Improving transcription agreement of non-native English speech corpus transcribed by non-natives

Hyuksu Ryu[1],Kyuwhan Lee[2], Sunhee Kim[2], Minhwa Chung[1]

[1] Department of Linguistics, Seoul National University, Seoul, Korea
[2] Center for Humanities and Information, Seoul National University, Seoul, Korea

{oster01, whannylee, sunhkim, mchung}@snu.ac.kr

Abstract

This paper proposes an economical and effective phonetic transcription method for dealing with a large amount of non-native English speech corpus. The method provides a consistent transcription agreement, although the corpus is transcribed by non-natives. To minimize the possibility of confusion in transcription process, forced aligned phone sequences and a set of possible mispronunciation candidate phones that Korean L2 learners are expected to make are given to the Korean transcribers for reference. The proposed method is evaluated by measuring the transcription agreement using Fleiss' kappa as well as percentage agreement. Furthermore, the transcription consistency is analyzed by comparing it to that performed on the English corpus transcribed by native speakers. As a result, a transcription agreement of 0.869 is achieved, while the Buckeye corpus transcribed by natives shows a transcription agreement of 0.803.

Index Terms: transcription method, transcription agreement, non-native transcriber, forced alignment

1. Introduction

Non-native speech recognition is necessary for developing a Computer Assisted Language Learning (CALL) or Computer Assisted Pronunciation Teaching (CAPT) system for L2 learning. Usually the performance of non-native speech recognition is lower than that of native speech recognition. For effective error detection and feedback in language education systems, we need to improve the performance of non-native speech recognition. Precise phonetic transcription of non-native speech corpus is necessary for enhancing the quality of acoustic and pronunciation models used for speech recognition.

In phonetic transcription, error patterns of non-natives should be indicated. For example, an English learner whose L1 is Korean frequently shows errors such as vowel epenthesis, monophthongization, nasalization, lateralization, and so on [1][2]. When a learner speaks 'cake', [k eɪ k], s/he usually shows pronunciation errors as shown below. These errors should be transcribed and used for improving non-native speech recognition performance.

'cake' [k eɪ k] a. [k e i k] (monophthongization)
b. [k eɪ k ɯ] (vowel epenthesis)
c. [k e i k ɯ] (combination of a & b)

Transcription of non-native speech corpus can be done by native speakers of target language or by non-natives. In both cases, transcribers should have phonetic knowledge about both L1 and L2 to show error patterns of non-native speakers. Bonaventura et al. [3] suggested a hybrid manual transcription method of non-native speech corpus, where both native and non-native transcribers participate in transcription. At first, a native trained phonetician marks mistakes of non-native speech. Afterwards, a non-native phonetician with good knowledge of the target language will listen again and confirm the marks. However, it is difficult to find native L2 transcribers, in L1 spoken country, who have also phonetic knowledge of L1. In this sense, the hybrid manual transcription method in [3] is an expensive procedure, especially when we need to transcribe a large amount of non-native speech corpus used for developing a non-native speech recognition system.

Transcription by natives has a strong point in that native transcribers can judge pronunciation of non-native speech intuitively. They can easily tell whether non-native speakers pronounced correctly or not. However, they have difficulties in finding exact L1-dependent error patterns. In contrast, transcription by non-natives has advantage that non-native transcribers catch L1-dependent error patterns better than native transcribers. However, transcription by non-natives has a problem that agreement of transcription is degraded [4].

There are many studies about transcription agreement itself [5][6][7]. Cucchiarini [5] used Dutch native speech corpus transcribed by native to propose feature-based distance matrices for calculating transcription agreement. Hacker [6] studied an automatic speech assessment system for non-native children. He used English corpus spoken by German children and adopted various coefficients such as kappa, the Pearson correlation, classification rate, strictness and the Krippendorff's alpha to measure transcription agreement between L2 non-native experts and a native teacher. In this study, the transcription is performed on word, text, and speaker levels, not on phone level. The agreement result shows that kappa is around 0.4, and other coefficients are between 0.7 and 0.8 on word level. On text and speaker levels, kappa is around 0.5 and others are around 0.8. Pitt et al. [7] presented transcription agreement and transcription consistency of the Buckeye corpus, which is a spontaneous English speech corpus transcribed by six natives. The study described categorical agreements (0.899 for consonants and 0.69 for vowels) in addition to an overall agreement of 0.803. Furthermore, this study showed the agreement of reduced vowel is extremely low (around 0.3), although transcribed by natives. However, there is no serious study for non-native speech transcription by non-native transcribers, which can provide a consistency of agreement.

In this paper, we propose an economical transcription method for dealing with a large amount of non-native speech corpus transcribed by non-natives, which provides an improved transcription agreement. Transcription is performed on non-native English speech corpus produced by Korean speakers and transcribed by native Koreans.

The remaining part of this paper is organized as follows. Section 2 describes corpus and transcribers used for the experiment, and then proposes our transcription method. Transcription agreement of the resulting annotations and comparisons with other corpus transcribed by natives are presented in Section 3, which is followed by conclusion in Section 4.

2. Methodology

2.1. Corpus and Transcribers

We used the ETRI (Electronics and Telecommunications Research Institute) English speech corpus produced by Korean speakers. It consists of 19,883 English sentences of read speech by 100 Korean adult speakers (52 males and 48 females). Seven experts with phonetics background have participated in the transcription of all sentences. They are all native Koreans who can speak English as L2 and have experience of transcription. To check the transcription agreement and consistency, 498 sentences (2,839 tokens) randomly selected are transcribed by all seven transcribers, whereas for others sentences one sentence is transcribed by one transcriber.

2.2. Phone Set

For phonetic transcription, the phone set in CMU Pronunciation Dictionary [8] is augmented with additional seven phones, five vowels and two consonants, to explain Korean's English pronunciation.

Table 1 shows vowels. In addition to CMU phone set, we added AXR, IX, AX, EU and O. In Table 1, 8 vowels (IY, UW, OW, AO, IX, AX, AXR, and ER) exist in only English (L2) phone system, not in Korean. AXR, IX and AX are reduced vowels. They represent unstressed syllables. ER is syllabic r and occurs only in stressed syllable, whereas AXR in unstressed syllable. OW is cardinal vowel and usually diphthongized as [o^{o}] in stressed and open syllables as in American English. For production of AO, the tongue is retracted and the lips are rounded. Many non-native speakers have difficulty not only in producing, but also in perceiving it as a distinctive sound. IY and UW are also cardinal vowels in English and mark the highest boundary for the front vowel and the back vowel, respectively [9]. On the other hand, EU and O are unique vowels in L1. EU is high-mid vowel in Korean, and is usually inserted after consonant in the coda in English spoken by Korean. O corresponds to Korean monophthong, and replaces very often English OW or AO. O is pure monophthong, and its formant structure changes little in a period of vowel. Diphthong vowels are EY [e^{j}], AY [a^{j}], OY [ɔ^{j}], and AW [ɑ^{ʊ}].

Table 1. *Monophthong Vowels.*

	Front		Back	
High	IY (i) IH (ɪ)		UH (ʊ) EU (ɯ)	UW (u)
Mid	IX (ɛ) ER (ɝ)	AXR (ɚ)	OW (o^{o})	O (o)
Low	EH (ɜ) AE (æ)	AX (ə)	AH (ʌ) AA (ɑ)	AO (ɔ)

Table 2 represents consonants used for the transcription. In addition to CMU phone set, we added DX and TS. Among consonants, five phones (F, V, TH, DH and Z) are used in English only. F and V are labio-dental fricatives. They are often ranked as one of the troublesome sounds in English for non-native speakers learning to pronounce English as L2. TH and DH are interdental fricatives. They pose problems for the non-native speaker of English because English is almost unique in having both interdental fricatives and they are relatively difficult to perceive [9].

Table 2. *Consonants.*

	Stop	Fricative/Affricate	Sonorant	Glide
Labial	P (p) B (b)	F (f) V (v)	M (m)	
Dental	T (t) D (d) DX (ɾ)	TH (θ) DH (ð) S (s) Z (z) SH (ʃ) ZH (ʒ) TS (ʦ)	L (l) R (r) N (n)	
Palatal		CH (ʧ) JH (ʤ)		Y (j)
Velar	K (k) G (g)		NG (ŋ)	W (w)
Glottal		HH (h)		

2.3. Transcription Method

In order to help the transcribers to perform a consistent transcription, we propose that forced aligned pronunciation sequences as well as expected pronunciation errors are given to them as reference. A transcription window is provided using the annotation function of Praat 5.2.19 [10]. The transcription is performed on the sound waves and spectrograms with four tiers; word tier, segment tier, candidate tier, and actual tier as shown in Figure 1.

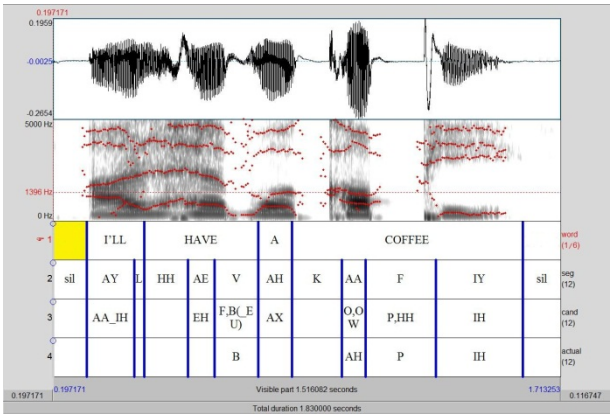

Figure 1: Transcription using Praat TextGrid

The word tier represents the words which are aligned to corresponding wave forms and spectrograms. HTK 3.4 [11] is used to automatically extract this information by forced alignment. We developed an acoustic model using a multi-pronunciation dictionary and a non-native speech corpus of 19,883 sentences described in Section 2.1.

The segment tier represents a forced aligned pronunciation sequence. The most optimized pronunciation sequence about non-native speech is chosen from multi-pronunciation dictionary by forced alignment and is converted to textgrid form of Praat with time alignment information.

Typical English L2 speech error patterns of Korean L1 speakers are provided as reference for Korean transcribers in the third tier, which is called the candidate tier. These error patterns are made by using knowledge-based rules [1] consisting of mapping rules and phonological rules that come from the difference of phoneme systems of L1 and L2.

The actual tier shows the result of transcription, where only phones different from the phones in the forced aligned phone sequence in the segment tier are marked by a transcriber. After listening to the sentence, if the transcriber judges that

spoken phone is different from the corresponding forced aligned phone in the segment tier, the recommended phone is marked in the actual tier. If not, nothing is marked. Deletion is marked by replacing a phone's label with a hyphen. Insertion is indicated by appending underbar and the inserted phone label to the leftmost neighboring phone, such as B_EU.

The candidate tier of Figure 1 shows that F, B, and B_EU are suggested references for V in the second word, 'HAVE'. This means that in Korean's English speech V tends to be mispronounced as F or B, and that vowel epenthesis, B_EU, is also a possible error in this position. This information given in the segment and candidate tiers helps the transcriber mark the real pronunciation as B in the actual tier. However, error patterns suggested in the candidate tier are not absolute options given to the transcriber to judge phones. The candidate tier works just as reference to help the labeler. For example, in case of AA of the fourth word COFFEE in Figure 1, the labeler transcribed AH as a mispronounced phone (over 100 cases), although only O and OW are suggested in the candidate tier.

2.4. Transcription Agreement

Agreement is measured in two ways. One is percentage agreement (PA) and the other is Fleiss' kappa. Measuring agreement is conducted by counting the number of phone labeling agreements for all pairs of seven transcribers. The number of all possible pairs of seven transcribers is the number of selecting two elements out of a set of seven elements ($_7C_2$). Hence, there are 21 pairs among seven transcribers. In these 21 pairs, we count agreement pairs and disagreement pairs. Agreement is calculated as (1).

For example, if a particular phone is labeled by six transcribers (T1-T6) as AH, but by one transcriber (T7) as AX, then the number of transcriber pairs who agree with each other is 15 (T1-T2~T6, T2-T3~6, T3-T4~6, T4-T5~6, T5-T6) and the number of transcriber pairs who disagree with each other is 6 (T1-T7, T2-T7, T3-T7, T4-T7, T5-T7, T6-T7). By (1), 15 pairs out of the 21 possible pair of transcriptions show percentage agreement as 0.714 (= 15/21).

$$\text{PA} = \frac{agreement\ pairs}{disagreement\ pairs + agreement\ pairs} \qquad (1)$$

Kappa is a measurement to get rid of chance agreement from percentage agreement. Chance agreement is probability of agreement at random. Fleiss developed kappa for multiple raters [12].

3. Results and Discussion

3.1. Overall Agreement

Overall phone level agreement is calculated on 9,327 phones from 498 sentences. Transcribers agreed on phone transcription in 86.90% of percentage agreement. Fleiss' kappa value obtained is 0.8685, which is similar to the percentage agreement. In case of phonetic transcription, as there are many phones to transcribe, the probability of random agreement between transcribers seems to be very tiny and can be ignorable.

In order to rule out the possibility of high agreement percentages because transcribers change nothing in the forced aligned result, percentage of changed symbol was calculated. The average percentage is 21.8%. [13] showed 10.5% of percentage of changed symbol as justified, therefore, our transcription procedure in which transcribers correct the forced aligned is also justified.

3.2. Categorical Agreement

In this section, we will see specific agreement by vowels and consonants. It is for observing agreement in detail in addition to overall agreement to compare with a result of agreement by native transcribers. In observing categorical agreement, agreement is calculated by percentage agreement.

3.2.1. Vowels

Vowels have 3,659 phones and 85.26% of agreement. The transcription agreement of monophthongs and diphthongs are shown in Table 3. Monophthongs have lower agreement than diphthongs. AO, AXR, ER, AX, OW, UW of monophthongs have lower agreement than average.

Table 3. *Agreement of monophthongs and diphthongs.*

Vowels	Agreement
Monophthong	0.837
Diphthong	0.961

The percentage agreement with regards to place of articulation is presented in Table 4. Vowel place of articulation consists of two dimensions; backness and height. Back vowel (0.758) has a lower agreement than front (0.893). In addition, low vowel (0.764) shows a lower agreement than high vowel (0.871). AO and AH in back and low vowel was confused with OW/O and AX, respectively, because AO and AH do not exist in L1.

Table 4. *Agreement classified by place of articulation.*

Place	Low	Mid	High	Sum
Front	0.822	0.904	0.921	**0.893**
Back	0.698	0.758	0.796	**0.758**
Sum	**0.764**	**0.850**	**0.871**	

3.2.2. Consonants

In our agreement experiment, consonants have 5,668 phones and it shows 87.91% of percentage agreement. Table 5 shows that agreement varied as manner of articulation. The agreement of fricative/affricate has the lowest agreement, and glide has the highest. Fricative/affricate has many phones that do not exist in Korean L1, such as DH, TH, V, F, and Z, while all English phones in glide are identical with those in Korean.

Table 5. A*greement classified by manner of articulation.*

Manner	Agreement
Stop	0.881
Fricative/Affricate	0.776
Sonorant	0.943
Glide	0.982

The percentage agreement with regards to place of articulation is presented in Table 6. Dental/alveolar shows lower consistency than any other classes. The class has Z, DH, and TH, which do not exist in L1. On the other hand, different categories do not have any unique English phones that do not exist in Korean. Glottal also shows low value like Dental/Alveolar. In case of glottal, a reason of low agreement is slightly different. When labelers transcribe wh- word like interrogatives, forced aligned sequence was given as HH W and labelers marked as HH W or – W (HH deletion). It makes low consistency of glottal.

Table 6. *Agreement classified by place of articulation.*

Place	Agreement
Labial	0.931
Dental/Alveolar	0.840
Palatal	0.948
Velar	0.961
Glottal	0.854

3.2.3. Comparison between transcription by native transcribers and by non-native

In this section, we will see in which aspect our method is effective in comparison with the Buckeye corpus, an English speech corpus transcribed by natives [7].

Table 7. *Comparison of Transcription Agreement by Natives (Buckeye corpus [7]) and Non-natives.*

	Buckeye (transcription by natives)		Experimental result (by non-natives)	
	N	agreement	N	agreement
Overall	2364	0.803	9327	0.869
Consonants	1457	0.899	5668	0.879
Vowels	907	0.69	3659	0.853

Table 7 shows that our experimental result in the proposed method of transcription has a higher degree of consistency. It means that our method using a forced aligned pronunciation sequence and a set of suggested error patterns in the candidate tier is effective for transcription by non-native labelers as a whole.

In consonants, transcription agreement of our data (87.91%) is similar to that of the Buckeye corpus (89.98%) [7] as shown in Table 7. Considering the result that agreement of transcription by non-natives is lower than by natives [4], our agreement value means that the proposed method is meaningful and contributes to supporting consistency of transcription. More specifically, degraded agreement was observed in some categories, such as fricative/affricate and dental/alveolar. Phones of L2 that do not exist in L1 are the reason of degradation.

Our data in Table 7 shows 85.26% of agreement in vowels. In comparison with it, the Buckeye corpus shows only 69 % of agreement [7]. Especially, reduced vowel has an exceedingly visible difference. The Buckeye corpus has about 30% of agreement. It means that reduced vowel is not easy for transcribers to transcribe and has high degree of confusion despite of transcription by natives. In other hand, we showed high level of consistency (81.7%). Our case explains that providing forced aligned pronunciation sequences to transcribers blocked possibility of confusion in advance.

4. Conclusion

We proposed an economical and effective transcription method for dealing with a large amount of non-native English speech corpus. To minimize the possibility of confusion in transcription process, forced aligned phone sequences and a set of possible mispronunciation candidate phones that Korean L2 learners are expected to make are given to the Korean transcribers for reference. The proposed method is evaluated by calculating the agreement using percentage agreement and kappa. The transcription consistency is analyzed by comparing it to that performed on the Buckeye corpus transcribed by natives. As a result, a 0.869 of the transcription agreement is achieved, while the Buckeye corpus shows 0.803 of agreement.

Experimental results showed that our method is effective for maintaining consistency of transcription in comparison with the Buckeye. Therefore, we conclude that the method proposed in this paper is helpful for transcription by non-natives. This can be extended to transcription of any languages, not only English, and by transcribers who speak any other language, not only Korean.

We expect that agreement of transcription could be negatively influenced when there is a fault in forced alignment. In our future work, recognition experiment will be performed to investigate negative influences for improving the method.

5. Acknowledgements

This work was supported by the Industrial Strategic Technology Development Program, 10035252, “Development of dialog-based spontaneous speech interface technology on mobile platform” funded by the Ministry of Knowledge Economy (MKE, Korea).

6. References

[1] Jang, T., "Construction of an English Speech Database for Korean Learners of English," *Language and Linguistics*, vol. 35, pp. 293-310, 2005 (in Korean).

[2] Hong, H., Kim, J. and Chung, M., "Effects of Korean Learners' Consonant Cluster Reduction Strategies on English Speech Recognition Performance," in *Proc. INTERSPEECH 2010*, Chiba, Japan, 2010.

[3] Bonaventura, P., Howarth, P. and Menzel, W., "Phonetic annotation of a non-native speech corpus," in *Proc. Conference of Integrating Speech Technologies in Learning (InSTIL)*, Dundee, U.K., 2000.

[4] Cole, R., Oshika, B.T., Noel, M., Lander, T. and Fanty, M., "Labeler agreement in phonetic labeling of continuous speech," in *Proc. ICSLP 1994*, Yokohama, Japan, 1994.

[5] Cucchiarini, C., "Assessing transcription agreement: methodological aspects," *Clinical Linguistics & Phonetics*, vol. 10, pp. 131-155, 1996.

[6] Hacker, C., "Automatic Assessment of Children Speech to Support Language Learning," Ph. D. Thesis, Engineering Faculty, Erlangen-Nurnberg University, NurnBerg, 2009.

[7] Pitt, M.A., Johnson, K., Hume, E., Kiesling, S. and Raymond, W., "The Buckeye corpus of conversational speech; labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, pp. 89-95, 2005.

[8] Carnegie Mellon University, The CMU Pronouncing Dictionary v. 0.7 a.
Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed on Nov. 15, 2009.

[9] Edwards, H.T., *Applied phonetics: the sounds of American English*, 3rd ed., Singular, 2002.

[10] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer (Version 5.2.21)” [Computer program]. Retrieved April 1, 2011, from http://www.praat.org/

[11] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D. and Povey, D., *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[12] Fleiss, J.L., Bruce, L. and Paik, M.C., *Statistical methods for rates and proportions*, 3rd ed., Wiley, 2003.

[13] Goddijn, S. and Binnenpoorte, D., “Assessing Manually Corrected Broad Phonetic Transcription in the Spoken Dutch Corpus,” in *Proc. 15th ICPhS*, Barcelona, Spain, 2003.