INFORMATICA MUSICALE

UNIVERSITA' DEGLI STUDI DI CATANIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA
LAUREA TRIENNALE IN INFORMATICA
A.A. 2018/19
Prof. Filippo L.M. Milotta

ID PROGETTO: 25

TITOLO PROGETTO: Il mondo degli assistenti vocali

AUTORE 1: Giurato Salvatore

Indice

1. Obiettivi del progetto

1.1 Storia degli assistenti vocali

Il primo dispositivo di riconoscimento vocale venne costruito nel 1952 nei laboratori di ricerca Bell e venne chiamato Audrey.

Audrey riusciva a capire le cifre da 0 a 9 a patto che vi fosse una pausa fra una cifra e l'altra.

Audrey non ebbe molto successo a causa delle sue grandi dimensioni, ai requisiti di alimentazione ed ai costi di produzione e manutenzione.

Nel 1962 un ingegnere IBM presentò la Shoebox, era simile ad una calcolatrice a comando vocale, riusciva a comprendere 10 cifre e 6 parole di controllo (+, -, totale, SUBTOTAL, falso e spento) ed era collegata ad una macchina che calcolava e stampava i problemi matematici di base.

Come per Audrey anche Shoebox riusciva a riconoscere le cifre e le parole solo se vi era una pausa tra una parola e l'altra.

Nel 1971 la DARPA (Defense Advanced Research Projects Agency) finanziò un progetto di riconoscimento vocale della durata di 5 anni che portò al lancio di Harpy.

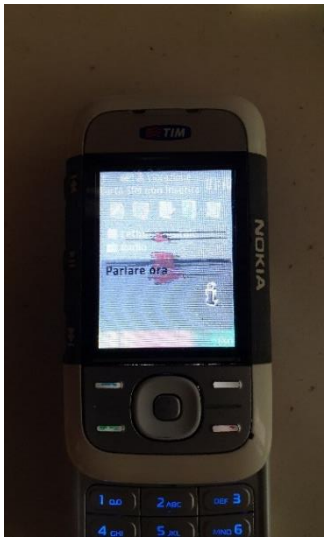
Harpy era una macchina che conteneva 1011 parole e poteva comprendere frasi, anche con parole iniziate e/o interrotte, Harpy inoltre quando non riusciva a comprendere qualcosa restituiva un messaggio di errore del tipo: "Non so cosa hai detto, per favore ripeti" come negli attuali assistenti vocali.

Nel 1986 venne rilasciato Tangora, un aggiornamento di Shoebox che si collegava ad una macchina da scrivere.

Tangora riusciva a riconoscere circa 20000 parole e, basandosi su ciò che era riuscito ad interpretare fino a quel momento, cercava di prevedere le parole successive, tuttavia non riusciva a adattarsi adeguatamente ai singoli utenti.

Negli anni 90 la tecnologia degli assistenti vocali arrivò all'interno dei call center; alcune aziende del settore realizzarono soluzioni che permettevano ai call center di gestire le chiamate in base alle interazioni vocali di chi chiamava.

Nel 1997 venne rilasciato il software NaturallySpeaking di Dragon che riusciva a riconoscere e trascrivere il linguaggio umano e senza fermarsi tra una parola e l'altra, scriveva in un documento digitale ad una velocità di 100 parole al minuto.



Nei primi anni 2000 alcuni cellulari permettevano l'avvio di una chiamata tramite un comando vocale o l'invio di un messaggio:

nell'immagine sulla destra vi è un Nokia 5300 (Anno uscita modello: 2006) in attesa di un comando vocale (ad esempio: "Chiama Bill").

Nell'anno 2010 venne rilasciato Watson da IBM che era un computer in grado di rispondere a domande basate sul linguaggio naturale; Watson riuscì a vincere il noto quiz statunitense Jeopardy; fino ad allora era fantascienza poter pensare una cosa simile.

Con la crescita del machine learning e dell'intelligenza artificiale l'efficienza del riconoscimento vocale è migliorata drasticamente favorendo la nascita degli assistenti vocali che vengono utilizzati in diversi ambiti.

1.2 Ambiti in cui è possibile usare gli assistenti vocali

Gli assistenti vocali di oggi sono in grado di interpretare il linguaggio naturale e dialogare con gli umani riuscendo a fornire le informazioni richieste dall'utente attraverso diversi comandi vocali.

Gli assistenti vocali si possono trovare negli Smartwatch, Smartphone, altoparlanti Smart e addirittura anche in alcune auto.

Tramite gli assistenti vocali è possibile:

- Avviare una chiamata,
  ad esempio, utilizzando il comando vocale da un iPhone:
  "Ehi Siri, chiama Mamma con il vivavoce";

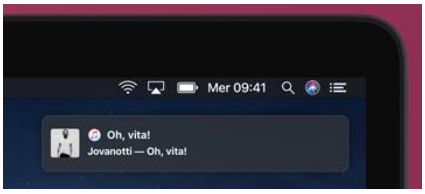
- Inviare un messaggio su WhatsApp,
  ad esempio, utilizzando il comando vocale da iPhone:
  "Ehi Siri, scrivi a Lia Longo su WhatsApp che arrivo fra 30 minuti";

- o inviare un semplice SMS;

- ottenere informazioni generali come sapere quando è l'alba o il tramonto, che ore sono in un altro paese con un fuso orario diverso, chiedere una traduzione, chiedere la conversione di una cifra da una valuta in un'altra o semplicemente chiedere il meteo della giornata o di un altro giorno;

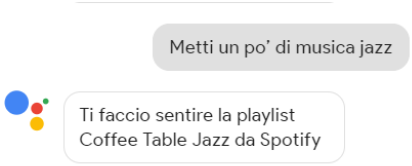
- pianificare la giornata;

- Avviare musica scegliendo la canzone, l'album o l'artista, ad esempio utilizzando il comando vocale:
"Ehi Siri, riproduci l'ultimo album di Jovanotti";



o facendo scegliere la musica al dispositivo, in questo caso il sistema si baserà sugli ascolti fatti di recente, ad esempio utilizzando il comando vocale: "Ehi Siri, fammi sentire qualcosa che mi piace";

oppure scegliendo il genere musicale, ad esempio utilizzando il comando vocale "Ok Google, Metti un po' di musica Jazz";



Se a casa si possiedono dei dispositivi Smart, quindi in grado di comunicare con gli assistenti vocali tramite una connessione a internet, è possibile gestirli utilizzando solo la voce:

- avviare un film o una serie TV se si ha una televisione connessa;

- accendere o spegnere le luci;
- accendere la caldaia a distanza;
- azionare una telecamera;
- avviare uno specifico ciclo di lavaggio di una lavatrice o di una lavastoviglie.

Le nuove auto di Mercedes e BMW sono dotate di un assistente vocale che permette di eseguire i comuni comandi come: richiedere informazioni, inviare messaggi o effettuare chiamate (se collegate con il cellulare) e di utilizzare comandi per la gestione dell'automobile:

- attivare il navigatore satellitare per raggiungere una destinazione;
- controllare i consumi e l'autonomia del carburante;
- salvare una stazione radio, cambiare la stazione radio;
- accendere l'aria condizionata e impostare la temperatura;
- accendere il riscaldamento e impostare la temperatura;
- accendere o spegnere i fari;
- accendere le luci da lettura che si trovano all'interno dell'auto;
- attivare il riscaldamento di uno o più sedili;
- attivare lo sbrinamento;

La possibilità di utilizzare un comando vocale in auto durante la guida dà la possibilità di poter svolgere qualunque compito senza dover staccare le mani dallo sterzo e senza dover distogliere lo sguardo dalla strada.

1.3 Perché si preferisce utilizzare la voce femminile?

Tutti gli assistenti vocali più conosciuti hanno delle voci femminili, in genere non si presta attenzione a questo particolare, quindi la domanda da porsi è: perché gli assistenti vocali più conosciuti hanno una voce femminile?

Ci sono diverse teorie che provano a dare una risposta a questa domanda:

- Nel 1980 il dipartimento dei trasporti degli Stati Uniti dopo aver fatto diversi sondaggi tra i piloti, stabilì che preferivano i sistemi di allarme automatico con voci femminili rispetto a quelli con voce maschile (a detta di molti piloti perché la voce femminile si sarebbe distinta dalla maggior parte delle altre voci in cabina di pilotaggio), in realtà dati empirici dimostrarono che non vi era una differenza significativa nel modo in cui i piloti rispondevano a voci maschili o femminili.
- Clifford Nass che fu uno dei precursori nel campo degli assistenti digitali scrisse in un suo libro che la nostra mentalità ci porta a pensare che una voce femminile ci aiuta a risolvere un problema da soli, mentre una voce maschile ci porta a pensare che la soluzione sarà imposta, creandoci un disagio.
- Karl Fredric MacDorman professore di informatica ed esperto di interazione uomo-computer fece una ricerca su come uomini e donne reagivano alle voci maschili ed alle voci femminili, il test consisteva nel far ascoltare a uomini e donne spezzoni di voci maschili e spezzoni di voci femminili. Quindi, somministrò ad un campione di persone, un questionario in cui veniva richiesto quali voci preferissero, ed eseguì un test che misurava le loro preferenze implicite; il risultato dei test fu che sia gli uomini che le donne preferivano voci femminili. MacDorman concluse lo studio dicendo: "I think there's a stigma for males to prefer males, but there isn't a stigma for females to prefer females" cioè pensa che ci sia una censura degli uomini a preferire una voce maschile rispetto ad una voce femminile mentre le donne non hanno nessuna censura a preferire le donne.

In conclusione, possiamo dire che le persone psicologicamente tendono a preferire la voce femminile perché la percezione che si ha da una voce femminile è migliore rispetto ad una voce maschile, tuttavia bisogna cercare di evitare che un utente con mentalità retrograda possa considerare l'assistente vocale con voce da donna come una segretaria a cui si può dire tutto senza che essa possa reagire.
Per evitare che ci siano abusi la soluzione può essere quella di poter scegliere fra almeno una voce maschile ed una voce femminile in base alla preferenza di ogni utente, e se l'utente dovesse commettere degli abusi, bisogna fare in modo che l'assistente vocale risponda a tono; le aziende hanno la responsabilità di dover garantire che i loro prodotti non siano vittime di stereotipi di genere.

1.4 Funzionamento degli assistenti vocali

Il primo passo per utilizzare gli assistenti vocali è quello di “svegliare” l’assistente vocale.

Gli assistenti vocali possono essere “svegliati” in diversi modi:

- Premendo un pulsante che si trova fisicamente sul dispositivo dove si trova l’assistente vocale
- Pronunciando una parola chiave, quindi questo tipo di assistenti vocali rimangono sempre in ascolto.
- Premendo un pulsante e/o pronunciando una parola chiave.

L’assistente vocale che viene svegliato quando si pronuncia una parola chiave utilizza un software di keyword spotting, il keyword spotting è un processo di speech-to-text simile al riconoscimento del parlato ma limitato al riconoscimento della parola chiave.

Per costruire un software di keyword spotting bisogna raccogliere una grande quantità di ore di registrazione della parola e bisogna fare in modo da avere il dataset il più eterogeneo possibile, cioè bisogna avere la registrazione della parola chiave fatta con diversi accenti, diverse pronunce e il più bilanciato possibile fra audio di utenti maschili e femminili.
Dopo aver creato il dataset contenente le diverse ore di registrazione, bisogna costruire una rete neurale che riesca a capire quando viene pronunciata quella parola chiave.
Per costruire la rete neurale il dataset viene diviso in training set, test set e validation test (ad esempio 80%, 10%, 10%).
Una rete neurale è un modello computazionale composto da neuroni artificiali ispirato alla semplificazione di una rete neurale biologica e si occupa di risolvere problemi di intelligenza artificiale e machine learning.

Una volta “svegliato” l’assistente vocale, bisognerà formulare una richiesta, per far sì che l’assistente vocale esegua un software di speech-to-text: sostanzialmente trasformerà l’audio della richiesta in testo.

Il software di speech-to-text spezza l’audio in piccole parti chiamate fonemi, un fonema è l’unità fonologica minima dotata di carattere distintivo: i fonemi nella lingua italiana sono 30 ma considerando le doppie diventano 45 e nella lingua inglese sono 44;

Il software analizza:

- l’acustica che è rappresentata dai fonemi che sono stati pronunciati e cerca le parole che li contengono;
- la pronuncia, perché dal modo in cui i fonemi sono stati pronunciati, catturara la variabilità fonetica della parola e capisce quali parole sono state pronunciate.

Il software, dopo aver trasformato le parole pronunciate in testo, deve interpretare cosa è stato detto in modo che la risposta da dare sia il più possibile coerente con la richiesta effettuata dall’utente, quindi ricerca tutte le possibili risposte alla richiesta e con un algoritmo di machine learning sceglie la risposta che ha la probabilità più alta di soddisfare l’utente.

Infine, dopo aver interpretato la richiesta dell’utente ed aver trovato una possibile risposta l’assistente vocale cerca di soddisfare la richiesta rispondendo alla domanda fatta o eseguendo uno dei compiti richiesti.

1.5 Informazioni sulla privacy

Ad aprile 2019 è stato pubblicato un articolo in cui si parla di Alexa, che è l'assistente vocale di Amazon, nel quale si dice che Alexa salva gli audio delle conversazioni.

I dipendenti di Amazon prendono dei campioni degli audio in modo da valutare le risposte e migliorarle per fornire un servizio sempre migliore.

Questa pratica di riascoltare le registrazioni non è solo ristretta ai dipendenti di Amazon ma anche a quelli delle altre aziende.

Nei casi in cui gli assistenti vocali vengono attivati da una parola chiave, quindi sono sempre in ascolto c'è la possibilità che possano intercettare una conversazione anche quando non è stata pronunciata la parola chiave e possano registrare la conversazione.
Anche in questo caso, le conversazioni possono essere ascoltate dai dipendenti dell'azienda dell'assistente vocale.

Gli audio salvati non sono legati all'account quindi non dovrebbe esserci la possibilità di risalire all'utente, ma c'è sempre la possibilità che nella registrazione venga rivelato un nome oppure una località da cui si possa risalire all'utente o alla posizione dell'utente: questa è sicuramente una grossa violazione di privacy.

In situazioni particolari, ad esempio quando un dipendente ascolta un audio divertente, lo potrebbe segnalare ai colleghi e quindi più persone potrebbero venire a conoscenza di un audio che non dovrebbe essere ascoltato da nessuno, oppure se un dipendente ascoltando un audio si rendesse conto che c'è una situazione di pericolo come la progettazione di un attentato o una situazione di violenza domestica, come dovrebbe comportarsi? Dovrebbe far finta di nulla oppure dovrebbe segnalarlo alle forze dell'ordine cercando anche di collegare l'audio al dispositivo da cui è stato registrato? Se da una parte con gli assistenti vocali non esiste più la privacy visto che qualunque conversazione potrebbe essere registrata dall'altra c'è l'occasione di avere più sicurezza.

Per trasformare queste violazioni di privacy in opportunità bisognerebbe trovare il giusto equilibrio fra privacy e sicurezza, che non è una cosa facile da fare.

2. Riferimenti Bibliografici

- https://www.theverge.com/ad/17855294/a-brief-history-of-voice-assistants: in questo articolo si parla della storia degli assistenti vocali a partire dai primi strumenti di riconoscimento vocale fino ad arrivare a Watson uno dei primi assistenti vocali come li conosciamo oggi.
- https://www.apple.com/it/siri/ : Il sito web della Apple nella sezione Siri spiega tutte le possibili funzionalità di Siri che è l'assistente vocale di Apple.
- https://assistant.google.com/intl/it_it/platforms/phones/: Il sito ufficiale dell'assistente di Google spiega le sue possibili funzionalità dividendo i diversi dispositivi: telefoni, Smart Speaker e indossabili.
- https://www.mercedesbenzclub.it/forum/viewtopic.php?t=116668: E' un forum di Mercedes in cui vengono elencati i possibili comandi per poter comunicare con l'assistente vocale di Mercedes per le auto e viene aggiornato ogni volta che viene scoperto un nuovo comando vocale.
- https://www.youtube.com/watch?v=C-gRJrDrICs: E' un video di una pubblicità di BMW in cui viene spiegato come funziona l'assistente vocale di BMW e vengono fatti alcuni esempi di comandi.
- https://www.theatlantic.com/technology/archive/2016/03/why-do-so-many-digital-assistants-have-feminine-names/475884/: in questo articolo viene citato Clifford Nass e si parla anche dello studio fatto nel 1980 sui piloti del dipartimento dei trasporti americano.
- https://www.internazionale.it/opinione/martin-caparros/2015/12/03/siri-sessismo: in questo articolo si parla di Clifford Nass e l'autore pensa che parlare con un assistente vocale sia sessismo.
- https://www.livescience.com/49882-why-robots-female.html: in questo articolo si parla dello studio di MacDorman.
- https://www.marshall.usc.edu/blog/how-do-digital-voice-assistants-eg-alexa-siri-work: questo articolo spiega i passaggi che compie un assistente vocale dopo aver ricevuto un comando vocale.
- https://www.youtube.com/watch?v=nydKXvL5ZKI: In questo video si parla dei possibili problemi di privacy legate agli assistenti vocali e di possibili opportunità.

3. Argomenti Teorici Trattati

In questo documento gli argomenti teorici trattati sono stati:

- la storia degli assistenti vocali;
- la psicoacustica quando si parla degli studi fatti sul perché si preferisce una voce femminile ad una voce maschile;
- il keyword spotting, utilizzato quando bisogna svegliare un assistente vocale.
- Cenni sulle reti neurali
- Come funzionano i software di speech-to-text
- Cos'è un fonema