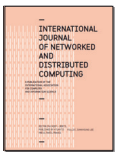
International Journal of Networked and Distributed Computing
Vol. **9**(**1**); January (2021), *pp.* 10–18
DOI: https://doi.org/10.2991/ijndc.k.201218.005; ISSN 2211-7938; eISSN 2211-7946
https://www.atlantis-press.com/journals/ijndc



Research Article

The Performance Evaluation of Continuous Speech Recognition Based on Korean Phonological Rules of Cloud-Based Speech Recognition Open API

Hyun Jae Yoo[1], Sungwoong Seo[1], Sun Woo Im[2], Gwang Yong Gim[1,*]

[1]*Department of IT Policy and Management, Graduate School, Soongsil University, Seoul, Korea*
[2]*Graduate School of Korean Language and Literature, Soongsil University, Seoul, Korea*

ARTICLE INFO

Article History

Received 09 October 2020
Accepted 18 November 2020

Keywords

Speech recognition
pronunciation dictionary
Korean phonological rules
cloud computing
Open API

ABSTRACT

This study compared and analyzed the speech recognition performance of Korean phonological rules for cloud-based Open APIs, and analyzed the speech recognition characteristics of Korean phonological rules. As a result of the experiment, Kakao and MS showed good performance in speech recognition. By phonological rule, Kakao showed good performance in all areas except for nasalization and Flat stop sound formation in final syllable. The performance of speech recognition of Korean phonological rules was good for /l/nasalization and /h/deletion. The speech recognition performance of phonological rule words accounted for a very high percentage of the whole words speech recognition performance, and the speech recognition performance of phonological rule was more different among companies than between speakers. This study hopes to contribute to the improvement of speech recognition system performance of cloud companies for Korean phonological rules and is expected to help speech recognition developers select Open API for application speech recognition system development.

1. INTRODUCTION

Speech recognition systems have significantly improved performance with cloud computing technology [1] and application of artificial intelligence [2]. The cloud-based speech recognition engine addresses the difficulties of developing speech recognition systems. By collecting large amount of speech data for development of speech recognition system, high performance computer for learning large volume speech data is not needed. Cloud-based speech recognition Open API has saved a lot of time, effort, and money to develop an applied speech recognition system. The improved performance and ease of development of speech recognition systems are being applied in a variety of areas. Speech recognition systems are largely divided into pre-processing and recognition units [3]. The recognition unit makes a word for the extracted speech information of the speech. The process of creating words either uses pronouncing dictionaries according to the characteristics of the speech recognition system [4–6], using information through deep learning of vocal information without pronouncing dictionaries [7,8]. The speech recognition system should accurately recognize phonological changes regardless of whether a pronouncing dictionary exists. In the speech recognition process, meaningful sentences should be made in terms of syllables by finding the exact morphemes. Therefore, assessing the recognition rate of speech recognition systems for phonological rules will help to understand the characteristics of speech recognition systems. This study aims to explore the characteristics of cloud-based speech recognition system's application of phonological rules and to present the criteria for selecting a high-performance cloud-based Open API for developing an applied speech recognition system. This study conducted a study on continuous speech recognition performance evaluation in accordance with the Korean phonological rules of the cloud-based speech recognition Open API. The composition of the paper described the related research on Korean phonological rules and cloud-based speech recognition Open API in Chapter 2, and Chapter 3 described experiment methods and test results as experiments. Chapter 4 summarizes the evaluation and meaning of experimental results and describes future research tasks.

2. BACKGROUND

2.1. Speech Recognition Overview

Speech recognition is a technique that converts a person's pronunciation into meaningful characters. The Korean Telecommunications Technology Association (TTA)'s information and communication terms dictionary describes speech recognition as "automatically identifying linguistic meaning contents from speech, and more specifically, it is a processing process that identifies words or series of words and extracts meanings by entering speech waveforms." The processing of speech recognition systems is divided into preprocessing and recognition units, as in Figure 1 [3]. In preprocessing, the input speech information is extracted, and in the

*Corresponding author. Email: gygim@ssu.ac.kr

recognition section, the speech information extracted from preprocessing is converted into words and the sentence is made. The creation of sentences uses pronouncing dictionaries and vocabulary dictionaries made in large TEXT coppers.

Factors affecting the performance of speech recognition system include noise elimination method, method of extracting speech characteristics, method of generating sound model, method of generating pronouncing dictionary, method of creating language model, and method of decoding network method. Pronouncing dictionary is a crucial factor in making speech information of speech recognition system a meaningful word. The pronouncing dictionary gives vocal information according to the heading. This vocalization information reflects phonetic rules to create a pronouncing dictionary. In particular, the difficulty of Korean speech recognition is to make a dictionary of pronunciation because there are so many rules that apply to the Korean pronunciation method in the generation of pronunciation.

2.2. Korean Phonological Rule

Phonological rules mean changing the predetermined pronunciation of the morpheme due to changes in phoneme and phenomena of change. The phonological process can be divided into official phonological processes and general phonological changes. The phonological process is divided into five parts [9] from the point of view of the syllabus: replacement, elimination, inclusion, condensation, and metathesis. Moreover, it is divided into essential phonological rules and veterinary rules, depending on the environment. Essential phonological rules are rules that must be applied in all conditioned environments, and optionally rules are rules that are both good and need not be applied in the same phoneme environment. Table 1 shows the classification and division

Figure 1 | Process of speech recognition system.

of public phoneme rules by phoneme process. Table 2 describes essential phonological rules and examples of related words in the synchronic phonological process, and Table 3 describes optionally phonological rules and examples of related words in the synchronic phonological process [9].

2.3. Prior Study on the Korean Pronunciation

Factors affecting the degradation of speech recognition rate in speech recognition systems include noise, completed pauses, repeat/repeat speech, pronunciation variation, stammering, and vocabulary diagram. Among these factors that reduce speech recognition is due to non-grammatical vocalization, except for noise factors. Many studies have been conducted on how to match spelling and pronunciation to reduce the error rate of speech recognition by non-grammatical speech. These methods include creating and using Grapheme to Phoneme (G2P) and learning pronunciation. The process of making a pronouncing dictionary is complex and has many maintenance limitations, so studies are being conducted on end-to-end speech recognition that does not require a pronouncing dictionary [10]. The existing pronouncing dictionary is a standard pronouncing dictionary based on linguistic standards, and the phonetic column is hand-written. However, this required professional knowledge of Korean phonological changes and required a lot of time and effort in writing. To solve these problems, we created a phonetic dictionary based on the Korean phonological rules [11]. This method performed particularly well in multi-pronunciation dictionaries. However, there is a problem with multiple pronunciations, which increases the size of the dictionary, increases the ambiguity of the perceived object at the recognition stage, and increases the congestion [12]. Thus, phonetic was extracted from two corpus of syllable unit and morpheme unit in consideration of phonological variation [4], and a pronouncing dictionary was created by establishing a new unit corpus in which morphological phonological variation was considered [5]. The size of the pronouncing dictionary decreased a lot and the error rate of the word also decreased. There was also a study without a

Table 1 | Classification and types of synchronic phonological rules

Synchronic phonological process	Formative phonological process	Formative phonological process of consonants	Replacement	Flat stop sound formation in final syllable, Nasalization, Liquidization, Place assimilation, Fortition
			Deletion	Simplification of Consonant cluster, /h/deletion, Geminate consonants Reduction, /t/deletion, /l/deletion, Nasal deletion
			Insertion	Gemination, /n/insertion, Homophony insertion
			Contraction	Aspiration, Fortition
		Formal phonological process for vowels and semi-vowels	Replacement	Vowel harmony, Umlaut, /j/semi-vowelization /w/semi-vowelization, Complete assimilation of vowel, front-vowelization, Vowel rounding
			Deletion	/ɯ/deletion, Same vowel elision, /j/deletion, /w/ deletion
			Insertion	/j/insertion
			Contraction	Vowel coalescence
	Joint process	Formative phonological process of consonants	Replacement	Implosive formation, Voicing (voicing assimiliation), Lateralization, Palatalization
		Formal phonological process for vowels and semi-vowels	Replacement	w-fronting (w front-vowelization)
			Deletion	/j/deletion

Table 2 | Explanation and examples of essential phonological rules

Phonological rules	Explanation	Example: Pronunciation variation (Before → After), Symbol (IPA)
Flat stop sound formation in final syllable	A phenomenon in which the obstruent changes from the final consonant to one of the final syllable neutralization /p, t, k/.	jʌp^h → jʌp, sot^h → sot
Nasalization	Final syllable neutralization /p, t, k/ assimilate into nasal /m, n, ŋ/ respectively in front of nasal sound.	pabman → pamman, padnɯn → pannɯn
Liquidization	/n/ encounters /l/ and turns into /l/.	mulnoli → mullori, ɕilnɛ → ɕillɛ
Fortition	A phonological phenomenon in which a plain consonant among obstruent is changed to a fortis in a certain environment.	tsabgo, → tsapk'o, midgo → mitk'o
Simplification of consonant cluster	In the case of a group of consonants consisting of two consonants, one of the two consonants is dropped out of the final consonant.	nʌgtsto, → nʌkt'o, ʌntsnɯn) → ʌnnɯn
/h/deletion	The predicate final consonant /h/ is a dropout phenomenon in front of a vowel.	nahɯn → naɯn, anha → aˈna
/t/deletion	/t/ is dropped in front of /s'/.	tsʌdzso → tsʌs'o, os sanda → os'anda
/l/deletion	The predicate final consonant /l/ are dropped in front ofthe first consonant /n/, thefinal consonant /n, l, m, p/, the pre-final ending '-ɯsi-, -ɯo-', the sentence-closingending ' -ɯo, -ɯma'.	manduɯl-m nida → manduɯmnida
Aspiration	A phenomenon in which the final syllable neutralization and the flat spirant become aspiration sounds when they meet /h/.	nohda → not^{h}a, anhgo → (ank^{h}o
Complete assimilation of vowel	In Gyeongsang dialect, the mediated vowel /ʌ/ is completely assimilated to the frontvowel.	tah-ʌmo → taːmo
/ɯ/deletion	/ɯ/ is dropped out under various conditions.	s'ɯ-ʌ → s'ʌ, h^{h}ɯ-ʌ → t^{h}ʌ
Same vowel elision	When vowel endings are connected after /a, ʌ/ terms, endings /a, ʌ/ are dropped out.	ka-ʌ → kaː, sʌ-ʌ → səː
/j/deletion	[j] is eliminated after the palatal sound [ɲ, ʃ, ʃ', ʎ].	tʌndzi-ʌ → tʌndzʌ, igi-e → iˈkeː
/w/deletion	In some dialects, when the conjugated form is connected with a bilabial sound, a tongue front sound and a double vowel /wa, w/ is eliminated.	po-a → pwa → paː
Lateralization	[r] is the lateralization of [l] at the final consonant or after [l].	oɾɯnparro → oɾɯnballo, tarrara → tallaɾa
Palatalization	A phenomenon in which dental sound [n], alveolar sound [s, s', l] change to palatal sound [ɲ, ʃ, ʃ', ʎ] in a front sound [i, j, ɥ], respectively.	kas'ni/kanni, jʌnlʎo/jʌlʎo

Table 3 | Explanation and examples of optionally phonological rules

Phonological rules	Explanation	Example: Pronunciation variation (Before → After), Symbol (IPA)
Place assimilation	A Phenomenon that /t, n/ is changed to /p, m/ in front of bilabial, /k, ŋ/ in front of dorsal (back). And /p, m/ is changed to /k, ŋ/ in front of dorsal (back).	mit^{h}p^{h}an → mibp^{h}an, tsip^{h}ko → tsik'o
Geminate consonants Reduction	In front of fortis and aspirate of stop and fricative sound, a flat (lax) sound /p, t, k/ is optionally deleted at same place	pabp^{h}ul → pap^{h}ul, tɯdtsa → tɯts'a
Gemination	In front of fortis and aspirate of stop and fricative sound, a flat (lax) sound /p, t, k/ is optionally inserted at same place	ap'a → app'a, ap^{h}asə → app^{h}asə
/n/insertion	A phenomenon that occurs optionally when the preceding word ends with a consonant and the word behind it starts with /i, j/ when a compound or derivative word is created	pamil → pamɲil, polil → polɲil
Umlaut	The back vowel /a, ʌ, o, u/ is changed to front vowel /ɛ, e, ø, y/ due to the influenceof the following front vowel 'i' or glide 'j'	pab-i → pɛˈbi, t'əg-i → t'egi
/j/semi-vowelization	When ending of vowel is connected behind predicate /i/, auslaut of predecate /i/ is optionally changed to semi-vowel /j/	p^h-ə → p^{h}jʌ
/w/semi-vowelization	When ending of vowel is connected behind predicate /o, u/, auslaut of predicate /o, u/ is optionally changed to semi-vowel /w/	po-a → pwa
Front-vowelization	Onset of postposition or ending /ɯ/ is changed to /i/ behind sibilant /s, s', ts, ts', ts^h/	os-əno → osino
/j/insertion	/j/ is inserted optionally when ending onset /ʌ/ is connected to predicate /i, e, ɛ, wi, ø/.	p^{h}i-ə → p^{h}jʌ

pronouncing dictionary. Although the G2P process that requires changes in phoneme and exception processing of Hangeul is necessary [13], there is a study that breaks down the method of recognizing through deep learning without the G2P process [7] and uses it as an output unit of sound model [8] by breaking it down into letters in initial, neutral, and ending. The method of using lettering showed better performance than pronouncing dictionaries. A new set of phonetic phonemes was created by clustering the ignited voices into a common spectral pattern to increase the discriminative power [6]. Pronouncing dictionaries using

common spectral patterns had an effect of reducing the relative word error rate of 8.9% in the phonetic speech than phonetic pronouncing dictionaries, and free speech data by about 7.0%. This study will be meaningful in evaluating the pronunciation treatment of the cloud company speech recognition system through the performance evaluation of Korean phonological rules for the cloud company speech recognition system.

2.4. Cloud Speech Recognition Open Application Programming Interface

Cloud-based speech recognition Open Application Programming Interface (API) is an application service in cloud computing environment. Cloud computing is a service that remotely orders and pays for computer resources (such as software, hardware, storage, etc.) and uses them [1,14]. Cloud computing has characteristics such as multitenancy, on-demand usage, usage measurement, elasticity, resilience, and ubiquitous access. The advantages of cloud computing are, first, low investment and lower maintenance costs. Second, the scalability of computer resources is good. Third, the service configuration is short. Fourth, availability and reliability are high. Fifth, rapid decision-making by the organization of the system configuration is reflected. The downside is, first, that it is vulnerable to security. The stability of data should be delegated to external companies. Second, it is difficult to transfer data when changing service provider. Third, data may be required to be disclosed in accordance with local regulations and regulations of the service provider [1,14]. Cloud-based speech recognition Open API is an API that enables speech recognition developers to develop speech recognition systems using the characteristics of cloud computing. The difficulty of developing a speech recognition system should be based on high-performance computers that can collect large-capacity speech data and learn large-scale speech data. However, cloud-based speech recognition Open API addresses the difficulties of developing speech recognition systems. The cloud-based Open API allows application speech recognition developers to implement desired application speech recognition systems quickly and easily. Companies providing cloud-based speech recognition Open API are represented by domestic Kakao Speech-to-Text system [15], SKT NUGU [16], Naver Clova Speech Recognition [17], GiGA Genie Speech Recognition [18], ETRI STT [19], and others, while foreign companies are Microsoft Azure Cognitive Speech Service [20], Amazon Transcribe [21], IBM Watson Speech to Text [22], and Google Cloud Speech-to-Text [23].

2.5. Prior Study on Cloud Speech Recognition Open API

Cloud-based speech recognition Open API supports development of application speech recognition system quickly and easily. Due to the convenience of development using cloud-based speech recognition Open API, applied speech recognition system is being established in various fields. Application speech recognition developers should choose the speech recognition Open API appropriate for their application speech recognition system, depending on the function and performance they want in developing the application speech recognition system. There are many cases of cloud speech recognition Open API performance evaluation studies to provide criteria for this choice. Cloud-based speech recognition Open API shows performance differences depending on the timing of research and the nature of learning data. The March 2017 study found that Google API was the best [24]. In August 2017, a study conducted experiments on numbers, Hangul, and sentences. The numbers were Kakao, and Naver performed well in Hangeul and sentences [25]. The October 2017 study conducted an experiment on sentences, and the main factors in sentences in which recognition errors occurred were words in Portuguese and English, acronyms, names and certain corporate terms. The Google Cloud Speech API had the highest accuracy. However, the speed was found to be the slowest [26]. In the December 2017 study, the Korean people's standard language and dialect were studied according to gender, age, and region. The accuracy of the sentences was measured based on spacing, props, surveys, and words according to the resulting sentences. Overall accuracy was good for Google, dialect was good for the Chungcheong and Jeolla dialects, and in the Gyeongsang dialects, sentences with large differences in intonation and pitch and unfamiliar Gyeongsang dialect words were not well recognized [27]. In the December 2018 study, Google showed moderate performance, unlike previous studies [28]. In the 2019 study, Korean and English sentences were recorded at a distance of 1, 3 and 5 m [29]. ETRI Open API in Korean, ETRI Open API in 1 m, Naver in 3 m, Naver Clova in 5 m, Microsoft Azure Speech Service in English, Microsoft Azure Speech Service in 1 m, Amazon Transcribe in 3 m, and ETRI Open API in 5 m showed high recognition rates [29]. According to prior research from 2017 to 2019, Google showed good performance in the beginning, but Microsoft Azure Speech Service and ETRI Open API showed good performance in the second half. In the preceding study, there is no case of speech recognition research on changes in Korean phonemes and changes. It would be a meaningful study to evaluate phonetic recognition of phonetic fever in identifying the characteristics of cloud-based speech recognition Open API.

3. EXPERIMENT

3.1. Experimental Method

The experiment tested the speech recognition performance of the cloud-based speech recognition Open API for Korean phonological rules. The Korean phonological rules selected 10 essential phonological rules (nasalization, /t/deletion, palatalization, /h/deletion, simplification of Consonant cluster, fortition, /l/nasalization, flat stop sound formation in final syllable, aspiration, liquidation) that occur during the public phonological process [9]. The cloud-based speech recognition Open API targeted seven domestic and foreign cloud companies (Kakao, ETRI, Naver, Microsoft, Google, IBM, Amazon, IBM). Speech data recorded a total of 100 sentences and 2560 phrases, 10 sentences each for 10 syllables by phonological rule, such as Table 4 [9,30,31]. Five speakers, male and female, participated in the recording, and the recording environment was recorded in a general office without soundproofing facilities. The format of the speech data was 16-bit PCM with a sampling of 16 kHz. Cloud-based speech recognition Open API did not consider any speech recognition options provided by cloud companies. The service method was chosen as a non-streaming method. The experimental equipment used a web program developed using PHP for

Table 4 | Examples of words by experimental phonological rules

Phonological rules	Examples: Pronunciation variation (Before → After), Symbol (IPA)
Nasalization	tsʌpnɯn → tsʌmnɯn, pʌpman → pʌmman, patnɯn → pannɯn, natsʰman → nanman, magnɯn → maŋnɯn,t'ʌgman → t'ʌŋman, tak'nɯn → taŋnɯn, ipʰman → imman, u:snɯn → u:nnɯn, mitnɯnta → minnɯnta.
/t/deletion	tsʌdzsoka → tsʌs'oka, mitsaoni → mis'a'o'ni, kas'ɯbnida → kas'ɯmnida, patsɯbnida → pas'ɯmnida, kɯɾɯsɛ'isgo → kɯ'ɾɯt ɕ'ik'o, mutsɯbnida → mus'ɯmnida, tɯtsɯbnida → tɯs'ɯmnida, tatsɯbnida → tas'ɯmnida,kətsɯbnida → kəs'ɯmnida, kotsɯbnida → kos'ɯmnida.
Palatalization	mitʰi → mitsʰi, patʰita → patsʰita, ku'ti → ku'dzi, katʰi → katsʰi, hɛ'dotiɾɯl → hɛ'dodziɾɯl, ma'ti i'ni → ma'dzi i'ni,kʌtʰi → kʌtsʰi, putʰida → putsʰida, satʰsatʰi → sa's'atsʰi, kuthjʌs'ta → kutsʰʌta.
/h/deletion	na'ha → na'a, ɕi'ɾhʌhanda → ɕi'ɾʌɦanda, ma:nhɯn → ma:nɯn, s'a'hida → s'a'ida, no'hinda → no'inda, k'ɯlhidaka → k'ɯlidaka, anhɯn → annɯn, a'nha → a'na, a'ɾha → a'ɾa, aɾhɯn → aɾɯn
Simplification of consonant cluster	nʌksto → nʌkt'o, ʌndznɯn → ʌnnɯn, ku:lmko → ku:mk'o, sa:lmmani → sa:mmani, haltʰtsi → halts'i, k'ɯlhnɯnda → k'ɯllɯnda, palktsi → pakts'i, nʌlbta → nʌlt'a, ɯlpʰko → ɯpk'o, alhnɯn → allɯn.
Fortition	kukputʰʌ → kukp'utʰʌ, tsapko → tsapk'o, papto → papt'o, mittsi → mits'i, u:sko → u:tk'o, a:nko → a:nk'o,o:mtsi → o:mts'i, kaltɯŋɯn → kalt'ɯŋɯn, multsilɯn → mults'ilɯn, solpaŋulɯl → solp'aŋulɯl.
/l/nasalization	nɯŋljʌkdo → nɯŋŋʌkdo, hamljaŋ → hamŋaŋ, ɯ'munlon → ɯ'munnon, hjʌpljʌkhajʌ → hjʌmŋʌkha'jʌ,homlʌnɯl → hom-nʌnɯl, ʌplodɯ → ʌpnodɯ, ɕimlilɯl → ɕimŋilɯl, pjə:ŋljʌlɯn → pjə:ŋŋʌlɯn, tsikljʌllo → tsiŋŋʌllo, ta:mljʌk → ta:mŋʌk.
Flat stop sound formation in final syllable	jʌpʰ → jʌp, tə:pʰko → tə:pk'o, ipʰto → ipt'o, nastwa → nat'o, is'ta → it'a, pits'to → pit'o, nohnɯn → nonnɯn, tak'nɯn → taknɯn, osman → onman, apʰman → amman.
Aspiration	mathjʌŋɯn → matʰjʌŋɯn, nohko → nokʰo, nohtaga → notʰaga, nəhtʌɾa → nətʰʌɾa, ma:nhkʌdɯn → ma:nkʰʌdɯn,k'ɯl-htsido → k'ɯltsʰido, tsohtsinɯn → tso'tsʰinɯn, pʌphakkwanɯn → pʌpʰak'anɯn, palkhjʌtsugi → palkʰjʌtsu'gi,iphaki → ipʰaki.
Liquidization	mulnori → mullori, s'alnunɯi → s'allunɯi, pulnɯŋ → pullɯŋ, ɕilnɛhwaɾɯl → ɕilleɦwaɾɯl, hultʰnɯn → hullɯn, tal-nimɯl → talʎiml, alhnɯn → allɯn, sonnanloka → sonna:lloka, ɕinlae'ke → ɕillae'ke, onlain → ollain.

general desktop computers. The evaluation method was measured in words. The recognition performance was verified by calculating the Word Error Rate (WER) in sentence, as shown in Equation (1). In Equation (1) S means Substitution, I mean Insertion, D means Delete, and *N* means the whole input phrase.

$$\text{Word Error Rate (WER)} = \frac{(S+I+D)}{N} \quad (1)$$

In the evaluation, the error rate of the whole words (sentence containing the phonological rule word) and the error rate of the phonological rule word were measured, respectively. Whole words consisted of sentences containing phonological rules (Ex: KOREAN-> "종이를 접는 방법들 배우고 싶다" / IPA->" tso'ŋiɾɯl tsʌmnɯn paŋbʌpɯl pɛ'ugo ɕipt'a" / ENGLISH->"I would like to learn how to fold paper", the phonological rule word is KOREAN->"접는" / IPA->"tsʌmnɯn"/ ENGLISH->"folding") [32].

3.2. Experiment Result

As a result of the experiment, as shown in Table 5, the error rate for all words on a per cloud company basis was 8.09% for Microsoft and 8.28% for Kakao, showing good performance. IBM 43.38% and Naver 19.02% showed poor performance. As shown in Table 6, the error rates for phonological rules were 18.00% for Kakao and 25.60% for Microsoft, which showed good performance. IBM 71.20% and Amazon 36.00% did not performance well. As shown in Table 5, the error rates of sentences containing phonological rules word on the basis of phonological rules were for /l/nasalization 12.32% and /h/deletion 13.20%, showing good performance. Palatalization 23.48% and aspiration 22.04% showed poor performance. In Table 6, the error rates for phonological rules word were /h/deletion 16.29% and /l/nasalization 20.57%, which showed good performance, while simplification of consonant cluster 61.43% and aspiration 49.14% showed poor performance. Table 7 show the ratio of the number of incorrect words in phonological rules words to the number of incorrect words in whole words. The ratio was as low as 31.34% for IBM and 35.11% for Naver, and as high as 61.84% for Microsoft and 50.57% for Google. As shown in Table 8, the speech recognition error rate for whole words by speaker was 14.79% to 20.34%, and the speech recognition error rate for phonological rule words by speaker was 30.43% to 40.00% as shown in Table 9.

4. CONCLUSION

In this paper, a study was conducted on continuous speech recognition performance in accordance with the Korean phonological rules of the cloud-based speech recognition Open API. First, the results of the experiment were compared and analyzed the speech recognition performance of the cloud-based speech recognition Open API. In Figure 2, the whole words error rate and phonological rule words error rate by cloud company both showed good performance for Kakao and MS, while IBM and Naver showed low performance. Looking at Table 7's ranking of error rates for corporate phonological rule phrases, Kakao showed good performance in all areas except nasalization and flat stop sound formation in final syllable, while Microsoft showed good performance in nasalization and Google showed good performance in flat stop sound formation in final syllable. Table 7's second-place group showed Google performing well in two areas of aspiration and liquidization, Amazon in two areas of fortition and final syllable neutralization, and Naver in /t/deletion and ETRI in /h/deletion. Cloud company's speech recognition Open API showed good performance for certain phonological rules. Second, the speech recognition characteristics of the Korean phonological rules were analyzed. In Figure 3, the whole words error rate and the phonological rule words error rate were both good for /h/deletion and /l/nasalization, and palatalization, simplification of consonant cluster, and aspiration

Table 5 | Speech recognition error rate for whole words by company (WER%)

	Amazon	ETRI	Google	IBM	Kakao	MS	Naver	Sum of wrong words	Total number of words	WER%
Nasalization	12%	13%	9%	44%	10%	8%	20%	309	1855	16.66
/t/deletion	12%	13%	13%	42%	8%	10%	18%	270	1610	16.77
Palatalization	22%	18%	17%	56%	13%	13%	25%	452	1925	23.48
/h/deletion	13%	4%	8%	45%	4%	4%	15%	231	1750	13.20
Simplification of consonant cluster	21%	22%	16%	43%	14%	11%	22%	415	1960	21.17
Fortition	16%	12%	8%	40%	9%	7%	13%	254	1715	14.81
/l/nasalization	11%	5%	6%	42%	5%	3%	15%	220	1785	12.32
Flat stop sound formation in final syllable	9%	11%	4%	44%	5%	5%	20%	270	1960	13.78
Aspiration	27%	23%	13%	41%	9%	16%	26%	378	1715	22.04
Liquidization	16%	11%	7%	46%	6%	4%	16%	251	1645	15.26
Sum of wrong words	404	343	261	1136	212	207	487	3050	17920	17.02
Total number of words	2560	2560	2560	2560	2560	2560	2560	17920	–	–
WER%	15.78	13.40	10.20	44.38	8.28	8.09	19.02	17.02	–	–

Table 6 | Speech recognition error rate for phonological rules by company (WER%)

	Amazon	ETRI	Google	IBM	Kakao	MS	Naver	Sum of wrong words	Total number of words	WER%
Nasalization	36.00%	32.00%	26.00%	76.00%	24.00%	20.00%	24.00%	119	350	34.00
/t/deletion	46.00%	44.00%	50.00%	92.00%	24.00%	44.00%	40.00%	170	350	48.57
Palatalization	42.00%	34.00%	36.00%	60.00%	8.00%	26.00%	34.00%	120	350	34.29
/h/deletion	14.00%	2.00%	8.00%	60.00%	0.00%	12.00%	18.00%	57	350	16.29
Simplification of consonant cluster	72.00%	56.00%	54.00%	88.00%	44.00%	52.00%	64.00%	215	350	61.43
Fortition	20.00%	26.00%	20.00%	60.00%	16.00%	16.00%	22.00%	90	350	25.71
/l/nasalization	24.00%	12.00%	12.00%	58.00%	6.00%	10.00%	22.00%	72	350	20.57
Flat stop sound formation in final syllable	14.00%	26.00%	6.00%	78.00%	18.00%	20.00%	40.00%	101	350	28.86
Aspiration	52.00%	72.00%	36.00%	60.00%	26.00%	42.00%	56.00%	172	350	49.14
Liquidization	40.00%	18.00%	16.00%	80.00%	14.00%	14.00%	22.00%	102	350	29.14
Sum of wrong words	180	161	132	356	90	128	171	1218	3500	34.80
Total number of words	500	500	500	500	500	500	500	3500	–	–
WER%	36.00	32.20	26.40	71.20	18.00	25.60	34.20	34.80	–	–

Table 7 | Ranking of WER% in terms of phonological rules by company

	1st		2nd		Average (%)
	Company	WER%	Company	WER%	
Nasalization	MS	20.00	Kakao, Naver	24.00	34.00
/t/deletion	Kakao	24.00	Naver	40.00	48.57
Palatalization	Kakao	8.00	MS	26.00	34.29
/h/deletion	Kakao	0.00	ETRI	2.00	16.29
Simplification of consonant cluster	Kakao	44.00	MS	52.00	61.43
Fortition	Kakao, MS	16.00	Amazon, Google	20.00	25.71
/l/nasalization	Kakao	6.00	MS	10.00	20.57
Flat stop sound formation in final syllable	Google	6.00	Amazon	14.00	28.86
Aspiration	Kakao	26.00	Google	36.00	49.14
Liquidization	Kakao, MS	14.00	Google	16.00	29.14
Total	Kakao	18.00	MS	25.60	34.80

were poor. Table 10 represents the ratio of the number of wrong words of phonological rules words to the number of wrong words of whole words by company. The ratio is very high, from at least 35.11% to up to 61.84%. Figures 4 and 5 show similar alignments in both speech recognition error rates for whole words and speech recognition error rates for phonological rule words. In other words, the speech recognition performance of the phonological rule words is affecting the speech recognition performance in the whole words. Figures 6 and 7 represent speech recognition error rates for phonological rule words by company and speaker, and the

Table 8 | Speech recognition error rate for whole words by speaker (WER%)

	A	B	C	D	E	Sum of wrong words	Total number of words	WER%
Nasalization	13.21%	24.26%	14.82%	14.29%	16.71%	309	1855	16.66
/t/deletion	15.53%	19.25%	17.39%	13.35%	18.32%	270	1610	16.77
Palatalization	20.78%	26.49%	19.74%	23.64%	26.75%	452	1925	23.48
/h/deletion	12.29%	16.29%	13.43%	10.00%	14.00%	231	1750	13.20
Simplification of consonant cluster	19.13%	24.49%	20.15%	21.68%	20.41%	415	1960	21.17
Fortition	12.83%	18.95%	11.37%	10.50%	20.41%	254	1715	14.81
/l/nasalization	12.04%	14.29%	13.73%	9.80%	11.76%	220	1785	12.32
Flat stop sound formation in final syllable	10.97%	15.82%	9.44%	12.76%	19.90%	270	1960	13.78
Aspiration	20.41%	24.78%	20.99%	19.53%	24.49%	378	1715	22.04
Liquidization	12.77%	17.93%	12.77%	10.64%	22.19%	251	1645	15.26
Sum of wrong words	539	729	552	530	700	3050	17920	17.02
Total number of words	3584	3584	3584	3584	3584	17920	–	–
WER%	15.04	20.34	15.40	14.79	19.53	17.02	–	–

Table 9 | Speech recognition error rate for words of phonological rules by speaker (WER%)

	A	B	C	D	E	Sum of wrong words	Total number of words	WER%
Nasalization	30.00%	45.71%	31.43%	27.14%	35.71%	119	350	34.00
/t/deletion	45.71%	47.14%	57.14%	44.29%	48.57%	170	350	48.57
Palatalization	31.43%	41.43%	34.29%	34.29%	30.00%	120	350	34.29
/h/deletion	12.86%	24.29%	15.71%	10.00%	18.57%	57	350	16.29
Simplification of consonant cluster	60.00%	70.00%	55.71%	62.86%	58.57%	215	350	61.43
Fortition	27.14%	35.71%	20.00%	15.71%	30.00%	90	350	25.71
/l/nasalization	28.57%	18.57%	22.86%	11.43%	21.43%	72	350	20.57
Flat stop sound formation in final syllable	24.29%	34.29%	27.14%	27.14%	31.43%	101	350	28.86
Aspiration	50.00%	48.57%	50.00%	48.57%	48.57%	172	350	49.14
Liquidization	24.29%	34.29%	27.14%	22.86%	37.14%	102	350	29.14
Sum of wrong words	234	280	239	213	252	1218	3500	34.80
Total number of words	700	700	700	700	700	3500	–	–
WER%	33.43	40.00	34.14	30.43	36.00	34.80	–	–

Table 10 | Ratio of phonological rules to whole words by company

	Amazon	ETRI	Google	IBM	Kakao	MS	Naver	Sum of wrong words
Total number of wrong words (A)	404	343	261	1136	212	207	487	3050
The number of words with wrong phonological rules (B)	180	161	132	356	90	128	171	1218
B/A ratio (%)	44.55	46.94	50.57	31.34	42.45	61.84	35.11	39.93

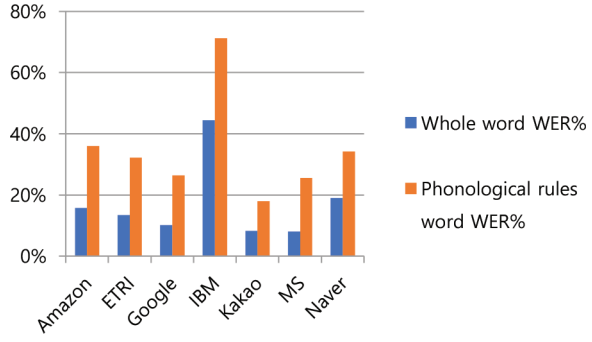


Figure 2 | Comparison of WER% by company.

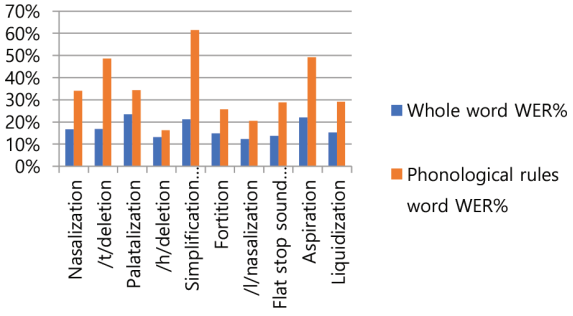


Figure 3 | Comparison of WER% by phonological rules.

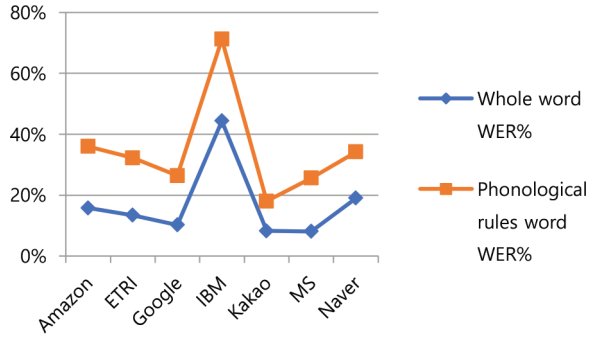


Figure 4 | Linear comparison of WER% by company.

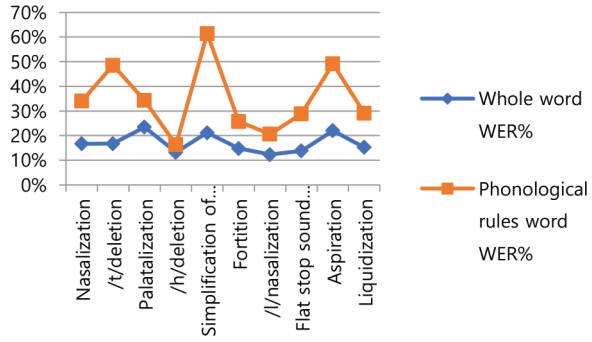


Figure 5 | Linear comparison of WER% by phonological rules.

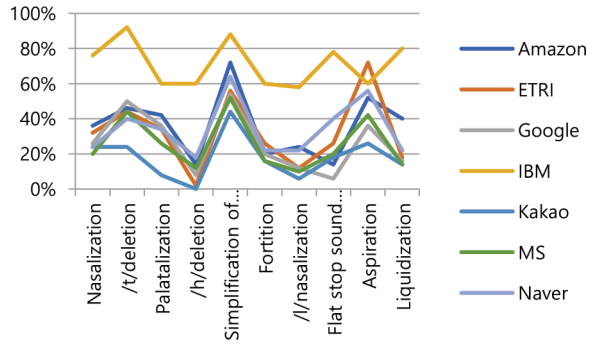


Figure 6 | Comparison of WER% of phonological rules by company.

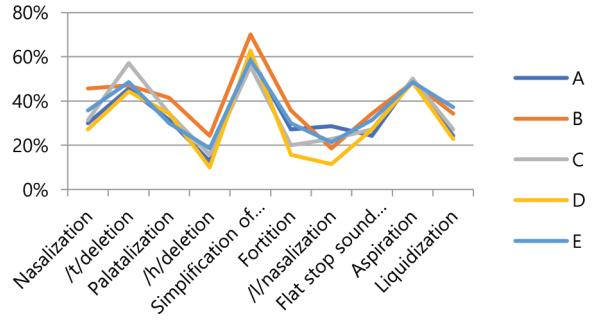


Figure 7 | Comparison of WER% for phonological rules by speaker.

linear shape in Figure 6 shows a more distracting linear form than in Figure 7. In other words, speech recognition performance for phonological rules can be attributed more to the speech recognition engine of cloud companies than to the speaker. Through this study, we confirmed that the speech recognition performance of the cloud-based speech recognition Open API for Korean phonological rule differs between companies, and that the speech recognition system of the same cloud company also has characteristics that show different performance by Korean phonological rule. According to the characteristics of speech recognition by Korean phonological rule, first, there was a difference in speech recognition performance by phonological rule. Second, speech recognition performance of phonological rule words had a significant impact on the overall speech performance. Third, speech recognition performance for phonological rule words was more different between companies than speakers. Therefore, this research will contribute to improving the Korean phonological rule speech recognition performance of the speech recognition engine of the cloud computing company and help speech recognition developers select the Open API to develop an applied speech recognition system.

A future task is to evaluate the performance of speech recognition on the optionally phonological rules of the synchronic phonological process. The result is expected to be different from the speech recognition rate of the essential phonological rule because the rules may or may not be applied in the same phoneme environment. Following the essential phonological rules of the synchronic phonological process, the study of the evaluation of speech recognition performance for the optionally phonological rules will be meaningful to improve the performance of speech recognition in Korean.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] J.H. Jeong, Current status and challenges of cloud computing, NARS Issue Rep. 313 (2017), 17–21.

[2] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (2012), 82–97.

[3] Trends and Prospects of Voice Recognition Technology, Korea Creative Content Agency's Cultural Technology (CT) In-depth Report, 11 (2011), 2020, Available from: https://www.kocca.kr/cop/bbs/view/B0000144/1756144.do?menuNo=.

[4] K.N. Lee, M.H. Chung, Morphological analysis of spoken Korean based on pseudo-morphemes, Proceedings of the Annual Conference on Human and Language Technology, Korean Institute of Information Scientists and Engineers, Busan, Korea, 10 (1998), pp. 396–404.

[5] J.U. Bang, S.H. Kim, O.W. Kwon, Performance of speech recognition unit considering morphological pronunciation variation, Phonet. Speech Sci. 10 (2018), 111–119.

[6] J.U. Bang, S.H. Kim, O.W. Kwon, Performance of Korean spontaneous speech recognizers based on an extended phone set derived from acoustic data, Phonet. Speech Sci. 11 (2019), 39–47.

[7] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, P. Nguyen, On the choice of modeling unit for sequence-to-sequence speech recognition, Proc. Interspeech 7 (2019), 3800–3804.

[8] M.H. Lee, J.H. Chang, Korean speech recognition based on grapheme, J. Acoust. Soc. Korea 38 (2019), 601–606.
[9] J.c. Bae, Opening of Korean Phonetics, third ed., (Hak)Shingu media & publishing, Gyeonggi Sungnamsi Jungwongu, Korea, 2018.
[10] W. Chan, N. Jaitly, Q. Le, O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition", Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, 2016, pp. 4960–4964.
[11] L.G. Nim, J.M. Hwa, Pronunciation dictionary for continuous speech recognition (in Korean), Proc. KIISE. Conf. 27 (2000), 197–199.
[12] P. Younghee, M. Chung, Pseudomorpheme-based Korean continuous speech recognition using tagged word bigram, Korean Inst. Inform. Sci. Eng. 26 (1999), 351–353.
[13] J.W. Yoo, A study on method of constructing pronunciation unit for continuous speech recognition, The Korean Electronics and Telecommunications Research Institute report, ETRI-94-03295, 1 (1995).
[14] L. Chang-Beom, Legal tasks for safe use and revitalization of cloud computing, Review of The Korea Institute of Information Security and Cryptology (Review of KIISC) 20 (2010), 32–43.
[15] Guide of Kakao Speech API, 2020, Available from: https://developers.kakao.com/docs/latest/ko/voice/.
[16] Guide of NUGU SDK Developers, 2020, Available from: https://developers-doc.nugu.co.kr/nugu-sdk.
[17] Guide of Clova Speech Recognition, 2020, Available from: https://www.ncloud.com/product/aiService/csr.
[18] Guide of GiGA Genie Speech Recognition API, 2020, Available from: https://apilink.kt.co.kr/api/menu/apiSpcDetail.do?apiSpcId=57.
[19] Guide of aihub Speech Recognition API, 2020, Available from: http://www.aihub.or.kr/ai_software/370#group00.
[20] Guide of Azure Speech to Text, 2020, Available from: https://azure.microsoft.com/ko-kr/services/cognitive-services/speech-to-text/.
[21] Guide of Amazon Transcribe, 2020, Available from: https://aws.amazon.com/ko/transcribe.
[22] Guide of Watson Speech to Text, 2020, Available from: https://www.ibm.com/kr-ko/cloud/watson-speech-to-text.
[23] Guide of Google Speech-to-Text, 2020, Available from: https://cloud.google.com/speech-to-text/.
[24] V. Këpuska, G. Bohouta, Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx), Int. J. Eng. Res. Appl. 7 (2017), 20–24.
[25] S.J. Choi, J.B. Kim, Comparison analysis of speech recognition open APIs' accuracy, Asia Pac. J. Multim. Serv. Converge. Art Human. Sociol. 7 (2017), 411–418.
[26] A.L. Herchonvicz, C.R. Franco, M.G. Jasinski, A comparison of cloud-based speech recognition engines, Computer on the Beach, 4 (2019), 366–375.
[27] H. Roh, K. Lee, A basic performance evaluation of the speech recognition APP of standard language and dialect using Google Naver and DaumKAKAO APIs, Asia Pac. J. Multim. Serv. Converge. Art Human. Sociol. 7 (2017), 819–829.
[28] I. Bobriakov, Comparison of the top speech processing APIs, 2018, Available from: https://activewizards.com/blog/comparison-of-the-top-speech-processing-apis.
[29] O. Hyun-woo, L. Koen-Nyeong, Y. Dongsuk, Performance comparison of open APIs for speech recognition, Journal of the Acoustical Society of Korea 2019 Spring Conference(Jeju, Korea), 5 (2019), Volume 38, No 1(s), P256.
[30] J. Lee, Lecture on Korean Phonology, Samkyung Munhwa Sa, Seoul Gangbukgu Miadong, Korea, 2014.
[31] J.h. Lee, G.h. Lee, S.j. Kim, Korean Pronouncing Dictionary, Jigu Publishing Co., Gyoha-eup, Paju-si, Gyeonggi-do, Korea, 2008.
[32] J. Laver, Principles of Phonetics, Cambridge University Press, New York, 1994, p. 561.