Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2914677 A1 2014/12/11

(21) 2 914 677

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) **Date de dépôt PCT/PCT Filing Date:** 2014/06/04
(87) **Date publication PCT/PCT Publication Date:** 2014/12/11
(85) **Entrée phase nationale/National Entry:** 2015/12/04
(86) **N° demande PCT/PCT Application No.:** US 2014/040906
(87) **N° publication PCT/PCT Publication No.:** 2014/197592
(30) **Priorité/Priority:** 2013/06/04 (US61/830,789)

(51) **Cl.Int./Int.Cl.** ***G10L 15/19*** (2013.01), ***G10L 13/02*** (2013.01), ***G10L 15/26*** (2006.01)

(71) **Demandeur/Applicant:**
IMS SOLUTIONS INC., US

(72) **Inventeurs/Inventors:**
CAMPBELL, DAVID NEIL, CA;
RAE, ROBERT ANDREW, CA;
ELGHAZAL, AKREM SAAD, CA;
SULPIZI, DANIEL JOHN VINCENT, CA

(74) **Agent:** GOWLING LAFLEUR HENDERSON LLP

(54) **Titre : INTERFACE HOMME-MACHINE AMELIOREE PAR LA RECONNAISSANCE DE MOTS HYBRIDE ET L'ADAPTATION DYNAMIQUE DE LA SYNTHESE DE LA PAROLE**

(54) **Title: ENHANCED HUMAN MACHINE INTERFACE THROUGH HYBRID WORD RECOGNITION AND DYNAMIC SPEECH SYNTHESIS TUNING**

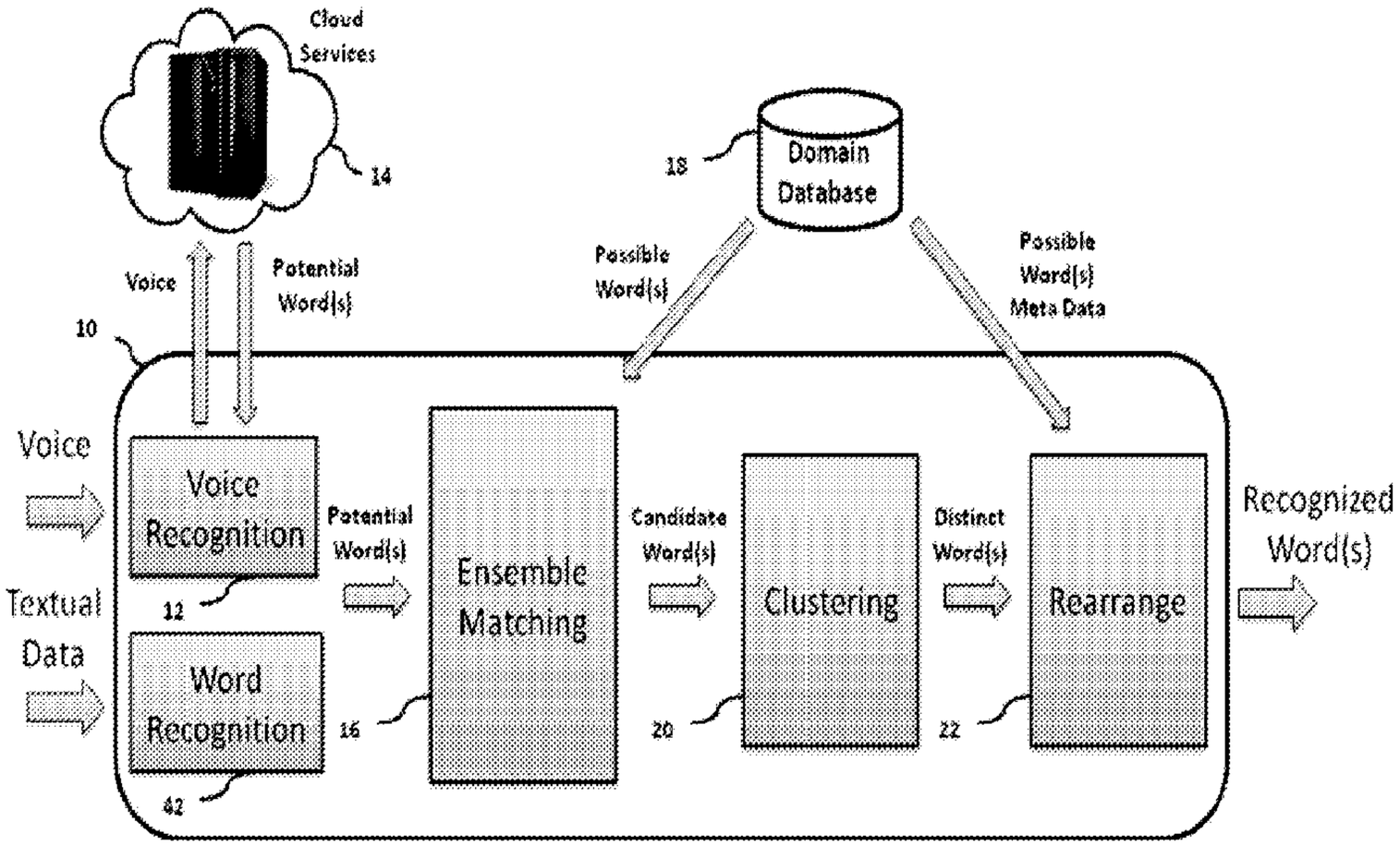

FIG. 1

(57) **Abrégé/Abstract:**
A human machine interface enables human users to interact with a machine by inputting auditory and/or textual data. The interface and corresponding method perform efficient look up of words, corresponding to inputted human data, which are stored in a domain database. The robustness of a speech synthesis engine is enhanced by updating the deployed pronunciation vocabulary dynamically. The architecture of the preferred embodiment of the former method includes a combination of ensemble matching, clustering, and rearrangement methods. The latter method involves retrieving suggested phonetic pronunciations for words unknown to the speech synthesis engine and verifying those through a manual or autonomous process.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 December 2014 (11.12.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/197592 A3

(51) International Patent Classification:
G10L 15/19 (2013.01) *G10L 13/02* (2013.01)
G10L 15/26 (2006.01)

(21) International Application Number:
PCT/US2014/040906

(22) International Filing Date:
4 June 2014 (04.06.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/830,789 4 June 2013 (04.06.2013) US

(71) Applicant: **IMS SOLUTIONS INC.** [US/US]; 1501 E Woodfield Road, Suite 113-E, Schaumburg, Illinois 60173 (US).

(72) Inventors: **CAMPBELL, David Neil**; Oakville, Ontario (CA). **RAE, Robert Andrew**; Elmira, ontario (CA). **EL-GHAZAL, Akrem Saad**; Waterloo, Ontario (CA). **SULPIZI, Daniel John Vincent**; Oakville, ONTario (CA).

(74) Agent: **CARLSON, John E.**; Carlson, Gaskey & Olds, P.C., 400 W. Maple, Suite 350, Birmingham, Michigan 48009 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(88) Date of publication of the international search report:
29 January 2015

(54) Title: ENHANCED HUMAN MACHINE INTERFACE THROUGH HYBRID WORD RECOGNITION AND DYNAMIC SPEECH SYNTHESIS TUNING

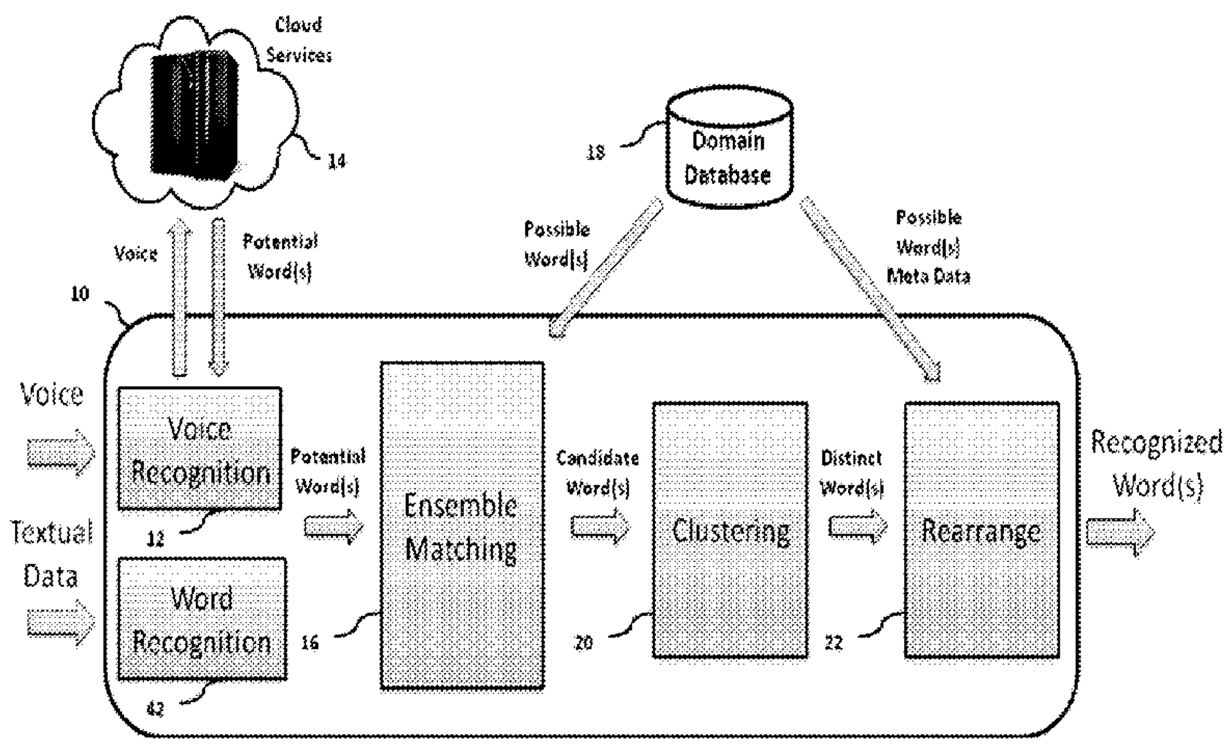

FIG. 1

(57) **Abstract**: A human machine interface enables human users to interact with a machine by inputting auditory and/or textual data. The interface and corresponding method perform efficient look up of words, corresponding to inputted human data, which are stored in a domain database. The robustness of a speech synthesis engine is enhanced by updating the deployed pronunciation vocabulary dynamically. The architecture of the preferred embodiment of the former method includes a combination of ensemble matching, clustering, and rearrangement methods. The latter method involves retrieving suggested phonetic pronunciations for words unknown to the speech synthesis engine and verifying those through a manual or autonomous process.

WO 2014/197592 A3

ENHANCED HUMAN MACHINE INTERFACE THROUGH HYBRID WORD RECOGNITION AND DYNAMIC SPEECH SYNTHESIS TUNING

BACKGROUND OF THE INVENTION

[0001] This application relates to enhanced human-machine interface (HMI), and more specifically two methods for improving user experience when interacting through voice and/or text. The two disclosed methods include a hybrid approach for human input transcription, as well as a robust text to speech (TTS) method capable of dynamic tuning of the speech synthesis process.

[0002] Automatic speech transcription of human input such as voice or text, is challenging due to the seemingly infinite domain of possible combinations, slang phrases, abbreviations, invented or derived phrases, and cultural dialects. Modern cloud-based recognition tools provide a powerful and affordable solution to the aforementioned problems. Nonetheless, they are typically inadequate when applied within a specific domain of application. As a result, efficient post-processing methods are required to map the recognition output provided by the aforementioned tools to a subset of words in a specific domain of interest.

[0003] Modern text to speech (TTS) technologies offer fairly accurate results where the targeted vocabulary is from a well-established and constrained domain. However, they might perform poorly when applied to more challenging domains containing new or infrequently used words, proper names, or derived phrases. Incorrect pronunciations of such words/phrases can make the product appear simple and naïve. On the other hand, many application domains, such as entertainment and sports, contain words that are transient and short lived in nature. Such volatile environments make it infeasible to employ manual tuning to keep pronunciation vocabularies up-to-date. Accordingly, automatic updating of the pronunciation vocabulary of TTS methods can significantly improve their flexibility and robustness in the aforementioned application domains.

SUMMARY OF THE INVENTION

[0004] Two methods for improving the user experience while interacting through voice and/or text are presented. The first disclosed method is a hybrid word look-up approach to match the potential words produced by a recognizer with a set of possible words in a domain database. The second disclosed method enables dynamic update of pronunciation vocabulary in an on-demand basis for words that are unknown to a speech synthesis system.

Together, the two disclosed methods yield a more accurate match for words inputted by a user, as well as more appropriate pronunciation for words spoken by the voice interface, and thus a significantly more user-friendly and natural human machine interaction experience.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] Figure 1 schematically illustrates a block diagram overview of the disclosed hybrid word look-up method.

[0006] Figure 2 schematically illustrates a block diagram overview of the disclosed dynamic speech synthesis engine tuning method.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0007] Figure 1 schematically illustrates the architectural overview for one embodiment of the disclosed hybrid look-up method as a word lookup-up system 10. Depending on its modality, the user input is fed to a voice recognition sub-system 12 or word recognition sub-system 42, which might operate by communicating wirelessly with a cloud-based voice/word recognition server 14, e.g. Google voice recognition engine. A set of potential words outputted by the voice recognition subsystem 12 are matched against the set of possible words, retrieved from a domain database 18, using an ensemble of word matching methods 16.

[0008] An ensemble of word matching methods 16 computes the distance between each potential word and each of the possible words. In an exemplary embodiment of the disclosed method, the distance is computed as a weighted aggregate of word distances in a multitude of spaces including the phonetic encoding, such as metaphone and double metaphone, string metric, such as Levenshtein distance, etc. The words are then sorted according to their computed aggregate distances and only a predefined number of top words are outputted as a set of candidate words and fed to a clustering method 20.

[0009] A set of candidate words are grouped into two segments by a clustering method 20. The first segment includes candidate words that are considered to be a likely match for the input user voice whereas the second segment contains the unlikely matches. The former category words are identified based on their previously computed aggregate distance by selecting the words that have a distinctly smaller distance. Consequently, the rest of the words are categorized as the second category. In a preferred embodiment of the

clustering method a well-known image segmentation approach, called Otsu method, can be used to identify a distinct set of words.

[0010] Before being presented to the user as a set of recognized words, a set of distinct words may be rearranged according to one or more of its associated metadata. The metadata are stored along with the set of possible words on a domain database 18 and include features such as frequently of usage, and user-defined or dynamically computed priority/importance, for each word. The rearrangement of words is particularly useful in disambiguation of distinct words with very close distinction level(s).

[0011] Figure 2 schematically illustrates the architectural overview of a speech synthesis system 40 that relies on the disclosed dynamic tuning method to update its vocabulary in an on-demand basis. A word recognition sub-system 42 extracts words contained in the input textual data. A speech synthesis engine 44 then converts the extracted words into a speech, to be played for a user. The speech synthesis engine groups words into two categories. The first category of words, referred to as native words, is those words that already exist in the phonetic vocabulary of a domain database 18. The second category of words, referred to as alien words, is those words that do not exist in the database 18.

[0012] For those words identified as alien, a cloud-based resource 14, such as the Collins online dictionary interface, is inquired to obtain one or more pronunciation phonetics suggestions. The obtained pronunciation phonetics could be represented using a phonetics markup language such as IPA or SAMPA. The suggested phonetics are presented to a human agent 46, e.g. a word is displayed on a screen while its suggested pronunciation is played out, to verify their validity. Alternatively, the suggested phonetic pronunciations can be validated using a software agent running on a local server 48. The confirmed pronunciation phonetics, along with their corresponding (previously) alien words, are then added to the domain database 18. This may be done in realtime (i.e. with the user possibly waiting a few seconds while the system confirms the pronunciation with the human agent 46, if there is not already sufficient words to be read to the user while the human verification is performed). Alternatively, this may be done offline, in which the case the user is presented with the best phonetic pronunciation available at the time, which is later validated by the human agent 46 and stored in the domain database 18.

[0013] The word-lookup system 10 may be a computer, smartphone or other electronic device with a suitably programmed processor, storage, and appropriate

communication hardware. The cloud services 14 and domain database 18 may be a server or groups of servers in communication with the word-lookup system 10, such as via the Internet.

[0014] In accordance with the provisions of the patent statutes and jurisprudence, exemplary configurations described above are considered to represent a preferred embodiment of the invention. However, it should be noted that the invention can be practiced otherwise than as specifically illustrated and described without departing from its spirit or scope.

CLAIMS

WHAT IS CLAIMED IS:

1. A method to perform word look-up based on human input including the steps of:

receiving human input;

performing initial recognition of the human input;

receiving metadata based upon the initial recognition;

prioritizing a plurality of possible words based upon the metadata; and

outputting a first word of the plurality of possible words based upon the prioritization.

2. The method in claim 1 wherein the human input is voice-based.

3. The method in claim 1 wherein the human input is text-based.

4. The method of claim 1 wherein the plurality of possible words and their associated metadata are stored in a domain database.

5. The method of claim 1 wherein recognition of human input voice data is performed using a voice recognizer to produce a set of potential words.

6. The method of claim 5 wherein a human input recognizer may be running locally or on a remote server residing in a cloud.

7. The method of claim 1 wherein a set of potential words are matched against the plurality of possible words using an ensemble of matching methods.

8. The method of claim 7 wherein an ensemble of nearest-neighbor methods operates by minimizing a weighted aggregate of potential to possible word distances.

9. The method of claim 8 wherein the potential to possible word distances are computed in two or more spaces.

10. The method of claim 9 where one space is phonetic encoding.

11. The method of claim 9 where one space is double metaphone encodings.

12. The method of claim 9 where one space is a natural edit distance.

13. The method of claim 1 wherein a set of candidate words is obtained by sorting a set of possible words according to their computed aggregate distance and outputting only a predefined number of top words.

14. The method of claim 1 wherein a set of produced candidate words are processed into two clusters of distinct words, and relevant/irrelevant words.

15. The method of claim 14 wherein the clustering is performed using a segmentation method.

16. The method of claim 14 wherein a set of produced distinct words can be rearranged according to their corresponding metadata of interest to produce a set of recognized words.

17. The method of claim 1 wherein the metadata includes frequency of usage.

18. A method of performing text to speech processing:

receiving text input including a plurality of words;

performing word recognition on the text input;

identifying native words of the plurality of words that already existing in a phonetic vocabulary; and

identifying alien words of the plurality of words that do not exist in the phonetic vocabulary.

19. The method of claim 18 wherein a vocabulary of words and their corresponding verified pronunciation are stored in the phonetic vocabulary.

20. The method of claim 19 further including the steps of dynamically retrieving through a remote server inquiry a suggested phonetic pronunciation for the alien word.

21. The method of claim 20 further including the step of validating a suggested phonetic pronunciation for the alien word by a human agent.

22. The method of claim 21 further including the step of adding the suggested phonetic pronunciation to the phonetic vocabulary based upon being validated by the human agent.

23. The method of claim 20 further including the step of validating a suggested phonetic pronunciation by a software agent.

24. The method of claim 21 further including the step of adding the suggested phonetic pronunciation to the phonetic vocabulary based upon being validated by the software agent.

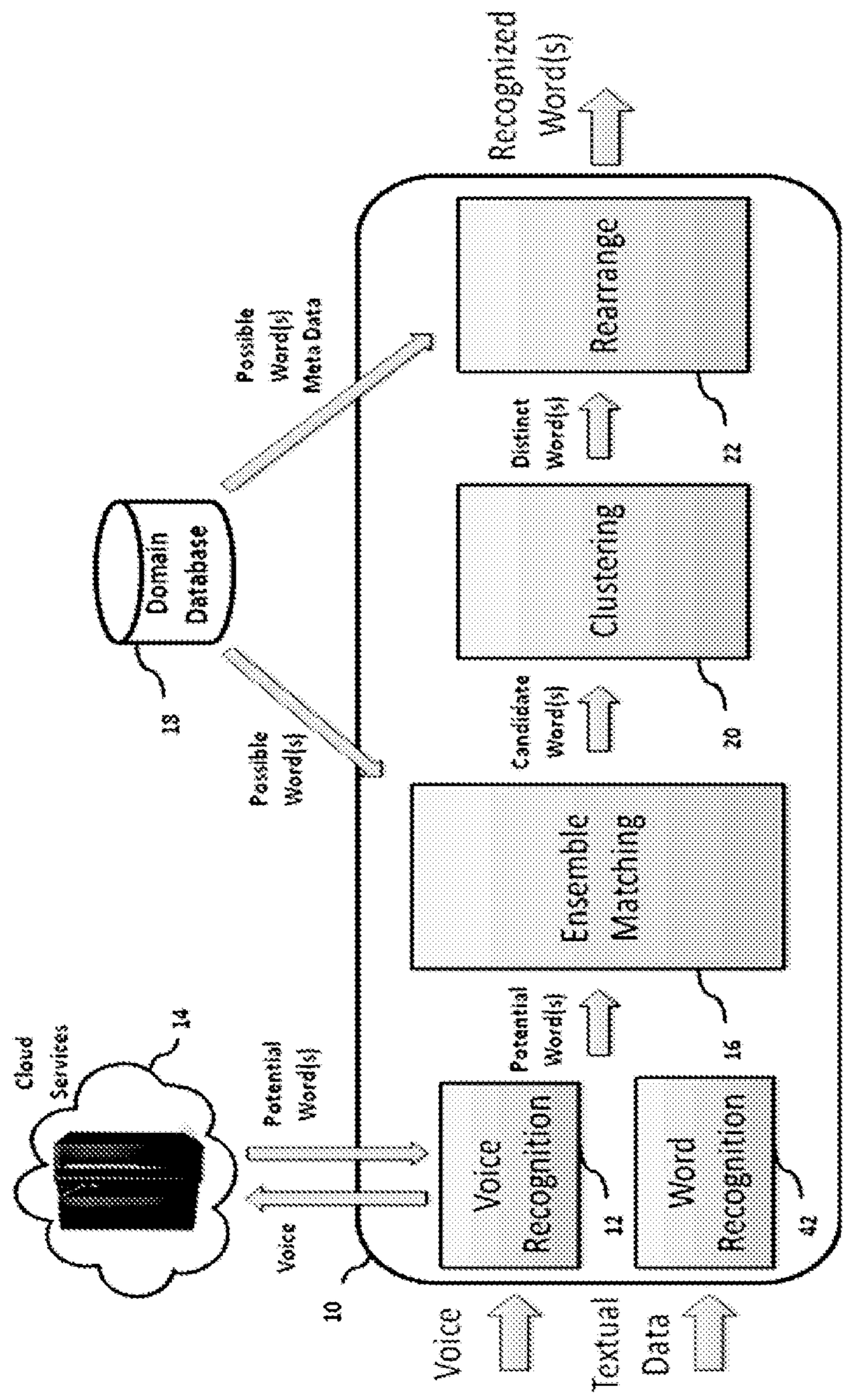
Recognized Word(s)
Domain Database
18
Possible Word(s) Meta Data
Possible Word(s)
Rearrange
22
Distinct Word(s)
Clustering
20
Candidate Word(s)
Ensemble Matching
16
Potential Word(s)
Cloud Services
14
Potential Word(s)
Voice
Voice Recognition
12
Word Recognition
42
10
Voice
Textual Data

FIG. 1

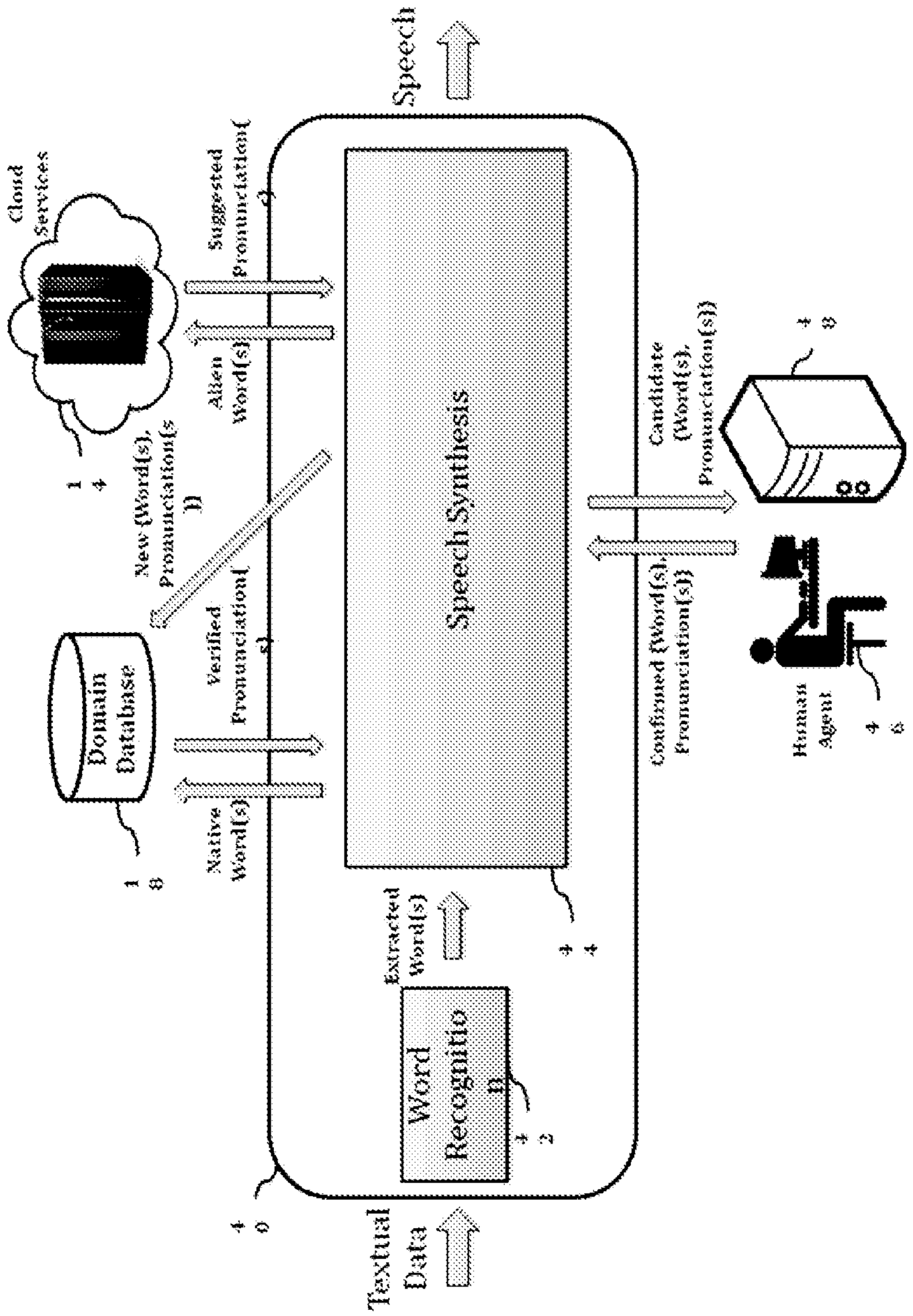
Speech
Suggested Pronunciation(
Cloud Services
Alien Word(s)
New (Word(s), Pronunciation(s))
Verified Pronunciation(
Domain Database
Native Word(s)
Speech Synthesis
Extracted Word(s)
Word Recognitio n
Textual Data
Candidate (Word(s), Pronunciation(s))
Confirmed (Word(s), Pronunciation(s))
Human Agent

FIG. 2

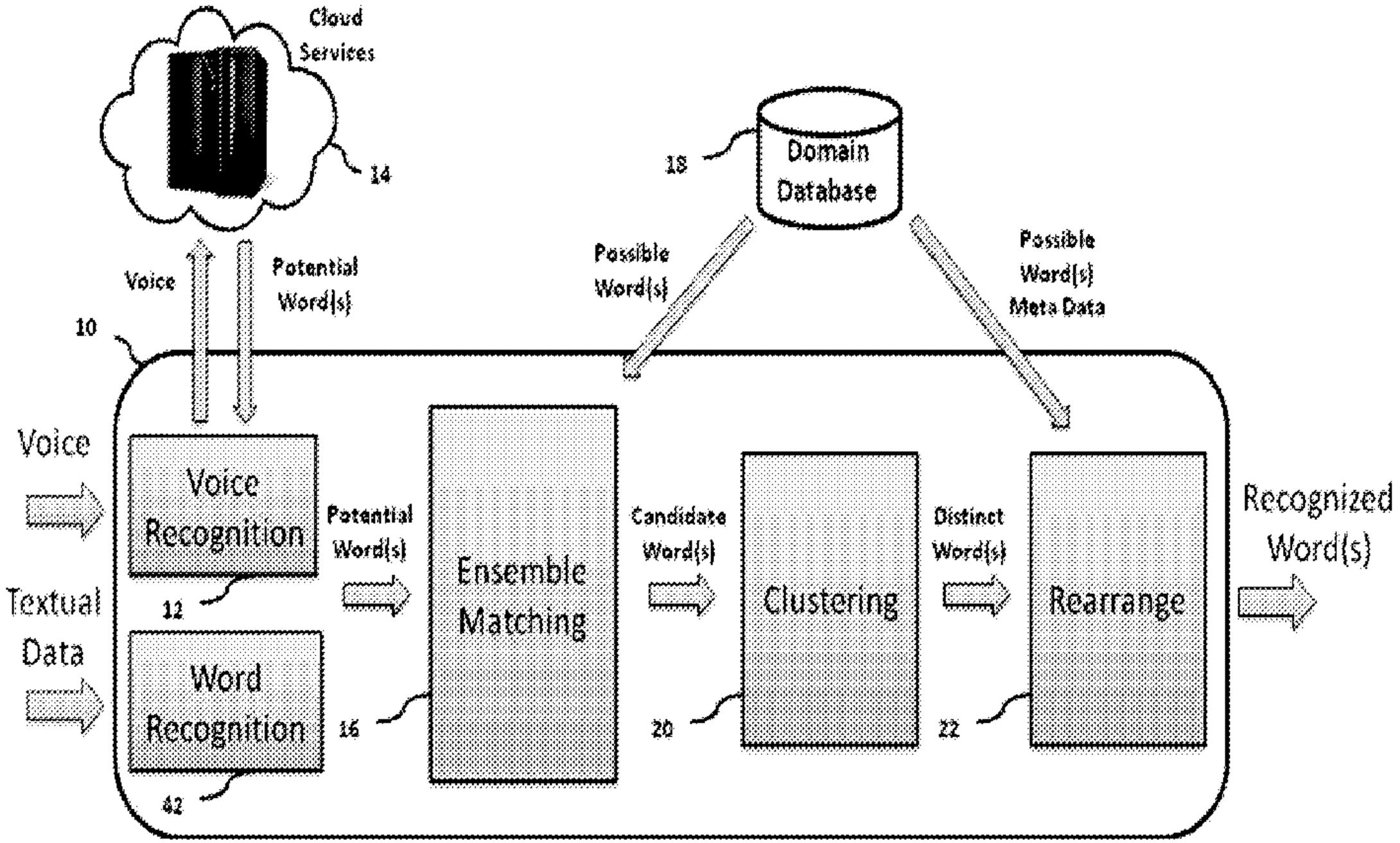
Cloud
Services
14
18
Domain
Database
Voice
Potential
Word(s)
Possible
Word(s)
Possible
Word(s)
Meta Data
10
Voice
Voice
Recognition
12
Potential
Word(s)
Ensemble
Matching
Candidate
Word(s)
Clustering
Distinct
Word(s)
Rearrange
Recognized
Word(s)
Textual
Data
Word
Recognition
42
16
20
22

FIG. 1