hand, no selection will take place and the output is

DICT_maxvar. The decision criterion is summarised in (2).

(2) a. The base phone string is always kept.
b. The probability threshold (T) is determined.
c. Discard the i'th variant (V_i) if $P(V_i|W) < T$.

The implication of the criterion is straightforward. If T=0.15, for example, the last two variants [b ɪ h ay n], [b ə h ay n] of the word 'behind' is discarded since its $P(V_i|W)$ is less than T.

The first statement (2a) states that even though the value $P(V_i|W)$ of a base pronunciation (V_{base}) falls lower than the cutoff threshold T, it is not discarded. This is necessary for two reasons. First, although speakers do not usually produce the citation form of a word, special situations may take place in speech communication: *e.g.*, confirmation, focusing, re-iteration, etc. Second, and more crucially, when T is set to be relatively high, it can happen that a word in the dictionary has no variant at all.[4] This means that the word network essential for any recognition process will be constructed without the relevant word item, eventually resulting in a system failure.

The only remaining problem to be solved is discovering the appropriate value T which determines the size of the variant dictionary. Unfortunately, it is not possible to draw a fixed universal value T that can be used to constitute a variation dictionary. This, however, is quite reasonable considering the variability of ASR systems. Given that recognition performance is supposed to be affected by so many variables such as the size of vocabulary, number of speakers, target language, types and qualities of input speech, etc., the size of a dictionary also has to be calibrated whenever a different recognition environment is given. In brief, the inevitable flexibility does not undermine the method suggested in the current study.

However, we have one more task to complete. As the current speech recognition environment is already fixed, the cutoff threshold (T) should be determined anyway. While generation of variants, as described in section 5, is performed based mainly on human investigation of speech variability whether in terms of literature survey or of acoustic phonetic observation, the selection process needs to be done automatically as the ultimate goal of this research is to enhance the performance of speech recognition. For this purpose and, at the same time, for verifying whether the variation dictionary works, speech recognition tests have been performed, which is described in the following section.

[4] In the current system, for example, when the value 0.4 was assigned to T, some words began to disappear out of the dictionary, as none of its variants survived T.

6. Performance verification

As has been mentioned in the preceding section, the aim of speech recognition experiments is two folds: (1) fixing the cutoff threshold T, and (2) deciding if the pronunciation variation dictionary is useful.

6.1 Speech recogniser

A phone unit speech recogniser is designed and constructed. For training phone models, a well-known statistical approach, the Hidden Markov Model (HMM) technique is adopted. A three-emitting-state left-to-right continuous HMM is established for each of 40 phones in terms of acoustic feature parameter estimations. Features are extracted from signals for each 10-msec frame by using 25-msec Hamming window with pre-emphasis coefficient 0.97. A 39 dimensional vector is allocated for each frame, which is composed of 12 *Mel Frequency Cepstral Coefficients* (MFCC), energy, and their first and second derivatives. In order to reflect that a brief pause can be located between any two words, a short-pause model with a single state is separately created. During the course of the entire processing, a useful tool called the Hidden Markov Model Toolkit (HTK v3.0: Young *et al*., 1996) was used.

The recogniser is used in three ways: two ways are already mentioned in the beginning of the section 6. The other use is as an automatic phone aligner which is necessary for the selection process described in section 5.3.

6.2 Evaluation

A total of 736 unseen sentence tokens are used as test data for performance verification. After these test data pass through the recogniser, the word accuracy is calculated as in (3), which is widely accepted as standard evaluation score for ASR performance.

(3) Accuracy(%) = 100 x (N-(Substitution+Insertion+Deletion))/N.

One final thing to mention is that roles of the language (or grammar) model, which is usually regarded as one of the necessary modules for speech recognition, is minimized in the current experiment so that the performance of acoustic models, the module which is mainly relevant to the current study, can be better revealed. Therefore, the word accuracies given below should not be interpreted as an index of the speech recogniser's utmost quality.

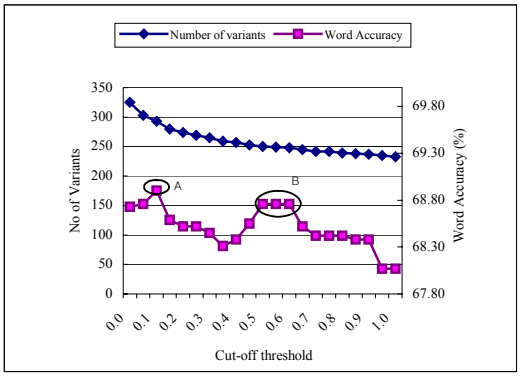
6.3 Results

Table 6 and Figure 2 summarise the performance of the speech recogniser with varied-size dictionaries.

Table 6. Word accuracy for each threshold

Cut-off Thresholds	Number of variants	Word Accuracy
0.00	325	68.73
0.05	303	68.76
0.10	293	68.90
0.15	280	68.59
0.20	274	68.52
0.25	269	68.52
0.30	265	68.45
0.35	259	68.31
0.40	257	68.38
0.45	253	68.55
0.50	250	68.76
0.55	249	68.76
0.60	248	68.76
0.65	245	68.52
0.70	242	68.42
0.75	242	68.42
0.80	239	68.42
0.85	238	68.38
0.90	237	68.38
0.95	235	68.07
1.00	233	68.07

Figure 2. Word accuracy as relative to varying thresholds: the highest accuracy (68.90%) is obtained at threshold (T= 0.1), while the lowest accuracy (68.07%) at T=1.0.



First of all, it is clearly shown, in terms of word accuracy, the variant dictionaries (T < 1.00) outperform the base-form dictionary (T=1.00). Especially when T=0.10, the best accuracy (68.90%) is obtained, which means the ultimate threshold of the current recognizer will be fixed to be 0.10. The upper line in Figure 2 representing values in the second column

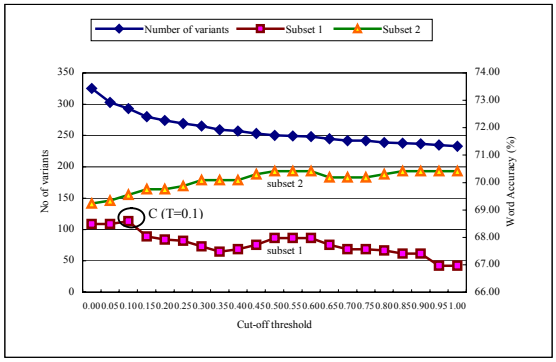
in Table 6 shows how the number of variants in the dictionary decreases, the smaller the cutoff value.

As for word accuracy, some fluctuations are shown but the general tendency is that the recogniser performs better when it works with a dictionary with more variants than less. It is not clear why the word accuracy line has two peak areas (A and B, in Figure 2). One possibility is that adding variants may have increased system confusability (Kessens *et al*. 2003: 521), deteriorating word accuracy at some specific area (*e.g.*, the valley between A and B). If this is the case, appropriate selection of variants becomes more important in that finding a balance between solving and introducing errors may influence the system in an unpredictable way.

One may point out the relatively small increase (0.83%), questioning the validity of the pronunciation variation dictionary in the current experiment. To further inspect this, a supplementary recognition test is contrived. The idea is that recognition performance is supposed to grow better when test data tokens are composed of sentences containing at least one word that has more than one variant in the dictionary. Therefore, the test data set used in the previous experiment is subdivided into two subsets. *Subset 1* is composed of 491 sentences in each of which pronunciation of one or more words are expanded as a variant either by RULES_pre or by RULES_des and specified in the dictionary. On the contrary, *Subset 2* includes 245 sentences without any word with pronunciation variants other than the base pronunciation. Consequently, it is anticipated that running recogniser with the *Subset 2* will result in deterioration of accuracy as the size of dictionary gets greater, *i.e.*, the cut-off threshold (T) becomes smaller.

Figure 3 summarises the recognition performance with those two test data subsets.

Figure 3. Recognition performance with two test data subsets: the best and worst accuracy for Subset 1 is 68.59% (T=0.1) and 66.96% (T=1.0), while those for *Subset 2* are 70.41% (T=1.0) and 69.23% (T=0.0), respectively.



The line representing accuracy on *Subset 1* shows the declining tendency, meaning that recognition performance worsens as the pronunciation dictionary shrinks. Note the best accuracy is still attained when T=0.1 as in the previous experiment with the whole test data. In addition, the distance between the best accuracy (68.59%, at T=0.1) at the peak point (C in Figure 3) and the lowest accuracy point (66.96%, at T=1.0) is greater than the distance between the best (68.90%, at T=0.1) and the worst (68.07%, at T=1.0) in the previous experiment (see Table 6 and Figure 2). The inclining trend of the line representing the performance on *Subset 2* also verifies the role of the pronunciation variation dictionary. When test utterances do not contain any word that produces variants, the dictionary with variants only increases confusability causing deterioration of recognition performance, which leads to the conclusion that not only generation but selection of variants is an indispensable process to provide an ASR system with the best-quality pronunciation dictionary.

The different performance of recognition with the three different data sets is also confirmed by statistical significance tests such as the single factor ANOVA ($F(2, 60)=239.09$, $p < 0.01$) followed by *Tukey*'s pairwise comparison .

7. Summary and suggestions

A comprehensive method of designing pronunciation variation has been introduced. The key procedures include (1) generating maximum variants, (2) adjusting the size of dictionary for efficiency, and (3) verifying performance through speech recognition tests.

Unlike previous research, both impressionistic top-down approaches and experimental bottom-up approaches are employed together to extract rules for generating maximum variants. Then, simple but effective relative frequency measure is used to systematically constrain variability and produce a task specific optimal-size dictionary. These methods appear fruitful since recognition performance has been enhanced, although slightly, thanks to the pronunciation dictionary.

The inevitable restriction of this study is the size and quality of data tokens. Judging from previous studies on pronunciation modelling, the suggested techniques in the current study are expected to work even better with a large amount of data with increased variability.

A systematically constructed pronunciation variation dictionary is useful in a variety of ways other than ASR. As pointed out by Tatham and Morton (2006: 261) lack of variability may cause a feeling of unnaturalness in synthesised speech. It implies that some rules that generate variations or some specific variants themselves can be directly used to make speech synthesiser produce more natural sounds.

The current study has paid attention to only segmental information in extracting rules for variability generation. The results of Bay and Ostendorf's

(2002) pilot experiments show that prosodic information can also be effectively used for pronunciation modelling and contribute in enhancing speech recognition performance, although not to the same extent as phonetic contexts. Considering that characteristics of English prosody are being constantly uncovered, and that methods of modelling and automatising are being developed, the task of using prosodic features for improved pronunciation modelling seems promising.

REFERENCES

BATES, REBECCA, and MARI OSTENDORF. 2001. Modelling pronunciation variation in conversational speech using prosody. *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, 17-22.

BINNENPOORTE, DIANA, CATIA CUCCHIARINI, LOU BOVES, and HELMER STRIK. 2005. Multiword expressions in spoken language: an exploratory study on pronunciation variation. *Computer Speech and Language* 19, 433-449.

BONAVENTURA, PATRIZIA, FILIPPO GALLOCCHIO, JEAN-FRANCOIS MARI, and GIORGIO MICCA. 1998. Speech recognition methods for non-native pronunciation variations, in Strik *et al*. 1998, 14-22.

BYRNE, WILLIAM, MICHAEL FINKE, SANJEEV KHUDANPUR, JOHN MCDONOUGH, HARRIET NOCK, MICHAEL RILEY, MURAT SARACLAR, CHARLES WOOTERS, and GEORGE ZAVALIAGKOS. 1997. Pronunciation Modelling for Conversational Speech Recognition: A Status Report from WS97, *Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, U.S.A.

GORONZY, SILKE, STEFAN RAPP, and RALF KOMPE. 2004. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communicatio*n 42.1, 109-123.

GREENBERG, STEVEN. 1997. The Switchboard Transcription Project, in Research Report #24, *Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA.

_____. 1998. Speaking in shorthand--a syllable-centric perspective for understanding pronunciation variation, in *Strik et al.* 1998, 47-56.

HEINE, HENRIK, GUNNAR EVERMANN, and UWE JOST. 1998. An HMM-based probabilistic lexicon, in Strik *et al*. 1998, 57-62.

KENSTOVICZ, MICHAEL. 1994. *Phonology in generative grammar*. Blackwell.

KESSENS, JUDITH M., CATIA CUCCHIARINI, and HELMER STRIK. 2003. A data-driven method for modeling pronunciation variation. *Speech Communication* 40.4, 517-534.

KREIDLER, CHARLES W. 1999. *The pronunciation of English: a course book in phonology*. Blackwell.

ROACH, Peter. 2004. *English Phonetics and phonology: a practical course*. Cambridge University Press.

SILVERMAN, DANIEL. 2006. *A critical introduction to phonology of sound, mind, and body*. Continuum.

STRIK, HELMER, JUDITH M. KESSENS, and MIRJAM WESTER (eds.) 1998. *Modeling pronunciation variation for automatic speech recognition*, Rolduc, The Nethelands. European Speech Communication Association, University of Nijemegen.

TATHAM, MARK, and KATHERINE MORTON. 2006. *Speech Production and perception*. Palgrave Macmillan.

WARD, WAYNE, HOLLY KRECH, XIUYANG YU, KEITH HEROLD, GEORGE FIGGS, AYAKO IKENO, DAN JURAFSKY, and WILLIAM BYRNE. 2002. Lexicon adaptation for LVCSR: speaker idiosyncracies, non-native speakers, and pronunciation choice. *ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation,* Estes Park, Colorado.

WESTER, MIRJAM. 2003. Pronunciation modeling for ASR-knowledge-based and data-derived methods. *Computer Speech and Language*, 17.1, 69-85.

YANG, QIAN, and JEAN-PIERRE MARTENS. 2000. On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR. *Proceedings of 11th ProRisc Workshop*, Veldhoven, The Netherlands, 589-593.

YOUNG, STEVE, J. JANSEN, DAVE OLLASON, and PHIL WOODLAND. 1996. *HTK Book.* Entropic.

Tae-Yeoub Jang
Department of English Linguistics
Hankuk University of Foreign Studies
270 Imun-dong, Dongdaemun-gu
Seoul 130-791, Korea
e-mail: tae@hufs.ac.kr

received: July 31, 2006
accepted: September 5, 2006