ACOUSTIC AND PRONUNCIATION MODEL ADAPTATION FOR CONTEXT-INDEPENDENT AND CONTEXT-DEPENDENT PRONUNCIATION VARIABILITY OF NON-NATIVE SPEECH

Yoo Rhee Oh, Mina Kim, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{yroh, kma58, hongkook}@gist.ac.kr

ABSTRACT

In this paper, we propose an acoustic and pronunciation model adaptation method for context-independent (CI) and context-dependent (CD) pronunciation variability to improve the performance of a non-native automatic speech recognition (ASR) system. The proposed adaptation method is performed in three steps. First, we perform phone recognition to obtain an n-best list of phoneme sequences and derive pronunciation variant rules by using a decision tree. Second, the pronunciation variant rules are decomposed into CI and CD pronunciation variation on the basis of context dependency. That is, some pronunciation variant rules that are dedicated to the specific phoneme sequences is classified into CI pronunciation variation, but others are classified into CD one. It is assumed here that CI and CD pronunciation variabilities are invoked by a different pronunciation space from the mother tongue of a non-native speaker and the coarticulation effects in a context, respectively. Third, the acoustic model adaptation is performed in a state-tying step for the CI pronunciation variability from an indirect data-driven method. In addition, the pronunciation model adaptation is completed by constructing a multiple pronunciation dictionary using the CD pronunciation variability. It is shown from the continuous Korean-English ASR experiments that the proposed method can reduce the average word error rate (WER) by 16.02% when compared with the baseline ASR system that is trained by native speech. Moreover, an ASR system using the proposed method provides average WER reductions of 8.95% and 3.67% when compared to the only acoustic model adaptation and the only pronunciation model adaptation, respectively.

Index Terms— Automatic speech recognition, non-native speech, pronunciation variability, acoustic model adaptation, pronunciation model adaptation.

1. INTRODUCTION

In spite of an increasing need for non-native automatic speech recognition (ASR), the recognition performance for a non-native ASR system degrades extremely when compared to a system that focuses solely on native speech [1]. There has been considerable research pertaining to non-native ASR reported, and they can be categorized into pronunciation modeling, acoustic modeling, and language modeling, and a hybrid modeling. First, pronunciation modeling applies the pronunciation variant rules to pronunciation models for non-native speech [2][3]. For example, several data-driven pronunciation modeling methods have been proposed by using a phoneme recognizer and a decision tree [4]-[7]. Second, acoustic modeling transforms and/or adapts the acoustic models to include the effect of non-native speech [2][8][9]. Third, language modeling handles the grammatical effects or speaking style of non-native speech [10]. Finally, a hybrid approach combines these three approaches for further improvement of ASR performance [11].

In this paper, we focus on a hybrid approach that combines an acoustic model and a pronunciation model adaptation method to improve the performance of a non-native ASR system. Especially, we analyze the pronunciation variability of non-native speech by using an indirect data-driven method, and adapt acoustic models and pronunciation models depending on the context-dependency of the pronunciation variability. To achieve our task, the pronunciation variability is first investigated with a non-native speech database (DB) in an indirect data-driven method based on a decision tree [12]. That is, we perform phone recognition to obtain an n-best phoneme sequences by using a development set, and derive pronunciation variant rules by using a decision tree, C4.5. Second, pronunciation variability is classified into either CI or CD pronunciation variability on the basis of context dependency. In other words, a pronunciation variant rule that occurs in the specific phoneme sequence (for example, the specific left and/or right phoneme) is classified as a CI pronunciation variant rule. Otherwise, the pronunciation variant rule is classified as a CD pronunciation variant rule. It is assumed here that CI pronunciation variability reflects a different pronunciation space between a mother tongue and a target language. Conversely, CD pronunciation variability covers coarticulation effects in a context. Third, an acoustic model adaptation [13] and a pronunciation model adaptation [12] are applied to reduce CI and CD pronunciation variability, respectively.

The organization of this paper is as follows. In Section 2, pronunciation variability is investigated in an indirect data-driven method, and is decomposed into CI and CD pronunciation variabilities. After that, we propose a hybrid acoustic and pronunciation model adaptation method for both CI and CD pronunciation variabilities in Section 3. In Section 4, the performance of a non-native ASR employing the proposed method is evaluated and compared with that using an acoustic model adaptation alone and a pronunciation model adaptation alone, respectively. Finally, we conclude and discuss our findings in Section 5.

2. DECOMPOSITION OF PRONUNCIATION VARIABILITY FOR NON-NATIVE SPEECH

2.1. Data-driven pronunciation variability analysis

To obtain pronunciation variability for non-native speech, an indirect data-driven method based on a decision tree is used, as shown in Fig. 1. First, each utterance in the development set of non-native speech is recognized by using a phoneme recognizer. The recognized n-best phoneme sequences are aligned using a dynamic programming algorithm with a reference phoneme sequence of the utterance, where the reference phoneme sequence is automatically obtained by using a CMU pronunciation dictionary [14] for the word of each utterance. From the alignment between the recognized phoneme sequence and the reference transcription, phoneme rule patterns are obtained as shown in Eq. (1):

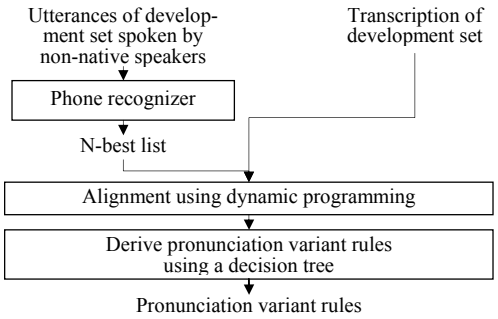

Figure 1: Procedure for obtaining pronunciation variability from non-native speech by using an indirect data-driven method based on a decision tree.

$$L_1 - L_2 - X + R_1 + R_2 \rightarrow Y \quad (1)$$

where X is a target phoneme that is to be mapped into Y, and the left and right phonemes in the reference transcription are L_1 and L_2, and R_1 and R_2, respectively. Second, pronunciation variation rules are derived from the variant phoneme patterns using a decision tree. In this paper, C4.5 is used for the decision tree, which is a software extension of the basic ID3 algorithm designed by Quinlan [15]. Their attributes are the two left phonemes, L_1 and L_2, and the two right phonemes, R_1 and R_2, of the affected phoneme X. The output class is the target phonemes, where one decision tree is constructed for each phoneme. After building the decision tree based on the established rule formulations and filtering the phoneme-to-phoneme mapping between the two transcriptions, we then construct pronunciation variant rule sets for each phoneme using options provided by C4.5. Eq. (2) shows a structure of a pronunciation variant rule set for a phoneme, $phoneme_{target}$:

Rule *rule_id*:
PrevPrev=p_1, *Prev*=p_2, *Next*=p_3, *NextNext*=p_4
→ class $phoneme_{variant}$
Default class: $phoneme_{default}$ (2)

where *rule_id* is an identifier of a pronunciation variant rule, and '*PrevPrev*=p_1, *Prev*=p_2, *Next*=p_3, and *NextNext*=p_4' are contexts in which the *rule_id* is applied. That is, the $phoneme_{target}$ is mapped into the $phoneme_{variant}$ if the context has the form of $p_1 - p_2 - phoneme_{target} + p_3 + p_4$. Otherwise, the $phoneme_{target}$ is mapped into the $phoneme_{default}$. A detailed explanation is described in [12].

2.2. Context-dependent and context-independent pronunciation variability

In this subsection, we describe how to classify a pronunciation variability into CI and CD pronunciation variability based on the result described in Section 2.1.

CD pronunciation variability is observed only in the limited and specific phoneme sequences such as allophones for a specific phoneme. For example, let us assume that a speaker utters 'this spring.' The pronunciation would be /DH IH S P R IH NG/ instead of /DH IH S S P R IH NG/ because the final phoneme /S/ of 'since' and the initial phoneme /S/ of 'spring' are adjacent. Except for these pronunciation variants due to coarticulation effects, the phoneme /S/ must be pronounced as /S/. Accordingly, the pronunciation variant rule sets for /S/ would be defined as follows:

Rule *S_rule1*:
PrevPrev=p_{11}, ***Prev=S***, *Next*=p_{31}, *NextNext*=p_{41}
→ class *sil*
Rule *S_rule2*:
PrevPrev=p_{12}, *Prev*=p_{22}, ***Next=S***, *NextNext*=p_{42}
→ class *sil*
⋮
Default class: */S/* (3)

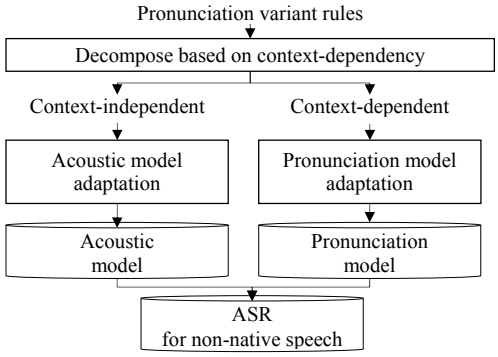

Figure 2: Schematic diagram of an ASR system for non-native speech constructed by combining an acoustic and pronunciation model adaptation method for CI and CD pronunciation variability, respectively.

Table 1. Comparison of CD and CI pronunciation variability.

	Pronunciation variability	
	Context-dependent (CD)	Context-independent (CI)
$phoneme_{target}$ vs. $phoneme_{default}$	Same	Different
Main reason	Coarticulation effect	Different Pronunciation space

As shown in Eq. (3), the phoneme /S/ is pronounced as /S/ except for the several exceptions that are identified in the pronunciation variant rule sets, *S_rule1*, *S_rule2* and etc. That is, the default class of the phoneme /S/ is /S/.

On the other hand, CI pronunciation variability is commonly observed in phonemes that do not exist in a mother tongue. For example, let us assume that a Korean utters 'five'. Since /F/ and /V/ do not exist in Korean, the Korean may mispronounce /F AY V/ as /P AY B/. This is because the Korean used to pronounce /F/ and /V/ as /P/ and /B/ that are similar phonemes in a Korean pronunciation space. In this case, the pronunciation variant rule sets for /F/ would be defined as follows:

Rule *F_rule1*:
PrevPrev=p_{11}, *Prev*=p_{21}, *Next*=p_{31}, *NextNext*=p_{41}
→ class $phoneme_{variant\_1}$
Rule *F_rule2*:
PrevPrev=p_{12}, *Prev*=p_{22}, *Next*=p_{32}, *NextNext*=p_{42}
→ class $phoneme_{variant\_2}$
⋮
Default class: */P/* (4)

As shown in Eq. (4), the most frequently pronounced phoneme for /F/ is /P/ for the Korean and therefore the default class for /F/ is mapped as /P/. In addition, the several exceptions for /F/ are mapped as the pronunciation variant rule sets, *F_rule1*, *F_rule2*, and etc.

As a result, the default class of the $phoneme_{target}$ is different from the $phoneme_{target}$ for CI pronunciation variability while the default class of the $phoneme_{target}$ is same as the $phoneme_{target}$ for CD pronunciation variability, which is described in the second row of Table 1.

3. COMBINATION OF ACOUSTIC AND PRONUNCIATION MODEL ADAPTATION FOR NON-NATIVE SPEECH

To improve the performance of a non-native ASR system, we propose a hybrid adaptation method by combining an acoustic model adaptation method for CI pronunciation variability and a pronunciation model adaptation method for CD pronunciation variability.

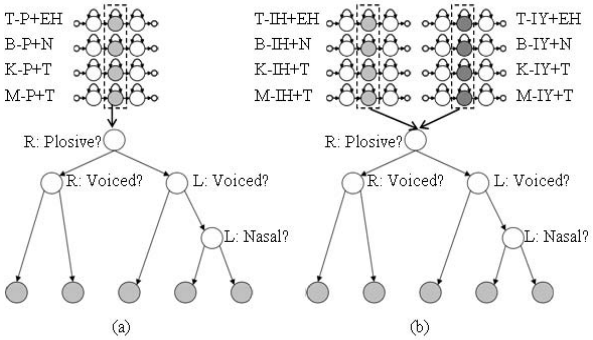

Figure 3**:** An illustration of the acoustic model adaptation in a state-tying step, presented in [13].

Fig. 2 outlines the procedure used in the proposed hybrid adaptation method. That is, we first obtain the CI and CD pronunciation variabilities from non-native speech. After that, we perform an acoustic model adaptation and a pronunciation model adaptation for CI pronunciation variability and CD pronunciation variability, respectively. The following two subsections briefly review the acoustic model adaptation method [13] and pronunciation model adaptation method [12], respectively, used for the proposed hybrid adaptation method.

3.1. Acoustic model adaptation

The acoustic model adaptation method presented in [13] is used to obtain CI pronunciation variability that is based on an indirect data-driven method. That is, the pronunciation variability from non-native speech is investigated in an indirect data-driven method and the acoustic model adaptation is performed in a state-tying step of acoustic modeling based on the pronunciation variability. Fig. 3 illustrates how the proposed acoustic model adaptation method works. Here we assume that a phoneme '*P*' has no variant phoneme but a phoneme '*IY*' is mapped into '*IH*', which is classified as a CI pronunciation variability. Fig. 3(a) shows a decision tree for the phone /P/ that has no pronunciation variants. In this case, the acoustic models for only /P/ are pooled on the root node of the decision tree. On the other hand, Fig. 3(b) shows a decision tree for the phoneme '*IY*' that has a pronunciation variant '*IH.*' That is, *−*−'*IY*'+*+*→ '*IH*' in Eq. (2), where * indicates any phoneme. From now on, it is denoted simply as */IY/→/IH/ for CI pronunciation variability*. In this case, the acoustic models of the triphones including both '*IY*' and '*IH*' as central phones are pooled on the root node of the decision tree. After clustering all the acoustic models using the decision tree, the models in each leaf node of the decision tree are tied with representative phonemes.

3.2. Pronunciation model adaptation

The pronunciation model adaptation method for CD pronunciation variability is based on the method proposed in [12], and is also performed as follows. The pronunciation variability from non-native speech is first investigated in an indirect data-driven method, which is the same in the acoustic model adaptation described in Section 3.1. Next, the pronunciation variant rules are derived from using a decision tree, as described in Section 2.1, allowing the CD pronunciation variability to be identified. Finally, pronunciation models are adapted from the derived pronunciation variant rules.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Baseline ASR system

A subset of the Wall Street Journal database [16], WSJ0, was used as the training set for the native-English ASR system. WSJ0 was a 5,000-word closed-loop task used to evaluate the performance of a large vocabulary continuous speech recognition system. The training set consisted of 7,138 utterances recorded by the Sennheiser close-talking microphone and several far-field microphones, where all the utterances were sampled at a rate of 16 kHz. As a recognition feature, we extracted 12 mel-frequency cepstral coefficients (MFCCs) with logarithmic energy for every 10 ms analysis frame, and concatenated their first and second derivatives to obtain a 39-dimensional feature vector. During training and testing, we applied cepstral mean normalization and energy normalization to the feature vectors.

The acoustic models were based on the 3-state left-to-right, context-dependent, 4-mixture, and cross-word triphone models, and trained using the HTK version 3.2 toolkit [17]. All the triphone models were expanded from 41 monophones, which also included a silence and a pause model, and the states of the triphone models are tied by employing a decision tree [18]. As a result, the acoustic models were composed of 8,360 triphones and 5,356 states, which is referred to as AM0 throughout this paper.

For non-native speech, we used a subset of the Korean-Spoken English Corpus (K-SEC) [19], which was composed of English pronunciations spoken by both Korean and native speakers. This database was divided into three parts: a development set, an evaluation set, and a test set. The development set was composed of utterances spoken by 1 Korean speaker, where the Korean speaker pronounced 1,103 isolated words. The evaluation set consisted of utterances spoken by 8 Koreans and 5 native speakers where each speaker utters 13 sentences, with an average number of about 7.6 words per sentence. The test set consisted of utterances spoken by 49 Koreans and 7 native speakers where each speaker spoke 14 sentences, with an average of 10.4 words per sentence.
In order to explore the behavior of the acoustic and pronunciation models due to differences between the target language and the mother tongue, we used only the texts from the evaluation set and the test set to construct a language model that is a backed-off bigram. The baseline pronunciation of each word was built from the CMU pronunciation dictionary [14] and the missing words in the CMU dictionary were transcribed manually. This baseline pronunciation model is referred to as PM0.

4.2. Acoustic model and pronunciation model adaptation based on context-independency

To derive pronunciation variability from the development set, we first performed phone recognition for each utterance of the development set by using the baseline ASR system to obtain a 20-best list. Second, we aligned each recognized phoneme sequence and the transcribed phoneme sequence. Third, the pronunciation variant rules were derived using the decision tree toolkit C4.5 [15]. Next, we decomposed the pronunciation variability based on the context-dependency. Finally, we identified the CI pronunciation variability as /G/→ /sil/, /L/→ /R/, /TH/→/DH/, /ZH/→/Z/. It is noted that /R/, /TH/, /DH/, /ZH/ cannot be pronounced in Korean, which proves that the CI pronunciation variability reflects the pronunciation structure to some degree.

Table 2 shows the average word error rates (WERs) of the baseline ASR system (*AM0* + *PM0*), an ASR system with the adapted acoustic models and the baseline pronunciation model (*adapted-AM*+*PM0*), an ASR system with the baseline acoustic model and the adapted pronunciation model (AM0+*adapted-PM*), and an ASR system with both the adapted acoustic models and the adapted pronunciation model (*adapted-AM* + *adapted-PM*) for the evaluation set. It was shown from the first row of Table 2 that the average WER of the system using *AM0* + *PM0* was 3.69%.

For the CI pronunciation variability, we applied the acoustic model adaptation method. As can be seen in the second row of Table 1, the average WER of the system using *adapted-AM* and *PM0* decreased to 2.73%. It was also interesting that the average WERs for native speech and non-native speech were all reduced. To investigate the effect of the acoustic model adaptation for the CI pronunciation variability, we repeated the procedure for deriv-

Table 2. Comparison of the average WERs (%) of the baseline ASR system and ASR systems with a different combination of adapted models for the evaluation set.

ASR system \ Speaker	Non-native	Native	Avg.	Relative WER Reduction (%)
Baseline (*AM0+PM0*)	6.47	0.91	3.69	-
adapted-AM+PM0	4.93	0.55	2.73	26.1
AM0+adapted-PM	4.66	0.73	2.69	27.1
adapted-AM adapted-PM	4.93	0.55	2.73	26.1

Table 3. Comparison of the average WERs (%) of the baseline ASR system and ASR systems with a different combination of adapted models for the test set.

ASR system \ Speaker	Non-native	Native	Avg.	Relative WER Reduction (%)
Baseline (*AM0+PM0*)	19.92	0.68	10.30	-
adapted-AM+PM0	18.12	0.88	9.50	7.8
AM0+adapted-PM	17.28	0.68	8.98	12.8
adapted-AM adapted-PM	16.51	0.78	8.65	16.0

ing the pronunciation variability and classified them into CI pronunciation variability and CD pronunciation variability. As a result, CI pronunciation variability such as /G/→/T/, /UH/→/AH/ was newly obtained.

For the CD pronunciation variability, the pronunciation model adaptation method was applied for the pronunciation dictionary. It was shown in the third row of Table 2 that the average WER was further reduced to 2.69%, which corresponded to a relative WER reduction of 27.1% when compared to the baseline ASR.

As proposed in Section 3, we combined the acoustic model adaptation and the pronunciation adaptation for the CI pronunciation variability and the CD pronunciation variability. However, the performance for the evaluation set was similar to that using only the acoustic model adaptation.

Table 3 shows the average WERs of the ASR systems using different combination of acoustic models and pronunciation models such as *AM0+PM0*, *adapted-AM+PM0*, AM0+*adapted-PM*, and *adapted-AM+adapted-PM* for the test set. It could be seen from Table 3 that the ASR systems using *adapted-AM+PM0* and *AM0+adapted-PM* could reduce the average WER by 7.8% and 12.8%, respectively, as compared to the baseline system. Moreover, the relative WER reduction of the ASR systems employing the *adapted-AM+adapted-PM* was 16.0%, compared to AM0+PM0.

It could be concluded here that the combination of the acoustic and pronunciation model adaptation methods for CI and CD pronunciation variability could improve the ASR performance further, when compared to the adapted acoustic model adaptation only and the pronunciation model adaptation only.

5. CONCLUSION

In this paper, we proposed a hybrid acoustic and pronunciation model adaptation method for context-independent (CI) and context-dependent (CD) pronunciation variability to improve the performance of a non-native ASR system. The decomposition of pronunciation variability into CI and CD pronunciation variabilities was based on the assumption that CI and CD pronunciation variabilities could be invoked by a different pronunciation space from the mother tongue of a non-native speaker and coarticulation effects in a context, respectively. The proposed acoustic and pronunciation model adaptation method was performed in three steps: the analysis step of non-native speech, the decomposition step into CI and CD pronunciation variability, and the adaptation step using a hybrid acoustic and pronunciation model adaptation method based on the decomposition results. It was shown from the continuous Korean-English ASR experiments that the proposed method could reduce the average WER by 16.02% when compared with the baseline ASR system that was trained by native speech. Moreover, an ASR system using the proposed method reduced the average WERs by 8.95% and 3.67% when compared with the WERs using the only acoustic model adaptation and the only pronunciation model adaptation, respectively.

6. ACKNOWLEDGEMENTS

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245) and by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2008-(C1090-0801-0017)).

7. REFERENCES

[1] D. V. Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," *Speech Communication*, vol. 35, no. 1, pp. 71-79, Aug. 2001.

[2] R. Gruhn, K. Markov, and S. Nakamura, "A statistical lexicon for non-native speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, pp. 1497-1500, Oct. 2004.

[3] A. Raux, "Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, pp. 616-616, Oct. 2004.

[4] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, vol. 29, nos. 2-4, pp. 225-246, Nov. 1999.

[5] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," in *Proc. Eurospeech*, Budapest, Hungary, pp. 463-466, Sept. 1999.

[6] I. Amdal, F. Korkmazasky, and A. C. Suredan, "Data-driven pronunciation modelling for non-native speakers using association strength between phones," in *Proc. ASRU*, Kyoto, Japan, vol. 1, pp. 85-90, Aug. 2000.

[7] S. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, vol. 42, no. 1, pp. 109-123, Sept. 2003.

[8] S. Steidl, G. Stemmer, C. Hacker, and E. Noth, "Adaptation in the pronunciation space for non-native speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, pp. 2901-2904, Oct. 2004.

[9] J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging." in *Proc. InSTIL/ICALL Symposium on Computer-Assisted Language Learning*, Venice, Italy, pp. 213–216, June 2004.

[10] J. Bellegarda, "An overview of statistical language model adaptation," in *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, pp. 165–174, Aug. 2001.

[11] G. Bouselmi, and I. Illina, "Combined acoustic and pronunciation modelling for non-native speech recognition," in *Proc. Interspeech*, Antwerp, Belgium, pp. 1449-1452, Aug. 2007.

[12] M. Kim, Y. R. Oh, and H. K. Kim, "Non-native pronunciation variation modeling using an indirect data driven method," in *Proc. ASRU*, Kyoto, Japan, pp. 231-236, Dec. 2007.

[13] Y. R. Oh, J. S. Yoon, and H. K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," *Speech Communication*, vol. 49, no. 1, pp. 59-70, Jan. 2007.

[14] H. Weide, *The CMU Pronunciation Dictionary, release 0.6*, Carnegie Mellon University, 1998.

[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, California, Morgan Kaufmann Publishers, 1993.

[16] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. DARPA Speech and Language Workshop*, Arden House, NY, pp. 357-362, Feb. 1992.

[17] S. Young, *et al*, *The HTK Book (for HTK Version 3.2)*, Microsoft Corporation, Cambridge University Engineering Department, Dec. 2002.

[18] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 307-312, Mar. 1994.

[19] S.-C. Rhee, S.-H. Lee, S.-K. Kang, and Y.-J. Lee, "Design and construction of Korean-spoken English corpus (K-SEC)," in *Proc. ICSLP*, Jeju Island, Korea, pp. 2769-2772, Oct. 2004.