US 20110010178A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2011/0010178 A1

(43) Pub. Date: Jan. 13, 2011

LEE et al.

(54) **SYSTEM AND METHOD FOR TRANSFORMING VERNACULAR PRONUNCIATION**

(75) Inventors: **Hyunjung LEE**, Seoul (KR); **Taeil Kim**, Seoul (KR); **Hee-Cheol Seo**, Seoul (KR); **Ji Hye Lee**, Seongnam-si (KR)

Correspondence Address:
H.C. PARK & ASSOCIATES, PLC
8500 LEESBURG PIKE, SUITE 7500
VIENNA, VA 22182 (US)

(73) Assignee: **NHN Corporation**, Seongnam-si (KR)

(21) Appl. No.: **12/831,607**

(22) Filed: **Jul. 7, 2010**

(30) **Foreign Application Priority Data**

Jul. 8, 2009 (KR) 10-2009-0062143

Publication Classification

(51) **Int. Cl.**
G10L 13/08 (2006.01)

(52) **U.S. Cl.** **704/260**; 704/E13.011

(57) **ABSTRACT**

Provided is a system and method for transforming vernacular pronunciation with respect to Hanja using a statistical method. In a system for transforming vernacular pronunciation, a vernacular pronunciation extracting unit extracts a vernacular pronunciation with respect to a Hanja character string, a statistical data determining unit determines a statistical data with respect to the Hanja character string by using statistical data of features related to a Hanja-vernacular pronunciation transformation, and a vernacular pronunciation transforming unit transforms the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

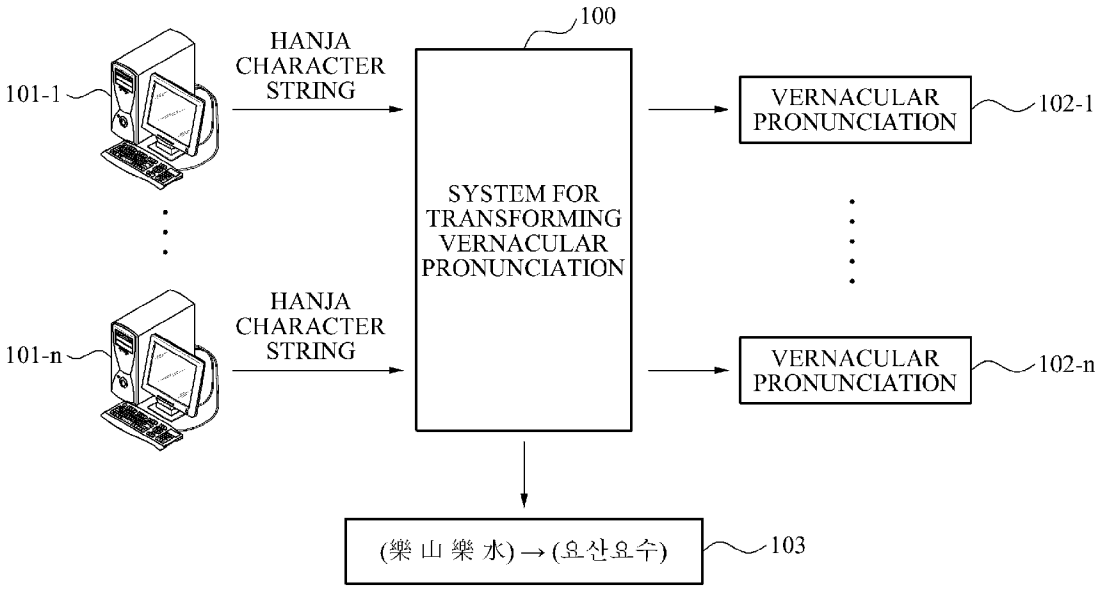


FIG. 1

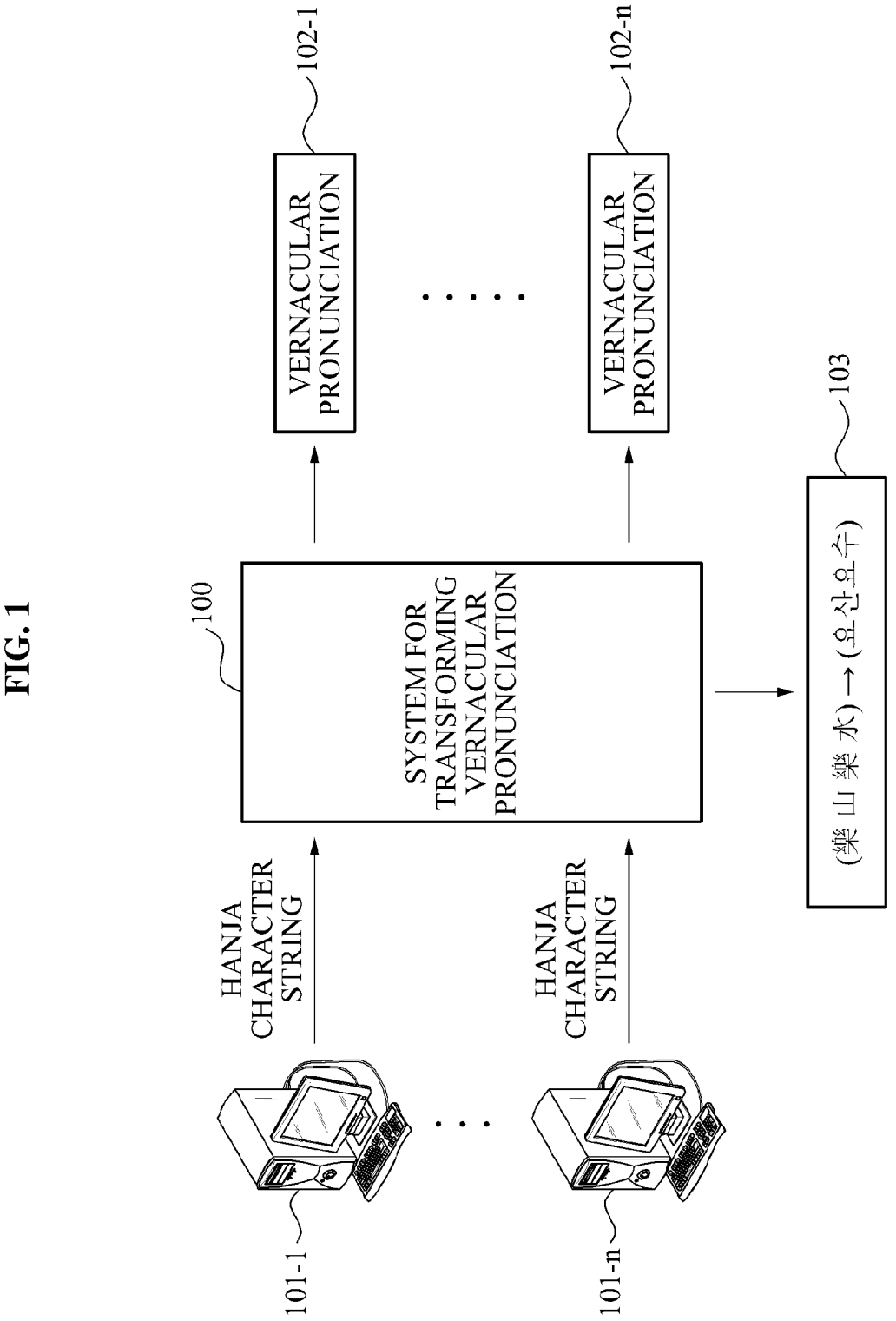

101-1
101-n
HANJA CHARACTER STRING
HANJA CHARACTER STRING
100
SYSTEM FOR TRANSFORMING VERNACULAR PRONUNCIATION
VERNACULAR PRONUNCIATION
VERNACULAR PRONUNCIATION
102-1
102-n
103
(樂 山 樂 水) → (요산요수)

FIG. 2

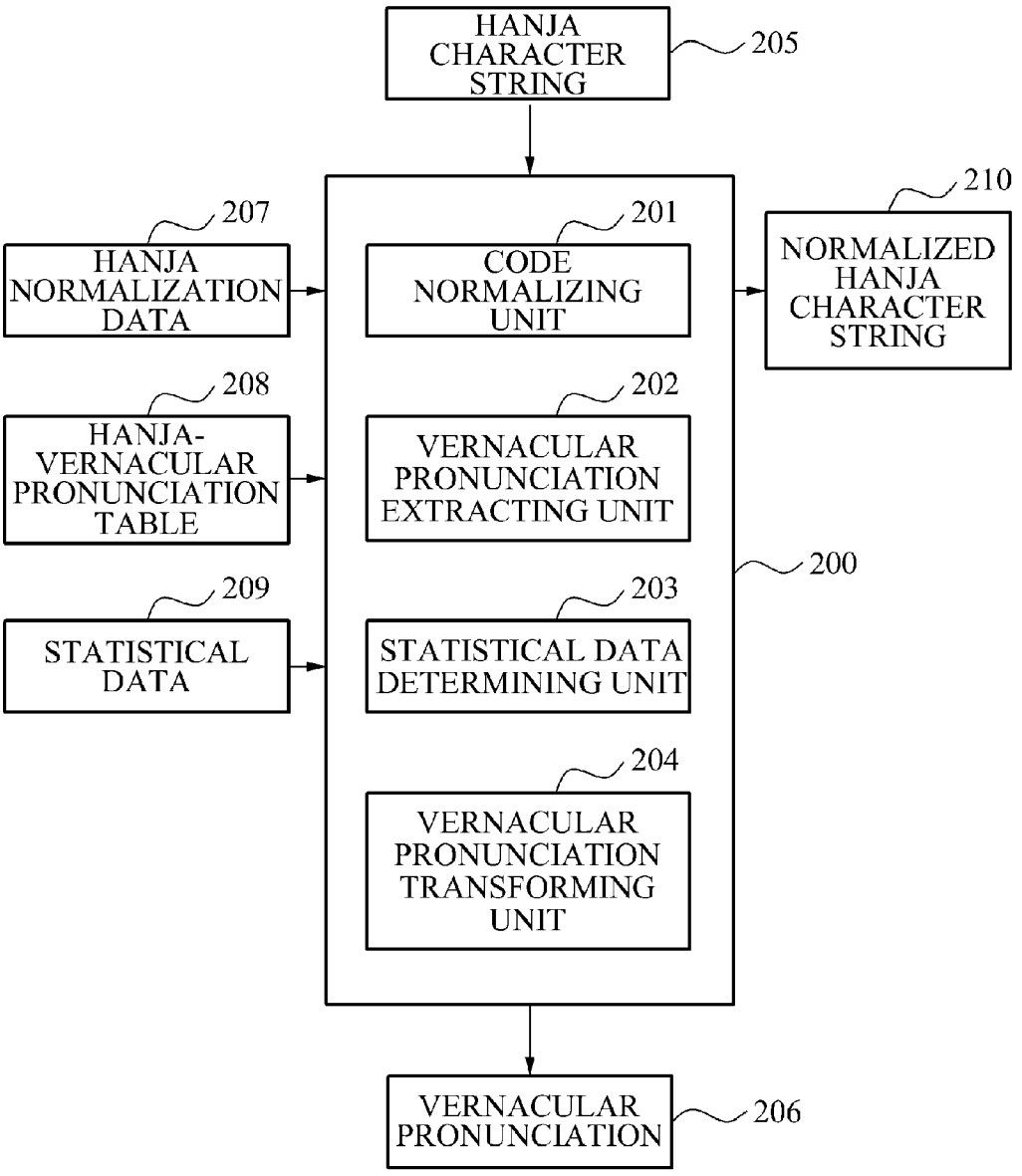

HANJA CHARACTER STRING
205
210
207
201
HANJA NORMALIZATION DATA
CODE NORMALIZING UNIT
NORMALIZED HANJA CHARACTER STRING
208
202
HANJA-VERNACULAR PRONUNCIATION TABLE
VERNACULAR PRONUNCIATION EXTRACTING UNIT
200
209
203
STATISTICAL DATA
STATISTICAL DATA DETERMINING UNIT
204
VERNACULAR PRONUNCIATION TRANSFORMING UNIT
VERNACULAR PRONUNCIATION
206

FIG. 3

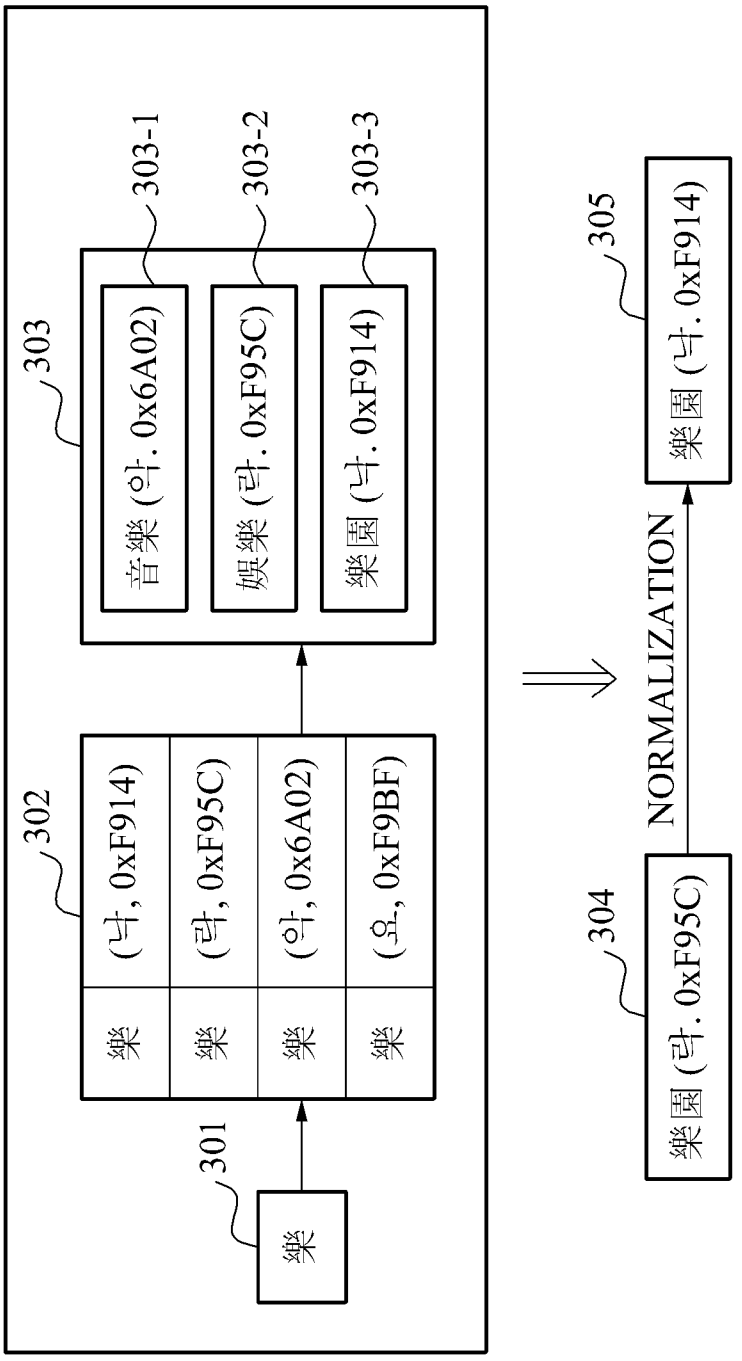

303
303-1
303-2
303-3
音樂 (악. 0x6A02)
娛樂 (락. 0xF95C)
樂園 (낙. 0xF914)
302
樂 (낙, 0xF914)
樂 (락, 0xF95C)
樂 (악, 0x6A02)
樂 (요., 0xF9BF)
301
樂
NORMALIZATION
305
樂園 (낙. 0xF914)
304
樂園 (락. 0xF95C)

FIG. 4

< HANJA-HANGUL PRONUNCIATION TABLE >

HANJA CHARACTER	HANGUL PRONUNCIATION
. . .	. . . . .
樂	낙 락 악 요
諾	낙 락
丹	난 단 란
寧	녕 령 영
怒	노 로
率	률 솔 수 율
. . .	. . . . .

FIG. 5

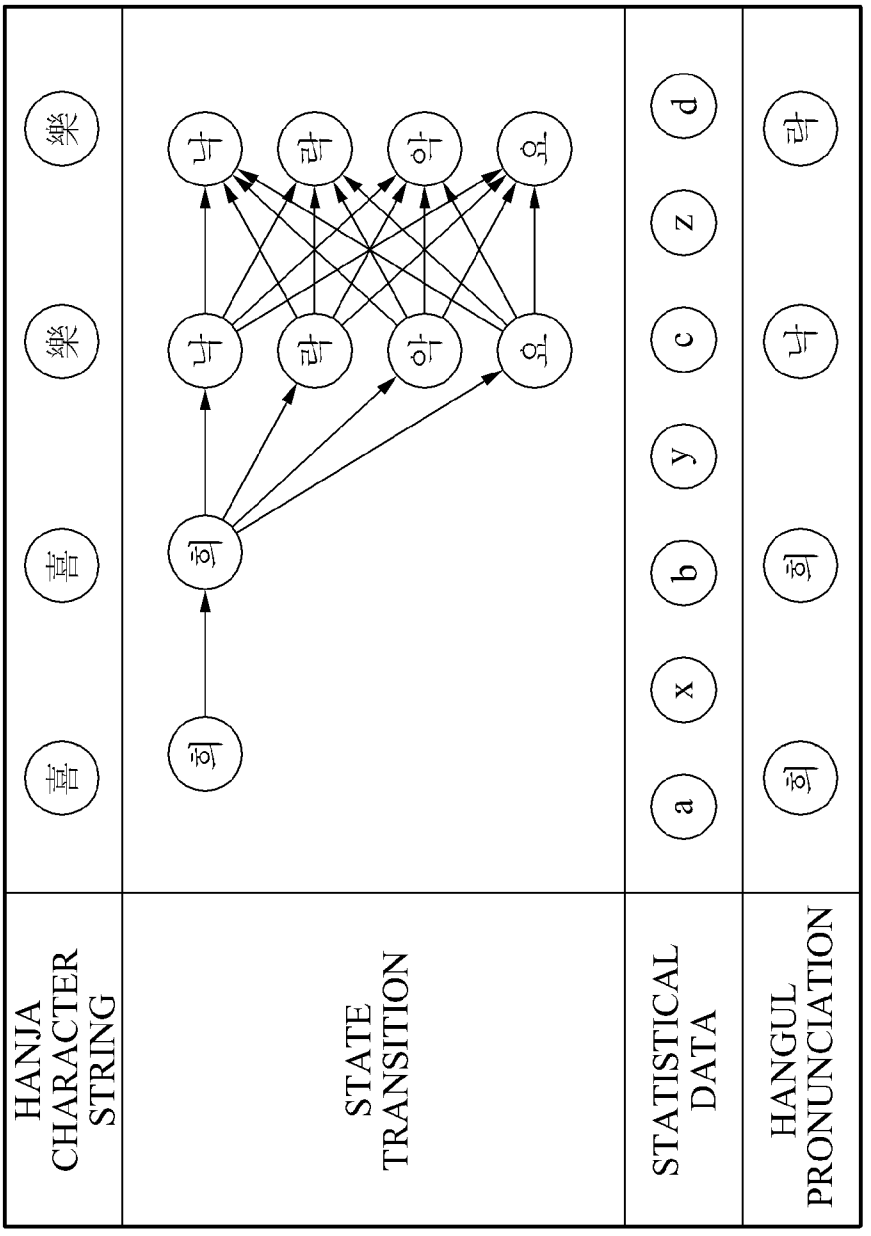
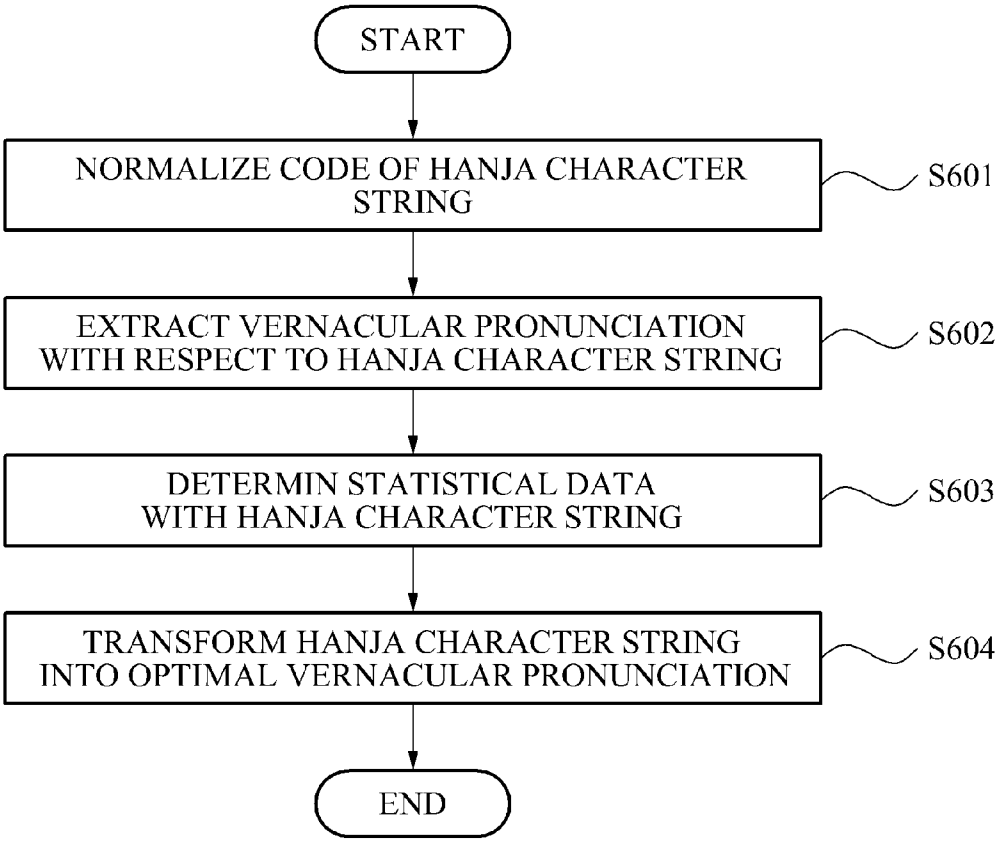

HANJA CHARACTER STRING
喜
喜
樂
樂
STATE TRANSITION
히
히
나
나
라
라
아
아
요
요
STATISTICAL DATA
a
x
b
y
c
z
d
HANGUL PRONUNCIATION
히
히
나
라

FIG. 6


START
NORMALIZE CODE OF HANJA CHARACTER STRING
S601
EXTRACT VERNACULAR PRONUNCIATION WITH RESPECT TO HANJA CHARACTER STRING
S602
DETERMIN STATISTICAL DATA WITH HANJA CHARACTER STRING
S603
TRANSFORM HANJA CHARACTER STRING INTO OPTIMAL VERNACULAR PRONUNCIATION
S604
END

SYSTEM AND METHOD FOR TRANSFORMING VERNACULAR PRONUNCIATION

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority from and the benefit of Korean Patent Application No. 10-2009-0062143, filed on Jul. 8, 2009, which is hereby incorporated by reference for all purposes as if fully set forth herein.

BACKGROUND

[0002] 1. Field

[0003] Exemplary embodiments of the present invention relate to a system and method for transforming vernacular pronunciation with respect to Hanja using a statistical method.

[0004] 2. Discussion of the Background

[0005] Hanja (Chinese characters) is used in various documents in Asian countries. In addition, Hanja is used in countries, such as the USA, that do not belong to the Hanja cultural is area. Particularly, text documents including Hanja are frequently used in programs using computers. However, some cases occur in which, for some users unfamiliar with Hanja, Hanja is transformed into a vernacular pronunciation in a word-processor program.

[0006] For example, in Korea, newspapers, legal documents, and the like in the past were frequently only written in Hanja. However, when Koreans search old newspapers or legal documents, they frequently search for Hanja by inputting the Hangul (Korean characters) pronunciation of the Hanja, as opposed to inputting the Hanja itself. As an example, '音樂' is searched for by inputting '음악' as a query.

[0007] In Japan, Hanja appears more frequently in documents as compared to Korea. However, the Japanese frequently search for Hanja by inputting Yomigana in place of the Hanja. As an example, '音樂' is searched for by inputting 'おんがく' as a query.

[0008] In China, Hanja appears more frequently in documents as compared with other Asian countries. Therefore, the Chinese frequently search for Hanja by inputting the Hanja itself as a query. However, some Chinese search for Hanja by inputting Pinyin in a query. As an example, '可口可乐' is searched for by inputting 'kekoukele' in a query.

[0009] In English-speaking countries, such as the USA, Hanja may be used in documents. However, a document can be easily searched for by transforming the Hanja used in the corresponding document into English and inputting the English as queries.

[0010] A related method for transforming the Hanja into vernacular pronunciation is performed using a conversion table. Vernacular corresponding to specific Hanja characters are stored in the conversion table. Then, if a Hanja character is inputted by a user, a vernacular corresponding to the Hanja character is presented.

[0011] Particularly, users may write documents or input search queries and not recognize is that heteronymous Hanja characters exist and that code values individually exist for each of the heteronymous Hanja character. The heteronymous Hanja character refers to a Hanja character with two or more pronunciations, for example, a Hanja character such as '樂' with Hangul pronunciations of '낙,' '락,' '악,' '요.' In Extended Unix Code—Korean (EUC-KR) or UNICODE, code values are individually determined for each of the heteronymous Hanja characters. Specifically, in UNICODE, four different code values, i.e., 樂(낙, 0xF914), 樂(락, 0xF95C), 樂(악, 0x6A02), and 樂(요, 0xF9BF), are provided for the Hanja character '樂.'

[0012] Accordingly, when the number of vernacular pronunciations capable of being transformed with respect to a Hanja character is at least one, the number of finally transformed vernacular pronunciations is also at least one. Therefore, it is necessary to reflect the user's original intention and derive the vernacular pronunciation suitable for the context and vernacular orthography.

[0013] Since multiple Hanja characters, each having various code values with respect to documents or queries, exist due to the heteronymous Hanja characters, there may occur a case in which all the documents or queries for a heteronymous Hanja character are not found. For example, if four documents are written respectively with 樂園(樂=0xF95C), 樂園(樂=0xF914), 樂園(樂=0x6A02) and 樂園(樂=0xF9BF), and, if a user searches documents by inputting 樂園 corresponding to 0xF95C, only one of the four documents may be found.

[0014] In Korea, if a Hanja character is transformed into a Hangul pronunciation without considering the Hangul orthography, such as the context and acrophony, an unintended result may be retrieved. For example, there may occur a case in which Hanja characters, such as '來日, are transformed into '래일' rather than '내일.' Since each country has a unique orthography, transformation of the Hanja into vernacular pronunciation in consideration of the orthography may be desired. Accordingly, more accurate transformation of the Hanja into vernacular pronunciation may be desired.

SUMMARY

[0015] Exemplary embodiments of the present invention provide a method and system in which a Hanja character string is transformed into a vernacular pronunciation using statistical data of features related to the Hanja-vernacular pronunciation transformation, thereby enhancing the accuracy of the finally derived vernacular pronunciation.

[0016] Exemplary embodiments of the present invention also provide a system and a method for transformation of a heteronymous Hanja character into a vernacular pronunciation suitable for the context and vernacular orthography by using statistical data.

[0017] Exemplary embodiments of the present invention also provide a system and a method for transformation of an accurate vernacular pronunciation even if a Hanja character string with an inaccurate code is inputted through a Hanja code normalization.

[0018] Exemplary embodiments of the present invention also provide a system and a method for enhancement of reliability of a vernacular pronunciation transformed with respect to a Hanja character string by accurately reflecting exceptional grammar, such as acrophony of Hangul, using statistical data.

[0019] Additional features of the invention will be set forth in the description which follows, and in part will be apparent from the description, or may be learned by practice of the invention.

[0020] An exemplary embodiment of the present invention discloses a system for transforming vernacular pronunciation, the system including a vernacular pronunciation extracting unit to extract a vernacular pronunciation with respect to

a Hanja character string, a statistical data determining unit to determine statistical data with respect to the Hanja character string by using statistical data of features related to a Hanja-vernacular pronunciation transformation, and a vernacular pronunciation transforming unit to transform the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

[0021] An exemplary embodiment of the present invention discloses a method for transforming vernacular pronunciation, the method including extracting a vernacular pronunciation with respect to a Hanja character string; determining statistical data with respect to the Hanja character string by using statistical data of features related to a Hanja-vernacular pronunciation transformation; and transforming the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

[0022] An exemplary embodiment of the present invention discloses a method for transforming vernacular pronunciation, the method including extracting a vernacular pronunciation with respect to a character string; determining statistical data with respect to the character string by using statistical data of features related to a language-vernacular pronunciation transformation; and transforming the character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

[0023] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory and are intended to provide further explanation of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] The accompanying drawings, which are included to provide a further understanding of the invention and are incorporated in and constitute a part of this specification, illustrate embodiments of the invention, and together with the description serve to explain the principles of the invention.

[0025] FIG. **1** is a diagram illustrating a process of transforming a vernacular pronunciation with respect to a Hanja character string through a system for transforming vernacular pronunciation according to an exemplary embodiment of the present invention.

[0026] FIG. **2** is a block diagram illustrating a system for transforming vernacular pronunciation according to an exemplary embodiment of the present invention.

[0027] FIG. **3** is a diagram illustrating a process of normalizing a Hanja character string according to an exemplary embodiment of the present invention.

[0028] FIG. **4** is a diagram illustrating an example of a Hanja-vernacular pronunciation table according to an exemplary embodiment of the present invention.

[0029] FIG. **5** is a diagram illustrating a method for transforming a vernacular pronunciation with respect to a Hanja character string according to an exemplary embodiment of the present invention.

[0030] FIG. **6** is a flowchart illustrating a method for transforming vernacular pronunciation according to an exemplary embodiment of the present invention.

DETAILED DESCRIPTION OF THE ILLUSTRATED EMBODIMENTS

[0031] The invention is described more fully hereinafter with reference to the accompanying drawings, in which exemplary embodiments of the invention are shown. This invention may, however, be embodied in many different forms and should not be construed as is limited to the exemplary embodiments set forth herein. Rather, these exemplary embodiments are provided so that this disclosure is thorough, and will convey the scope of the invention to those skilled in the art. In the drawings, the size and relative sizes of layers and regions may be exaggerated for clarity. Like reference numerals in the drawings denote like elements.

[0032] A method for transforming vernacular pronunciation may be performed by a system for transforming vernacular pronunciation.

[0033] FIG. **1** is a diagram illustrating a process of transforming a vernacular pronunciation with respect to a Hanja character string through a system **100** for transforming vernacular pronunciation according to an exemplary embodiment of the present invention. Although described herein with respect to Hanja, aspects of the present invention are not limited thereto such that features described herein may be applied to other languages and characters.

[0034] If a user inputs via at least one of terminals **101-1** to **101-*n*** a Hanja character string including at least one Hanja character, the system **100** can transform the Hanja character string into a vernacular pronunciation **102-1** to **102-*n***. The vernacular may be differently determined based on the language written in a document provided by or to the system **100**. For example, if the system **100** provides or is provided with a Hangul document, the vernacular may be determined as Hangul.

[0035] In this case, the Hanja character string includes at least one Hanja character. Hanja characters included in a text document may be transformed into vernacular pronunciations in a program (a program for a personal computer (PC), a program for a server, a program for the Internet, and the like) using a computer. For example, if a user inputs '情報檢索' as a Hanja character string, the system **100** may transform the Hanja character string into '정보검색' that is a vernacular pronunciation **102-1** to **102-*n***. If the user inputs a Hanja character string as a search query, the amount of search results is relatively small when the Hanja character string inputted to a search engine is searched for as is. Hence, the system **100** transforms the Hanja character string into the vernacular pronunciation **102-1** to **102-*n*** so that the search engine can derive more appropriate search results.

[0036] If a Hanja character string is included in a text document, the system **100** transcribes a vernacular pronunciation **102-1** to **102-*n*** with respect to the Hanja character string at the point at which the corresponding Hanja character string is positioned so that the user can more conveniently read the text document. As can be seen in a transformation example **103** of FIG. **1**, if a Hanja character string, i.e., '樂山樂水' is included in the text document, the system **100** may transform the Hanja character string into a Hangul pronunciation, i.e., '요산요수.'

[0037] The system **100** uses data obtained by statistically analyzing the data transformed into a vernacular pronunciation with respect to a given Hanja character string, thereby providing a more accurate vernacular pronunciation. Also, the system **100** provides vernacular pronunciation suitable for the context and vernacular orthography, thereby providing a more accurate vernacular pronunciation.

[0038] FIG. **2** is a block diagram illustrating a system for transforming vernacular pronunciation according to an exemplary embodiment of the present invention. Referring to FIG. **2**, the system **200** may include a code normalizing unit **201**, a

vernacular pronunciation extracting unit **202**, a statistical data determining unit **203**, and a vernacular pronunciation transforming unit **204**.

[0039] The code normalizing unit **201** normalizes the code of a Hanja character string **205** including a heteronymous Hanja character having a same form and different codes. As an example, the code normalizing unit **201** may normalize the code of the Hanja character string **205** by transforming the heteronymous Hanja character as a representative Hanja character. In this case, the code normalizing unit **201** may normalize the code of the Hanja character string **205** using Hanja normalization data **207**.

[0040] As a result, a normalized Hanja character string **210** normalized by the code normalizing unit **201** can be derived. However, if the Hanja character string **205** includes no heteronymous Hanja character, the code normalizing unit **201** may not operate. The operation of the code normalizing unit **201** will be described in detail with reference to FIG. **3**.

[0041] The vernacular pronunciation extracting unit **202** extracts a vernacular pronunciation with respect to a Hanja character string using a Hanja-vernacular pronunciation table **208**. The Hanja-vernacular pronunciation table **208** may include pairs or multiples of vernacular pronunciations for respective Hanja characters. That is, according to the Hanja-vernacular pronunciation table **208**, a vernacular pronunciation may correspond to each of the Hanja characters.

[0042] However, if one or more pronunciations correspond to the same Hanja character, the vernacular pronunciation may be transformed to be suitable for the context and vernacular orthography. Accordingly, the system **200** can enhance the accuracy of the vernacular pronunciation transformed using statistical data transformed into the vernacular from the Hanja.

[0043] The statistical data determining unit **203** determines statistical data with respect to a Hanja character string using the statistical data of features related to the Hanja-vernacular pronunciation transformation **208**.

[0044] As an example, the statistical data determining unit **203** may determine statistical data with respect to the Hanja character string **205** using statistical data **209** that is extracted from data in which the Hanja and vernacular are represented together and corresponds to meaningful features with respect to the Hanja-vernacular transformation. The statistical data determining unit **203** may determine the syllable probability and transition probability with respect to syllables of a vernacular pronunciation **206** related to the Hanja character string **205**.

[0045] That is, the statistical data determining unit **203** may more accurately determine the vernacular differently pronounced with respect to the same Hanja character depending on conditions by using various statistical data transformed into the vernacular with respect to the Hanja. The process of using statistical data will be further described with reference to FIG. **5**.

[0046] The vernacular pronunciation transforming unit **204** transforms the Hanja character string **205** into an optimal vernacular pronunciation **206** using the extracted vernacular pronunciation and the determined statistical data. As an example, the vernacular pronunciation transforming unit **204** may determine a vernacular pronunciation **206** having a maximum probability of the vernacular pronunciation to be transformed with respect to the Hanja character string **205**.

[0047] In this case, the vernacular pronunciation transforming unit **204** may transform the Hanja character string **205** into the vernacular pronunciation **206** based on a Hidden Markov Model, but aspects are not limited thereto such that other models may be used. The vernacular pronunciation transforming unit **204** may transform the Hanja character string **205** into the vernacular pronunciation **206** having an optimal path with respect to the Hanja character string **205** by applying a Viterbi algorithm to Hanja character strings that are repeatedly processed.

[0048] FIG. **3** is a diagram illustrating a process of normalizing a Hanja character string according to an exemplary embodiment of the present invention.

[0049] Although a Hanja character string may not be transformed into a vernacular pronunciation, words each having various code values exist in documents or queries due to heteronymous Hanja characters. Hence, a search may not be performed. Therefore, the code of a Hanja character string including a heteronymous Hanja character with the same form and different codes may be normalized.

[0050] For example, a Hanja list of four different codes with the same form and different Hangul pronunciations may be derived from '樂' **301**. If the '樂' **301** is inputted as 樂(요, 0xF9BF) **302**, a search result **303** including 音樂(악, 0x6A02) **303-1**, 娛樂(락, 0xF95C) **303-2** and 樂園(낙, 0xF914) **303-3** may not be retrieved. Therefore, the system for transforming vernacular pronunciation may perform normalization with respect to a Hanja character string including a heteronymous Hanja character.

[0051] Vernacular pronunciations with respect to a heteronymous Hanja character may be differently defined for different countries, regions, and/or populations. For example, the '樂' may be pronounced as '낙', '락', '악', or '요' in Hangul. However, '樂' the may be pronounced as 'がく(音樂, おんがく)' or 'らく(らくょう)' in Japanese. In addition, '乐' may be pronounced as 'yue' or 'le' in Chinese.

[0052] As an example, the system may normalize the code of a Hanja character string by transforming a heteronymous Hanja character into a representative Hanja character. In this case, the system may normalize the code of a Hanja character string using a normalization data built through a Hanja dictionary. That is, although a user inputs 樂園(락, 0xF95C) **304**, the system may normalize the '樂' that is a heteronymous Hanja character and transformed as a representative Hanja character. Then, the system may derive a normalized Hanja character string **305**.

[0053] The system may solve the problem of data scarcity in a statistical model through the normalization process of a Hanja character string. Also, the system may transform a vernacular pronunciation with a Hanja character used with a code unsuitable for the context and vernacular orthography.

[0054] FIG. **4** is a diagram illustrating an example of a Hanja-vernacular pronunciation table according to an exemplary embodiment of the present invention. Particularly, FIG. **4** illustrates an example of a Hanja-Hangul pronunciation table. The description of FIG. **4** may be analogically applied for other pronunciations, languages, and/or characters.

[0055] The Hanja-Hangul pronunciation table may include pairs or multiples of vernacular pronunciations for respective Hanja characters. Particularly, the Hanja-Hangul pronunciation table may be applied to a case in which one Hanja character has a plurality of Hangul pronunciations. As can be seen in FIG. **4**, '樂' the may be pronounced as '낙', '락', '악', and '요' in Hangul.

[0056] For example, if a Hanja character '寧' is included in a Hanja character string inputted by a user, the system for

transforming vernacular pronunciation may extract Hangul pronunciations '녕', '령', and '영' with respect to the Hanja character '寧' using the Hanja-Hangul pronunciation table.

[0057] A Hanja-Japanese pronunciation table may include Japanese pronunciations 'がく and 'らく' with respect to the Hanja character '樂.' In addition, a Hanja-Chinese pronunciation table may include Chinese pronunciations (Pinyin) 'yue' and 'le' with respect to the Hanja character '乐.'

[0058] FIG. **5** is a diagram illustrating a method for transforming a vernacular pronunciation with respect to a Hanja character string according to an exemplary embodiment of the present invention. Referring to FIG. **5**, it is assumed that a Hanja character string '嬉嬉樂樂' is inputted. The system for transforming vernacular pronunciation may transform vernacular pronunciations with respect to characters constituting the Hanja character string by using a Hanja-vernacular pronunciation table. As an example, '喜' may be transformed into '희,' and '樂' may be transformed into '낙', '락', '악', and '요.'

[0059] The system may determine statistical data with respect to a Hanja character string using the statistical data of features related to the Hanja-vernacular pronunciation transformation. As an example, the system may determine statistical data with respect to the Hanja character string using statistical data that is extracted from data in which the Hanja and vernacular are represented together and corresponds to features with respect to the Hanja-vernacular transformation.

[0060] The features may be varied depending on grammar and orthography. The features with respect to the Hanja-Hangul transformation may include the following probabilities:

[0061] Probability that a current Hangul pronunciation appears together with a current Hanja character (e.g., probability that '樂' is transformed into '요')

[0062] Probability that a current Hangul pronunciation appears together with a previous Hangul pronunciation (e.g., probability that '요' appears before '산')

[0063] Probability that a current Hanja character appears together with a previous Hangul pronunciation (e.g., probability that '요' appears before '山')

[0064] Probability that a current Hangul pronunciation appears together with a Hangul pronunciation before the previous Hangul pronunciation (e.g., probability that '요' appears before another '요' with a Hangul pronunciation interposed therebetween)

[0065] Probability that a current Hanja character appears together with a Hangul pronunciation before the previous Hangul pronunciation (e.g., probability that '요' appears before '樂' with a Hangul pronunciation interposed therebetween)

[0066] Probability that if a current Hanja character is '不' and the following Hanja pronunciation is starting with 'ㅈ' or 'ㄷ', the '不' is pronounced as '부'

[0067] Probability that if a current Hanja character is '來' and its current position is placed at the head of a word, the '來' is pronounced as '내' (acrophony)

[0068] Probability that when a current Hanja character is '來' and its current position is placed at the end of a word, the '來' is pronounced as '래'

[0069] The probability for the aforementioned features may be statistically determined using data from blogs, documents, web pages, and the like, in which the vernacular and Hanja are represented together. Particularly, various acrophonies exist in Hangul pronunciations, and many exceptions for the acrophonies also exist. Hence, it is possible to enhance the accuracy of Hangul pronunciations transformed using a statistical data that is extracted from data in which the Hanja and vernacular are represented together and corresponds to features with respect to the Hanja-vernacular transformation. Since unique orthographies exist in other countries, regions, and populations, like the Korean acrophony, statistical data suitable for conditions of each country, region, or population may be derived using features that reflect the unique orthographies.

[0070] The below details may be used as features applied to the statistical data. As an example, the acrophonies for Hangul pronunciations and their exceptions are as follows:

[0071] If a Hangul pronunciation having an initial sound of 'ㄴ' appears at the beginning of a word, the 'ㄴ' is pronounced as 'ㅇ' (e.g., 여자 (女子), 연세 (年歲), 요소 (尿素), 익명 (匿名), ...)

[0072] If a Hangul pronunciation having an initial sound of 'ㄹ' appears at the beginning of a word, the 'ㄹ' is pronounced as 'ㅇ' (e.g., 양심 (良心), 역사 (歷史), 예의 (禮儀), 용궁 (龍宮), 유행 (流行), . . .)

[0073] If a Hangul pronunciation having an initial sound of 'ㄹ' appears at the beginning of a word, the 'ㄹ' is pronounced as 'ㄴ' (e.g., 낙원 (樂園), 내일 (來日), 노인 (老人), 뇌성 (雷聲), 누각 (樓閣), . . .)

[0074] Acrophony exists in derivative words and compound words (the boundary between words exists in a word phrase) (e.g., 落花流水 (낙화유수), 修學旅行 (수학여행), 新女性 (신여성), . . .)

[0075] Exceptions of the acrophony (e.g., 구름양 (量)/노동량 (量), 운율 (律)/법률 (律), 진열 (列)/행렬 (列), 의논 (論)/토론 (論), . . .)

[0076] The system may determine statistical data with respect to a Hanja character string. As an example, the system may calculate the syllable probability and transition probability with respect to syllables of a vernacular pronunciation related to a Hanja character string, thereby determining the statistical data with respect to the Hanja character string. Referring to FIG. **5**, '희,' '희,' '낙', '락', '악' and '요,' and '낙', '락', '악', and '요,' transformed into Hangul pronunciations with respect to a Hanja character string '嬉嬉樂樂' may be configured as respective states.

[0077] In this case, the probability that a Hanja character corresponding to any one syllable in the Hanja character string is transformed into a vernacular pronunciation may be defined as a syllable probability. For example, the probability that a Hanja character '喜' is transformed into a Hangul pronunciation '희' may be defined as a syllable probability with respect to the Hanja character '喜.' In addition, the probability that a Hanja character '樂' is transformed into a Hangul pronunciation '낙' may be defined as a syllable probability with respect to the Hanja character '樂.' In FIG. **5**, the syllable probabilities that are statistical data determined with respect to the Hanja character string may be determined as "a," "b," "c" and "d," respectively.

[0078] The probability the vernacular pronunciation of a next Hanja character appears with respect to the vernacular pronunciation of a specific Hanja character in the transition of a state may be defined as a transition probability. For example, the probability that the Hangul pronunciation of a Hanja character '喜' is '희' and the Hangul pronunciation of another Hanja character '喜' that appears after the former Hanja character '喜'is '희' may be defined as the transition probability of the another Hanja character '喜.' Also, the probability that the Hangul pronunciation of a Hanja charac-

ter '喜' is '희' and the Hangul pronunciation of a Hanja character '樂' that appears after the other Hanja character '喜' is '악' may be defined as the transition probability of the Hanja character '樂' that appears after the another Hanja character '喜'. In FIG. **5**, the transition probabilities that are statistical data determined with respect to the Hanja character string may be determined as "x," "y" and "z," respectively.

[0079] The system may transform the Hanja character string into the optimal vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data. As an example, the system may determine a vernacular pronunciation with the maximum probability that the Hanja character string is transformed into a desired vernacular pronunciation using the syllable probability and transition probability, which are statistical data. In this case, the system may transform a Hanja character string into a vernacular pronunciation using a Hidden Markov Model.

[0080] For Korean, the Hanja character string may be transformed into a Hangul pronunciation. For Japanese, the Hanja character string may be transformed into a Yomigana (よみがな) or (ふりがな) pronunciation. For Chinese, the Hanja character string may be transformed into a Pinyin pronunciation. In this case, the Pinyin may be obtained by transcribing Chinese pronunciations into Roman characters.

[0081] In the case of English-speaking countries, regions, and/or populations such as the USA and the UK, the Hanja character string may be transformed into Romaji (transcription of Japanese into Roman characters) or Pinyin (transcription of Chinese into Roman characters). For example, 'I like 壽司' may be transformed into 'I like sushi ' as the transcription in Roman characters. As another example, '劉備 visited' may be transformed into 'Liu Bei visited' as the transcription in Pinyin.

[0082] As an example, the system may transform a vernacular pronunciation with respect to a Hanja character string using a Hidden Markov Model according to the following Expression 1.

$$\Gamma(C) = \underset{K}{\operatorname{argmax}} P\langle K \mid C\rangle = \underset{K}{\operatorname{argmax}} P(K, C) \quad \text{[Expression 1]}$$

$$\begin{aligned} P(K, C) &= P(k_{1,n}, c_{1,n}) \\ &= P(c_1)\cdot P\langle k_1 \mid c_1\rangle \cdot P\langle c_2 \mid c_1, k_1\rangle \cdot P\langle k_2 \mid c_{1,2}, k_1\rangle \cdot \\ &\quad P\langle c_3 \mid c_{1,2}, k_{1,2}\rangle \cdot P\langle k_3 \mid c_{1,3}, k_{1,2}\rangle \Lambda \cdot \\ &\quad P\langle c_n \mid c_{1,n-1}, k_{1,n-1}\rangle \cdot P\langle k_n \mid c_{1,n}, k_{1,n-1}\rangle \\ &\approx \prod_{i=1}^{n} P\langle c_i \mid c_{i-M,i-1}, k_{i-J,i-1}\rangle \cdot P\langle k_i \mid c_{i-L,i}, k_{i-I,i-1}\rangle \end{aligned}$$

[0083] In this case, C denotes a Hanja character string, and K denotes a vernacular pronunciation. Also,

$$\prod_{i=1}^{n} P\langle c_i \mid c_{i-M,i-1}, k_{i-J,i-1}\rangle$$

is a syllable probability, and $P\langle k_i|c_{i-L,i},k_{i-I,i-1}\rangle$ is a transition probability.

[0084] Then, the vernacular pronunciation finally transformed with respect to the Hanja character string may be determined according to the following Expression 2.

$$\underset{k_{1,n}}{\operatorname{argmax}} \prod_{i=1}^{n} P\langle c_i \mid c_{i-2,i-1}, k_{i-2,i-1}\rangle \cdot P\langle k_i \mid c_{i-1,i}, k_{i-2,i-1}\rangle \quad \text{[Expression 2]}$$

[0085] That is, the system may determine a vernacular pronunciation with the maximum combination of the syllable probability and transition probability with respect to a given Hanja character string. In this case, the system may transform the Hanja character string into the vernacular pronunciation having the optimal path with respect to the Hanja character string by applying the Viterbi algorithm to Hanja character strings that are repeatedly processed.

[0086] Through the above described processes, the vernacular pronunciation with respect to the Hanja character string '囍嬉樂樂' can be determined as '희희낙락' as shown in FIG. **5**.

[0087] FIG. **6** is a flowchart illustrating a method for transforming vernacular pronunciation according to an exemplary embodiment of the present invention. The system for transforming vernacular pronunciation may normalize the code of a Hanja character string in operation **S601**. As an example, the system may normalize the code of a Hanja character string including a heteronymous Hanja character with a same form and different codes. In this case, the system may normalize the code of the Hanja character string by transforming the heteronymous Hanja character to a representative Hanja character using normalization data. Here, the normalization data may be built from a dictionary.

[0088] The system may extract a vernacular pronunciation with respect to the Hanja character string in operation **S602**. As an example, the system may extract the vernacular pronunciation with respect to the Hanja character string by using a Hanja-vernacular pronunciation table that includes pairs or multiples of vernacular pronunciations for respective Hanja characters. In this case, when the Hanja character string passes through the normalization process, the system can extract the vernacular pronunciation with respect to the normalized Hanja character string.

[0089] The system may determine a statistical data with respect to the Hanja character string by using statistical data of features related to the Hanja-vernacular pronunciation transformation in operation **S603**. As an example, the system may determine statistical data with respect to the Hanja character string using statistical data that is extracted from data in which the Hanja and vernacular are represented together and corresponds to features with respect to the Hanja-vernacular transformation. In this case, the system may determine the syllable probability and transition probability with respect to syllables of the vernacular pronunciation related to the Hanja character string.

[0090] The system may transform the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data in operation **S604**. As an example, the system may determine a vernacular pronunciation with a maximum probability of the vernacular pronunciation to be transformed with respect to the Hanja character string.

[0091] In this case, the system may transform the Hanja character string into the vernacular pronunciation based on a Hidden Markov Model. Particularly, the system may transform the Hanja character string into the vernacular pronunciation having an optimal path with respect to the Hanja

character string by applying a Viterbi algorithm to Hanja character strings that are repeatedly processed.

[0092] Details that are not described in FIG. **6** may be understood by referring to the descriptions of FIGS. **1** to **5**.

[0093] The method according to an exemplary embodiment of the present invention may include non-transitory computer-readable media including program instructions to implement various operations embodied by a computer. The media may also include, alone or in combination with the program instructions, data files, data structures, and the like. The media and program instructions may be those specially designed and constructed for the purposes of the present invention, or they may be of the kind well-known and available to those having skill in the computer software arts. Examples of non-transitory computer-readable media include magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD ROM disks and DVD; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory (ROM), random access memory (RAM), flash memory, and the like. Examples of program instructions include both machine code, such as produced by a compiler, and files containing higher level code that may be executed by the computer using an interpreter. The described hardware devices may be configured to act as one or more software modules in order to perform the operations of the above-described embodiments of the present invention.

[0094] It will be apparent to those skilled in the art that various modifications and variation can be made in the present invention without departing from the spirit or scope of the invention. Thus, it is intended that the present invention cover the modifications and variations of this invention provided they come within the scope of the appended claims and their equivalents.

What is claimed is:

1. A system for transforming vernacular pronunciation, the system comprising:

a vernacular pronunciation extracting unit to extract a vernacular pronunciation with respect to a Hanja character string;

a statistical data determining unit to determine statistical data with respect to the Hanja character string by using statistical data of features related to a Hanja-vernacular pronunciation transformation; and

a vernacular pronunciation transforming unit to transform the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

2. The system of claim **1**, wherein the vernacular pronunciation extracting unit extracts the vernacular pronunciation using a Hanja-vernacular pronunciation table that includes vernacular pronunciations for respective Hanja characters.

3. The system of claim **1**, further comprising:

a code normalizing unit to normalize a code of the Hanja character string including a heteronymous Hanja character with a same form and different codes,

wherein the vernacular pronunciation extracting unit extracts the vernacular pronunciation with respect to the Hanja character string of which the code is normalized.

4. The system of claim **3**, wherein the code normalizing unit normalizes the code of the Hanja character string by transforming the heteronymous Hanja character into a representative Hanja character.

5. The system of claim **1**, wherein the statistical data determining unit determines the statistical data with respect to the Hanja character string by using statistical data extracted from data in which the Hanja and the vernacular are represented together and corresponds to features with respect to the Hanja-vernacular transformation.

6. The system of claim **1**, wherein the statistical data determining unit determines a syllable probability and a transition probability with respect to a syllable of the vernacular pronunciation related to the Hanja character string.

7. The system of claim **1**, wherein the vernacular pronunciation transforming unit determines the vernacular pronunciation having the maximum probability of the vernacular pronunciation to be transformed with respect to the Hanja character string.

8. The system of claim **7**, wherein the vernacular pronunciation transforming unit transforms the Hanja character string into the vernacular pronunciation based on a Hidden Markov Model.

9. The system of claim **8**, wherein the vernacular pronunciation transforming unit transforms the Hanja character string into the vernacular pronunciation having an optimal path with respect to the Hanja character string by applying a Viterbi algorithm to Hanja character strings that are repeatedly processed.

10. A method for transforming vernacular pronunciation, the method comprising:

extracting a vernacular pronunciation with respect to a Hanja character string;

determining statistical data with respect to the Hanja character string by using statistical data of features related to a Hanja-vernacular pronunciation transformation; and

transforming the Hanja character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

11. The method of claim **10**, wherein the extracting the vernacular pronunciation comprises:

extracting the vernacular pronunciation using a Hanja-vernacular pronunciation table that includes vernacular pronunciations for respective Hanja characters.

12. The method of claim **11**, further comprising normalizing a code of the Hanja character string including a heteronymous Hanja character with a same form and different codes,

wherein the extracting the vernacular pronunciation with respect to the Hanja character string comprises extracting the vernacular pronunciation with respect to the Hanja character string of which the code is normalized.

13. The method of claim **12**, wherein the normalizing the code of the Hanja character string comprises:

normalizing the code of the Hanja character string by transforming the heteronymous Hanja character into a representative Hanja character.

14. The method of claim **10**, wherein the determining the statistical data with respect to Hanja character string comprises:

determining the statistical data with respect to the Hanja character string by using statistical data extracted from data in which the Hanja and the vernacular are represented together and corresponds to features with respect to the Hanja-vernacular transformation.

15. The method of claim **10**, wherein the determining the statistical data with respect to Hanja character string comprises:

determining a syllable probability and a transition probability with respect to a syllable of the vernacular pronunciation related to the Hanja character string.

16. The method of claim **10**, wherein the transforming the Hanja character string into the vernacular pronunciation comprises:

determining a vernacular pronunciation having the maximum probability of the vernacular pronunciation to be transformed with respect to the Hanja character string.

17. The method of claim **16**, wherein the transforming the Hanja character string into the optimal vernacular pronunciation comprises:

transforming the Hanja character string into the vernacular pronunciation based on a Hidden Markov Model.

18. The method of claim **17**, wherein the transforming the Hanja character string into the optimal vernacular pronunciation comprises:

transforming the Hanja character string into the vernacular pronunciation having an optimal path with respect to the Hanja character string by applying a Viterbi algorithm to Hanja character strings that are repeatedly processed.

19. A non-transitory computer-readable medium in which a program for performing the method of claim **10** is recorded.

20. A method for transforming vernacular pronunciation, the method comprising:

extracting a vernacular pronunciation with respect to a character string;

determining statistical data with respect to the character string by using statistical data of features related to a language-vernacular pronunciation transformation; and

transforming the character string into a vernacular pronunciation using the extracted vernacular pronunciation and the determined statistical data.

* * * * *