INTERSPEECH 2009 BRIGHTON

Improving phone recognition performance via phonetically-motivated units

Hyejin Hong, Minhwa Chung

Department of Linguistics, Seoul National University, Seoul, Korea

souble1@snu.ac.kr, mchung@snu.ac.kr

Abstract

This paper examines how phonetically-motivated units affect the performance of phone recognition systems. Focusing on the realization of /h/, which is one of the most frequently error-making phones in Korean phone recognition, three different phone sets are designed by considering optional phonetic constraints which show complementary distributions. Experimental results show that one of the proposed sets, the *h-deletion* set improves phone recognition performance compared to the baseline phone recognizer. It is noteworthy that this set needs no additional phonetic unit, which means that no more HMM is necessary to be modeled, accordingly it has the advantage in terms of model size. Besides, it obtains competent performance compared to the baseline system in terms of word recognition as well. Thus, this phonetically-motivated approach dealing with improvement of phone recognition performance is expected to be used in embedded solutions which require fast and light recognition process.

Index Terms: phone recognition, phonetic constraint, phone set

1. Introduction

Current large vocabulary speech recognition systems have made dramatic advances, however, still their performance degrades further below that of humans. Several methods have been suggested to introduce all linguistic information available into ASR systems. As one of these methods dealing with this, layered ASR systems which split up the search engine into two separate layers, acoustic decoding and lexical decoding, have been proposed [1][2][3]. This method can be used effectively in embedded solutions since it enables faster and lighter recognition procedure. In the case of the layered systems, the first layer, phone recognizer, is one of the most crucial factors affecting systems' overall performance. To improve phone recognition performance, examining frequently occurring errors which lead to poor performance and dealing with these errors are strongly required.

In this paper, as an exploratory study, we shall show considering phonetically-motivated units can be successful to improve phone recognition performance. First, we build a baseline phone recognizer to detect phones which lead to poor performance and need to be improved by analyzing confusion matrix. One of Korean phones showing the poorest performance is selected and its error-occurring contexts are analyzed based on phonetic knowledge. Then, phonetic units are designed by considering three Korean phonetic constraints, which complementarily operate in the same context, but are not obligatorily applied. The effects of the proposed phonetically-motivated units are investigated by performing phone recognition experiments using a large vocabulary PBS (Phonetically Balanced Sentences) continuous speech corpus.

The remaining part of this paper is organized as follows. Section 2 describes our baseline phone recognizer and its performance as well as commonly occurring errors. Three phone sets are proposed by introducing different units based on phonetic constraints in Section 3. Experimental results using the phonetic units are presented in Section 4. Section 5 concludes the paper.

2. Baseline phone recognizer

2.1. Settings

Our baseline phone recognizer is constructed in a way as follows. The baseline phone set, which consists of 38 units including 19 consonantal units, 17 vowel units and 2 units for silence and short pause, is shown in Table 1. Each phonetic unit is modeled as a three state, left-to-right HMM (Hidden Markov Model) using HTK (HMM Toolkit) version 3.4 [4]. The number of Gaussian mixtures used for the output probability density of an HMM is increased up to 64. Pronunciation dictionary and transcriptions are automatically generated from G2P (Grapheme-to-Phoneme) converter [5], and edited by a labeler with Korean phonetic knowledge, if needed. For the main purpose which investigates the effects of phonetic units, the language model is refined with less effort, not as in word recognition.

Table 1. *Baseline phone set.*

Consonants			Vowels		
Korean orthography	IPA symbol	PLU symbol	Korean orthography	IPA symbol	PLU symbol
ㅂ	/p/	P	ㅏ	/ɑ/	AA
ㅃ	/p͈/	PP	ㅓ	/ʌ/	AX
ㅍ	/pʰ/	PH	ㅗ	/o/	OW
ㄷ	/t/	T	ㅜ	/u/	UW
ㄸ	/t͈/	TT	ㅡ	/ɯ/	WW
ㅌ	/tʰ/	TH	ㅣ	/i/	IY
ㄱ	/k/	K	ㅐ/ㅔ	/e/	EY
ㄲ	/k͈/	KK	ㅑ	/jɑ/	JA
ㅋ	/kʰ/	KH	ㅕ	/jʌ/	JX
ㅅ	/s/	S	ㅛ	/jo/	JO
ㅆ	/s͈/	SS	ㅠ	/ju/	JU
ㅎ	/h/	H	ㅒ/ㅖ	/je/	JE
ㅈ	/tɕ/	Z	ㅘ	/wɑ/	WA
ㅉ	/tɕ͈/	ZZ	ㅝ	/wʌ/	WX
ㅊ	/tɕʰ/	CH	ㅚ/ㅙ/ㅞ/	/we/	WE
ㄹ	/l/	L	ㅟ	/wi/	UI
ㅁ	/m/	M	ㅢ	/ɯi/	WI
ㄴ	/n/	N			
ㅇ	/ŋ/	NX			
Others					
silence		sil	short pause		Q

For all experiments presented in this paper, the phonetically balanced continuous speech corpus which consists of 45,000 sentences read by 450 native Korean

10.21437/Interspeech.2009-504

speakers (100 sentences per speaker) is used. The training data consists of 43,000 sentences and a speaker-disjoint set of 2,000 sentences is set aside for testing. The statistics of the corpus used in our experiments are shown in Table 2.

Table 2. *Statistics of the corpus used in experiments.*

	Training	Test
Sentence	43,000	2,000
Word	843,628	37,749
Vocabulary	35,868	6,216
Phone	2,826,699	125,487
Words per sentence	19.6	18.9
Phones per word	3.4	3.3

The speech data are parameterized into 12th order MFCCs and log energy, and their first and the second deltas are incorporated.

2.2. Baseline performance

2.2.1. Overall performance

The overall performance of the baseline recognizer is presented in Table 3.

Table 3. *Overall baseline performance.*

Number of Gaussian mixtures	PER (%)
1	71.52
2	64.76
4	55.52
8	48.48
16	41.96
32	37.91
64	34.23

2.2.2. Individual phone recognition performance

To investigate which phone degrades the recognition performance the most, confusion matrix analysis is performed. The performance of individual phone recognition is assessed using the percentage correct (%c) in the row of the confusion matrix, which means how many times a phone instance was correctly recognized [4]. We inspect all errors considering phone frequencies besides the performance, since one's impact on the overall performance may not be considerable if it does not occur frequently. Considering the number of insertions and that of deletions, PER increase of each phone is calculated. Only five fundamental phones which considerably reduce the performance, occupying over 2.5% of the whole phones and showing the poorest performance, are shown in Table 4.

Table 4. *The most common error-making phones (32 Gaussian mixtures).*

Phone	%c	Frequency (% of total phones)	PER increase (% of total errors)
EY	54.0	5,398 (4.30)	2.30 (6.07)
H	63.8	4,100 (3.27)	2.59 (6.83)
Z	64.6	4,703 (3.75)	1.40 (3.70)
K	66.1	8,835 (7.04)	3.14 (8.28)
P	66.2	3,310 (2.64)	1.73 (4.57)

These five phones were also indicated as the main error-making phones in [6]. These phones increase the overall absolute PER by 11.16%, which means 29.44% of total phone recognition errors. This implies that some specific phones occupying a large portion of total phones failed to be correctly recognized.

Then why these certain phones show poor performance? Two possibilities seem plausible. First, certain phones keep quite small acoustic or phonetic distance from other phones, so that they fail to be correctly recognized, but are substituted. In the case of 'EY', 41.17% of errors are occurred as substitutions into diphthongs 'WE' and 'JE', which are comprised of 'EY' and a semi-vowel, [j] and [w], respectively. Besides 'EY', other monophthongs often misrecognized as their corresponding diphthongs. It might be because that the duration of each semi-vowel is relatively short compared to other segments in general, accordingly, distinctive phonetic characteristics between diphthongs and their matching monophthongs are intricate to capture. Finally, it might lead to many substitution errors between monophthongs and diphthongs. Same situations are found in 'Z', 'K', 'P', as well. Three units for unvoiced lax obstruents, 'Z', 'K', 'P' are likely to misrecognized as other tense 'ZZ', 'KK', 'PP' and aspirated consonants 'CH', 'KH', 'PH', which keep close phonetic distance from their corresponding lax consonants. In the case of these lax consonants, some proportions of these confusions seem to be removed by considering allophonic realizations. Our previous work [7], which introduced Korean consonantal allophonic units, voiced units, confirmed that these allophonic units are effective to improve the performance.

Glottal fricative /h/, however, seems to be in different case from the previously discussed phones. Its phone recognition performance was not considerably improved but still poor when its voiced unit was considered, that is, it needs to receive more attention than other obstruents. The second possibility arises at this point. Some phones may be hard to be modeled for their characteristics per se, having tendency to be weakened or diversity of phonetic realizations according to their contexts. Actually, /h/ is likely to be weakened especially in spontaneous speech. Moreover, its phonetic realizations vary in different contexts, which makes hard to capture discriminative clue to recognize it correctly. Thus, more phonetic considerations on these phones are necessary for improving the performance of these phones including /h/.

Among the five common error-making phones, we shall focus on /h/ for the reason discussed above. /k/ increases the overall PER the most (3.14%), however, /h/ increases absolute PER by 2.59%, which is still high. Accordingly, we give priority to /h/ and scrutinize its error-occurring contexts in the next.

2.2.3. Frequently error-occurring contexts of /h/

We analyze all errors in /h/ recognition. The deletion of /h/ is the most significant type of errors, which means that 59.14% of /h/ recognition errors are deletions, and 21.28% of all deletion errors are from /h/. This is comparable to [6].

It is worthy to note that /h/ is likely to be misrecognized especially between sonorant sounds including vowel, semi-vowel, nasal and lateral. This was indicated in [6], as well. Figure 1 shows an example of misrecognized /h/ between sonorants.

Transcription	K	OW	**NX**	**H**	**AA**	K
Recognition	K	OW	**NX**		**WA**	K

Figure 1: *An example of misrecognized /h/ between sonorants.*

Note that not only 'H' is deleted, but following 'AA' is also substituted. As shown, misrecognition of /h/ affects its adjacent phones as well.

The realizations of Korean /h/ between sonorants, however, have been rarely investigated in terms of both phonetics and speech technology, except for [8], a phonetics-based quantitative study. It is reported that 69.03% of /h/ is completely deleted between sonorants in average, and 12.67% is realized as voiced sound. 18.30% of /h/ makes following vowel devoiced [8]. Thus, /h/ has three different realizations, showing complementary distributions, between sonorant sounds. Focusing on the most frequently error-occurring contexts of /h/, between sonorants, phonetic units are proposed to improve the overall performance and /h/ recognition performance in the next section.

3. Phonetic unit refinements

Based on the realization of /h/ as described above, alternate phone sets are designed by adding phonetically-motivated units to the baseline phone set presented in Table 1. This paper focuses on the following three phonetic constraints of the realizations of /h/, which are applied optionally, not obligatorily:

- *h-voicing*: /h/ is realized as voiced [ɦ] between sonorants.
- *Vowel devoicing*: /h/-following vowel is realized as [V̥], when /h/ is between sonorants.
- *h-deletion*: /h/ is deleted between sonorants.

Considering these optional phonetic constraints, the phonetic units which are to be added to the baseline phone set is given in Table 5.

Table 5. *Added units derived by phonetic constraints on the realizations of /h/.*

Phone set	Number of units	Applied constraint		
		Korean orthography	IPA symbol	PLU symbol
h-voicing	39 (baseline +1)	ㅎ	[ɦ]	H1
vowel devoicing	56 (baseline +17)	ㅏ	[ɑ̥]	AA0
		ㅓ	[ʌ̥]	AX0
		ㅗ	[o̥]	OW0
		ㅜ	[u̥]	UW0
		ㅡ	[ɯ̥]	WW0
		ㅣ	[i̥]	IY0
		ㅐ/ㅔ	[e̥]	EY0
		ㅑ	[jɑ]	JA0
		ㅕ	[jʌ]	JX0
		ㅛ	[jo]	JO0
		ㅠ	[ju]	JU0
		ㅒ/ㅖ	[je]	JE0
		ㅘ	[wɑ̥]	WA0
		ㅝ	[wʌ̥]	WX0
		ㅚ/ㅙ/ㅞ/	[we̥]	WE0
		ㅟ	[wi̥]	UI0
		ㅢ	[ɯi̥]	WI0
h-deletion	38 (baseline +0)	-	-	-

These refined units are used to model the realization of /h/ and its following vowel. Figure 2 shows an example of transcriptions using the refined phone sets.

Korean orthography	공학 'engineering'					
Baseline	K	OW	NX	H	AA	K
h-voicing	K	OW	NX	**H1**	AA	K
vowel devoicing	K	OW	NX	H	**AA0**	K
h-deletion	K	OW	NX	-	AA	K

Figure 2: *An example of transcriptions using baseline and refined phone sets.*

Note that the phonetic unit for /h/ and its following vowel are different in four phone set. The *h-voicing* set has the unit for voiced /h/, 'H1', while the *vowel devoicing* set has the unit for devoiced vowel unit, 'AA0'. In the case of the *h-deletion* set, the unit for /h/ is removed from the transcription to represent deleted /h/ between sonorants.

4. Experimental results

With three proposed phone sets given in section 3, phone recognition experiments are performed. All experimental setups are unchanged as in the previous baseline experiments described in section 2, except for the phone set and its corresponding pronunciation dictionary.

Table 6 shows the experimental results of phone recognition. Compared to the baseline, the performance is degraded in two sets, the *h-voicing* set and the *vowel devoicing* set, while the *h-deletion* set obtains an absolute reduction in PER by 0.75%, which means 1.37% of relative improvement. One possible factor leading the *vowel devoicing* set to remarkable degradation is that it has 17 more units than the baseline set.

Table 6. *Experimental results using the proposed phone sets (32 Gaussian mixtures).*

Phone set	PER (%)	Absolute reduction in PER (%)	Relative reduction in PER (%)
Baseline	37.91	-	-
h-voicing	38.77	-0.86	-1.95
vowel devoicing	39.65	-1.74	-3.35
h-deletion	37.16	+0.75	+1.37

To examine the effects of each phone set more in detail, the number of errors is given in Table 7. Compared to the baseline set, the *h-voicing* set has more insertion errors while deletion errors decrease as in the *vowel devoicing* set, which shows the worst performance. Substitutions neither increase nor decrease significantly in both sets. In the case of the *h-deletion* set, however, deletion and substitution errors remarkably decrease, while insertions increase to a certain degree. The increase of insertion errors may be on account of realization of /h/, which was modeled as deleted one in the system. Actually, 833 more instances of /h/ are inserted when the *h-deletion* set is used compared to the baseline set. Though, as shown in Table 6, the *h-deletion* set obtains the best performance, reducing the absolute PER by 0.75%.

Table 7. *Number of insertion, substitution and deletion errors.*

Phone set	Insertions	Deletion	Substitution
Baseline	10,492	6,625	30,457
h-voicing	11,934	6,170	30,543
vowel devoicing	13,987	5,389	30,376
h-deletion	10,889	4,964	29,760

Now, we are in a position to inspect the proposed set by comparing the performance in terms of the overall performance, /h/ recognition performance and the overall performance excluding /h/.

Table 8. *Performance comparison of proposed sets (32 Gaussian mixtures).*

	PER (%)	PER excluding /h/ (%)	%c of /h/
Baseline	37.91	36.52	63.80
h-voicing	38.77	36.40	74.90
vowel devoicing	39.65	40.30	61.72
h-deletion	**37.16**	**35.90**	**85.30**

As shown in Table 8, the *h-voicing* set obtains improved performance excluding /h/ and that of /h/. The overall performance, however, degrades. The improvement in the percentage correct (%) of /h/ and PER excluding /h/, but the reduction in overall performance means that added phonetic units for voiced /h/ makes more hits of /h/, however, it leads to more insertions of /h/. The number of /h/ insertions drastically increase in the *h-voicing set*, from 863 insertions in the baseline set to 2,622. In the case of the *vowel devoicing* set, it shows no improvements. The number of hits increases, however, since the number of insertions enormously increases, the overall performance is worsened. On the contrary, the *h-deletion* set seems to be the best set among the proposed sets; it shows the best performance in terms of PER excluding /h/ and the performance of /h/ recognition as well as the overall performance. It obtains 0.62% absolute reduction in PER excluding /h/. The performance of /h/ recognition measured by the percentage correct (%c) is improved by 21.5% in the *h-deletion* set compared to the baseline.

It is noteworthy that no additional phonetic unit is added to the *h-deletion* set, which means that no more HMM is required to be modeled compared to the baseline recognizer.

We performed additional word recognition experiments with the same corpus used in phone recognition experiments to investigate the effectiveness of the set in other recognition domains. Although phone recognition performance closely correlates with word recognition performance, it may not exactly match. Thus, we perform word recognition experiments to show that our phonetically-motivated units provide the reliable source of phonetic information which is used to decode a given data.

Our experimental results of word recognition are given in Table 9. When 4 Gaussian mixtures are used, the *h-deletion* set shows 26.96% of WER, while the baseline set obtains WER 27.13%. Increasing Gaussian mixtures to 16, the baseline set outperforms the *h-deletion* set by 0.02%. Accordingly, in terms of word recognition task as well, this set shows competent performance compared to the baseline, as given in Table 9.

Table 9. *WER (%) of the baseline and the h-deletion set.*

	4 Gaussian mixtures	16 Gaussian mixtures
Baseline	27.13	20.59
h-deletion	26.96	20.61

5. Conclusions

This paper examines the effects of phonetically-motivated units on the performance of Korean phone recognition systems, especially for embedded-solutions. One of the most frequently misrecognized phones, glottal fricative /h/ is focused, especially when it is realized between sonorants. By considering three optional phonetic constraints on the realizations of /h/, which are not obligatory, three different phone sets are designed. Experimental results show that the *h-deletion* set, which do not require any additional phonetic unit to the baseline phone set and do not increases model size, outperforms compared to the baseline phone recognizer using phoneme-based phone set. In terms of word recognition task, it shows competent performance compared to the baseline system. Thus, this phonetically-motivated approach seems to be successful to improve phone recognition performance, and it would be used in any phone recognizer, especially requiring fast and light recognition process such as in embedded solutions.

In our future research, phonetic characteristics of conversational speech and their effects will be investigated with performing experiments in the real embedded solution conditions.

6. Acknowledgements

This paper was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy (MKE).

7. References

[1] Demuynck, K., Laureys, T., Van Compernolle, D. and Van hamme, H., "Flavor: a flexible architecture for LVCSR", in *Proc. EUROSPEECH 2003*, 1973-1976, 2003.

[2] Casar, M. and Fonollosa, J.A.R., "Analysis of HMM temporal evolution for automatic speech recognition and utterance verification", in *Proc. INTERSPEECH 2006*, 613-616, 2006.

[3] Kim, S.H., Hwang, K., Jeon, H., Jeong, H. and Park, J., "Development of embedded fast/light phoneme recognizer for distributed speech recognition", in *Proc. Korea Information Processing Society 2007 Spring Conference*, 395-396, 2007.

[4] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.

[5] Lee, K.N. and Chung, M., "Morpheme-based modeling of pronunciation variation for large vocabulary continuous speech recognition in Korean," *IEICE Transactions on Information and Systems*, 90(7), 1063-1072, 2007.

[6] Kim, S.H., Jeon, H.B. and Park, J., "Recognition error analysis of phone recognizer", in *Proc. The 23th Korean Speech Communication and Signal Processing Workshop*, 2006.

[7] Hong, H., Kim, S. and Chung, M., "Effects of allophones on the performance of Korean speech recognition", in *Proc. INTERSPEECH 2008*, 2410-2413, 2008.

[8] Cha, J., Jung, M. and Shin, J., "A study on the realization of /h/ between sonorant sounds", in *Proc. Korean Society of Phonetic Sciences and Speech Technology 2003 Spring Conference*, 48-51, 2003.