Technical Disclosure Commons

Defensive Publications Series

October 02, 2017

ASSISTANT VOCALIZED OUTPUT OF WORD COMBINATIONS

Google Inc.

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation

http://www.tdcommons.org/dpubs_series/733

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
ASSISTANT VOCALIZED OUTPUT OF WORD COMBINATIONS

ABSTRACT

A virtual, intelligent, or computational assistant (e.g., also referred to simply as an “assistant”) is described that is configured to properly pronounce word junctures when converting text to speech (e.g., when synthesizing audio data for output to a user). Some example word junctures that the assistant may properly pronounce include, but are not limited to, false geminates, affricates, and other such letter/word combinations. For instance, when performing text to speech on the text “black cat”, the assistant may determine that the consecutive combination of the words “black” and “cat” is a false geminate because the last consonant phoneme in “black” is the same consonant phoneme at the start of “cat” (i.e., black ends with the consonant phoneme /k/ and cat starts with the consonant phoneme /k/). As such, the assistant may pronounce the text “black cat” differently than the separate pronunciations of “black” and “cat.” Specifically, the assistant may avoid repeating the consonant phoneme /k/ when pronouncing “cat.”

DESCRIPTION

Assistants execute on counter-top devices, mobile phones, automobiles, and many other type of computing devices. Assistants output useful information, responds to users’ needs, or otherwise performs certain operations to help users complete real-world and/or virtual tasks. Some assistants may perform text to speech (TTS) operations to read text aloud by playing audio data synthesized based on the text. In regular person-to-person speech, it may be considered
proper for some combinations of words to be pronounced differently than the words on their own.

The example system shown in FIG. 1 provides an assistant that pronounces combinations of words differently based on context when performing TTS. For example, when the assistant is reading aloud text that includes a particular combination of words, the assistant may determine whether to pronounce the particular combination of words differently than the pronunciations of the words included in the particular combination of words.

FIG. 1
The system of FIG. 1 includes one or more external systems and computing devices A–N communicating across a network with each of computing devices A–N executing an assistant that performs operations involving groups of people. The network of FIG. 1 represents a combination of any one or more public or private communication networks, for instance, television broadcast networks, cable or satellite networks, cellular networks, Wi-Fi networks, broadband networks, and/or other type of network for transmitting data (e.g., telecommunications and/or media data) between various computing devices, systems, and other communications and media equipment. Computing devices A–N represent any type of computing device, or other system that is configured to execute an assistant and communicate on a network. The external systems represent any type of cloud computing environment, mainframe, server, or other computing system that is configured to support the assistants executing at computing devices A–N.

Computing devices A–N can be personal computing devices. In some examples, the external systems and/or computing devices A–N may be shared assets of multiple users. Examples of computing devices A–N mobile phones, tablet computers, wearable computing devices, countertop computing devices, home automation computing devices, laptop computers, desktop computers, televisions, stereos, automobiles, and any and all other type of mobile and non-mobile computing device that is configured to execute an assistant. For example, computing device A may be a countertop assistant device and computing device N may be a mobile phone or automobile infotainment system.

An assistant executes across any combination of external systems one or more of computing devices A–N to provide assistant services to users of computing devices A–N. Examples of assistant services include: setting up reminders, creating calendar entries, booking
travel, online ordering, sending messages or other communications, reading text aloud, controlling
television, lights, thermostats, appliances, or other computing devices, providing navigational
instructions, or any other conceivable task or operation that may be performed by an assistant.

As a user interacts with the assistant, the assistant may obtain personal information about the user. Examples of personal information include: habits, word or phrase selections, voice samples, routines, preferences, notes, lists, contacts, communications, interests, location histories, and other types of user information. After receiving explicit permission from the user, the assistant may store, the personal information at user information data stores and in the course of providing assistant services, make use of the personal information stored at the user information data stores.

The external systems and computing devices A–N and the assistant treat the information stored at the information stores so that the information is protected, encrypted, or otherwise not susceptible to unauthorized use. The information stored at the information data stores may be stored locally at each of computing devices A–N and/or remotely (e.g., in a cloud computing environment provided by the external systems and which is accessible via the network of FIG. 1).

Further to the descriptions below, a user may be provided with controls allowing the user to make an election as to both if and when the assistant, the computing device, or the computing systems described herein can collect or make use of supplemental data (e.g., user information or contextual information about a user’s social network, social actions or activities, profession, a user’s preferences, or a user’s current location), and if and when the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways
before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity may be treated so that no personally identifiable information can be determined for the user, or a user’s geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what supplemental data is collected about the user, how that supplemental data is used, and what supplemental data is provided to the user.

In operation, the assistant may perform TTS to read text aloud to a user. For instance, the assistant may read messages, responses to questions, documents, or any other text aloud to the user. When the assistant is performing TTS, the assistant may determine the pronunciations of words based on their neighboring words (i.e., the assistant may determine pronunciations of word junctures). For instance, when the assistant would independently pronounce a particular word in a first way (e.g., “you”), the assistant may pronounce the particular word differently based on context (e.g., “you” may be pronounced differently if proceeded by the word “hit” or another word that ends in “t”).

Some example word junctures that the assistant may properly pronounce include, but are not limited to, false geminates, affricates, and other such letter/word combinations. An example of a false geminate that the assistant may properly pronounce is “hip pocket” in-which the assistant may avoid repeating the /p/ consonant phoneme twice (i.e., the assistant may only pronounce the /p/ consonant phoneme at the end of “hip” and not pronounce the /p/ consonant phoneme at the start of “pocket”). An example of an affricate that the assistant may pronounce is “hit you” in-which the assistant may pronounce the combination of “hit” and “you” differently (e.g., the assistant may pronounce the combination as “hit-chu”).
The assistant may determine the proper pronunciations in a number of ways. As one example, when reading a sentence aloud, the assistant may determine the pronunciation for the entire sentence before outputting the corresponding audio. As another example, when reading a current word in a sentence aloud, the assistant may consider the following and/or proceeding N (e.g., 1, 2, 3, etc.) words when determining the pronunciation of the current word.

The assistant may contextually apply the word juncture pronunciation techniques. For instance, the assistant may apply the above word juncture pronunciation techniques when reading sentences but not when reading words in a table (or other words subject to syntactic constraints). The word juncture techniques may be stored in a lookup table, and identified for use based on the first word or word combinations in the target phrase to be spoken. The appropriate word juncture techniques may also or alternatively be identified by evaluating the phonetics of a target phrase. Machine learning techniques may be applied to identify likely word combinations to which juncture techniques are to be applied, with frequency of use of phrases, transcription of user commands, and other inputs use to assistant with training of models for the machine learning. Furthermore, the use of juncture techniques may be based on the vocalization model being employed for TTS. That is, for an assistant voice output with one accent or dialect, a certain first set juncture techniques may be utilized, while for a different accent or dialect of the voice assistant a second, different set of juncture techniques may be utilized.

By pronouncing word junctures properly, the assistant may be able to read text aloud in a less mechanical manner and/or sound like an actual person. As such, the assistant may allow for less awkward and smoother user interactions. The above examples are just some use cases for the assistant architecture shown in FIG. 1, the assistant architecture has many other applications and use cases.