POLITECNICO DI TORINO

Collegio di Ingegneria Informatica, del Cinema e Meccatronica

Corso di Laurea Magistrale
in Ingegneria Informatica

Tesi di Laurea Magistrale

Analisi e applicazione delle nuove tecnologie digitali a supporto dell'interazione Banca - Cliente attraverso assistenti vocali digitali

Politecnico
di Torino

Relatore
prof. Luca Cagliero

Candidato
Davide Fersino

Anno Accademico 2022/2023

INDICE

I. INTRODUZIONE

II. ASSISTENTI DIGITALI E DISPOSITIVI SMART

1. *Cosa sono, utilizzi e benefici*
2. *Struttura e funzionamento*
    a. *Speech-to-Text*
    b. *Text-to-Speech*
    c. *Natural Language Understanding*

III. L'INTERAZIONE BANCA - CLIENTE

1. *L'interazione attuale e i benefici degli assistenti digitali*
2. *Analisi delle soluzioni dei principali istituti bancari*

IV. ANALISI E RISULTATI DEL PROGETTO

1. *Panoramica*
2. *Architettura e logica di funzionamento*
    a. *Limiti e rischi*
    b. *Varianti proposte*
    c. *Le biometriche vocali*

V. CONCLUSIONI

VI. BIBLIOGRAFIA

ELENCO DELLE FIGURE

Figura 1: Numero di utilizzatori di assistenti vocali negli Stati Uniti dal 2017 al 2022. Fonte [41]……...6

Figura 2: Percentuale di utenti ad utilizzare assistenti vocali per effettuare acquisti negli Stati Uniti dal 2018 al 2020. Fonte [42]……..7

Figura 3: Visualizzazione grafica dell'elaborazione di un comando fornito ad un assistente digitale. ……10

Figura 4: Architettura di un generico sistema di riconoscimento vocale ……..12

Figura 5: Schema di un HMM per la classificazione di fonemi…….14

Figura 6: Reticolo di parole generato da un HMM……16

Figura 7: Rappresentazione grafica di una RNN standard, la variante a destra mostra il funzionamento in funzione del tempo……...……...19

Figura 8: Rappresentazione grafica della cella di una rete LSTM…….20

Figura 9: Rappresentazione grafica del funzionamento di uno strato di self-attention che utilizza le matrici di query, key e value…….26

Figura 10: Visuallizazione grafica di come le teste di un Transformer identifichino le relazioni a distanza nella seuqenza. Si può notare come l'uscita per il verbo "making" ha identificato forti relazioni con "laws", "2009", "more" e "difficult", andando a delineare le informazioni importanti che lo riguardavano. Ogni colore nel disegno rappresenta l'uscita di una delle teste del modello. Fonte [14]. ……..……..29

Figura 11: Architettura della rete Transformer usata in [17] come codificatore audio e codificatore delle etichette. ……...…….. 34

Figura 12: Architettura di un generico sistema di sintesi vocale……...……36

Figura 13: Tre modelli architetturali a confronto. BERT usa dei Transformer bidirezionali. OpenAI GPT usa Transformer che sfruttano solo il contesto passato.

ELMo concatena due blocchi separati di LSTMs uno che sfrutta solo il contesto passato della sequenza di ingresso e uno solo quello futuro. Fonte [20]……....41

Figura 14: Rappresentazione grafica degli ingressi in una rete BERT. Gli ingressi del modello sono generati sommando la versione codificata del vettore di ingresso alle informazioni extra di: frase di appartenenza e posizione nella stessa. Fonte [20]…….42

Figura 15: Percentuale di compagnie che riscontrano i benefici menzionati in figura dall'adozione degli assistenti digitali raggruppate per settore. 1. Riduzione di oltre il 20% del costo del servizio clienti, 2. Oltre quattro ore persona risparmiate ogni giorno, 3. Riduzione di oltre il 20% delle chiamate all'assistenza clienti, 4. Incremento di oltre il 20% nella percentuale di utenti che utilizzano gli assistenti digitali, 5. Riduzione di oltre il 20% del tasso di abbandono dei clienti, 6. Incremento di oltre il 20% delle interazioni gestite. Fonte [43]……...46

Figura 16: Industrie che incrementeranno gli usi e le applicazioni delle tecnologie vocali nel giro di 3-5 anni (dal 2021). Fonte [44]……..……...51

Figura 17: Rappresentazione grafica semplificata dell'interazione con l'applicazione per Alexa distribuita da Capital One. La rappresentazione è stata semplificata escludendo gli elementi relativi alla connessione tra il dispositivo Amazon e l'account utente Capital One, l'interazione completa di una soluzione di questo tipo sarà disponibile nel quarto capitolo. …….……...54

Figura 18: Principali traguardi nel numero di interazioni di Erica, l'assistente di Bank of America dal lancio alla fine del 2022. Fonte [26]……...……..56

Figura 19: Quantità di smart speaker spediti nel mondo, raggruppati per venditore, dal 2016 al 2022. Fonte [45]……..…….59

Figura 20: Oggetto JSON di esempio inviato dall'assistente Alexa al servizio di fulfillment nel momento in cui il comando richiesto dall'utente sia stato abbinato ad uno degli intenti gestibili dalla skill ricevente. In questo esempio in particolare, è stato chiesto al servizio di gestire l'intento di nome "CreditRefillIntent" e come si può notare non è stato valorizzato nella richiesta nessuno dei valori aggiuntivi (slots). ……62

Figura 21: Rappresentazione tramite PlantUml della procedura di account linking effettuata nel prototipo. ……..…….66

Figura 22: Rappresentazione grafica dell'architettura della skill creata come prototipo per l'assistente Amazon Alexa. ……69

Figura 23: PSD di due oratori differenti presi come campioni. Fonte [47]……....74

ELENCO DELLE TABELLE

Tabella 1: Caratteristiche personali imputate dell'inutilizzo di sevizi di Online Banking. Fonte [22]…….48

Tabella 2: Valutazione dei servizi di Online Banking da parte dei gruppi di utilizzatori e non. Fonte [22]……...49

Tabella 3: Risultati ottenuti dai primi 5 classificati nella NIST CTS SRE. Fonte [29]..75

I. INTRODUZIONE

Giorno dopo giorno la tecnologia penetra sempre più a fondo nelle nostre vite, andando a semplificare e velocizzare la maggior parte delle nostre attività.

Vite sempre più frenetiche e prive di tempo ci hanno spinto a delegare molte delle nostre interazioni giornaliere alla tecnologia. È una tendenza che ha preso vita con l'avvento degli smartphones, che ci hanno permesso di avere tutto a portata di click ovunque noi andiamo, e che ora sta lentamente evolvendo spostandosi sulle tecnologie vocali.

Con tecnologie vocali ci riferiamo a quella serie di prodotti che basano il loro funzionamento sulla voce umana, tra questi, ad oggi, gli assistenti vocali digitali rappresentano il prodotto più usato. Quest'ultimi sfruttano la voce umana come mezzo per impartire un'istruzione ad un dispositivo. La richiesta di un utente viene elaborata da algoritmi di riconoscimento vocale, per ottenerne una versione scritta, e poi analizzata in algoritmi di comprensione del linguaggio naturale per categorizzarla e di conseguenza identificare le intenzioni dell'utente.

Ad oggi queste tecnologie sono presenti in molti dei nostri dispositivi, infatti la maggior parte degli smartphones, dispositivi indossabili e altoparlanti intelligenti, permettono di utilizzare tali prodotti. La consapevolezza degli utenti riguardo agli assistenti digitali è in continua crescita, così come il loro utilizzo, non vi sono ad oggi particolari prove che facciano pensare che questa tecnologia non sia il futuro dell'interazione tra uomo e macchina, sebbene non si possano definire propriamente maturi al momento.

Come si nota da alcuni sondaggi condotti da PwC in *"Consumer Intelligence Series: Prepare for the voice revolution"* [1], l'adozione di questi assistenti è ancora limitata ad attività molto semplici. La maggior parte degli utenti che gli utilizzano, che siano su smartphone o dispositivi appositi, ne fanno un uso basilare, principalmente per ricerche su internet, domande veloci, riproduzione musicale, oppure per ricevere informazioni riguardanti il tempo, notizie o il traffico. Decisamente minore è la

percentuale di utenti disposti ad utilizzare gli assistenti digitali per attività che coinvolgono l'acquisto di beni o più in generale la gestione di attività monetarie. Questo è dovuto principalmente a due ragioni, la complessità stessa di alcune operazioni e soprattutto la mancanza di fiducia nella sicurezza dell'operazione, avendo paura che il dispositivo possa sbagliare.

Nonostante le suddette criticità il pensiero comune riguardo questi dispositivi è che siano la maniera più veloce, facile ed intelligente per tenersi al passo con le attività di tutti i giorni, anche se per ora con un occhio scettico per ciò che riguarda la gestione del denaro.

Con l'assistenza della società IRISCUBE Reply S.r.l. si sono analizzate le tecnologie e i metodi di integrazione ad oggi disponibili per tali dispositivi, con particolare interesse verso l'integrazione di quest'ultimi con i servizi bancari.

È stata in primo luogo analizzata l'architettura dei prodotti in questione, andando a comprendere le differenze, qualora ce ne fossero tra i dispositivi offerti dai principali produttori.

Sono state poi esaminate e valutate le principali modalità di gestione dei dati utente e dell'autenticazione dello stesso, con un'attenzione particolare proprio al tema della sicurezza, essendo il tema centrale in un contesto come quello bancario.

Proprio su questo tema anticipiamo che sono state individuate varie criticità non propriamente risolvibili con i supporti tecnologici presenti ad oggi in tali assistenti.

Tenendo a mente quest'ultime sono state proposte e implementate alcune varianti alternative alla struttura originale, in cui si è cercato di portare la sicurezza del sistema ad un livello conforme agli standard del settore, con ovvi compromessi sull'esperienza utente.

Si definiscono a questo punto due famiglie di soluzioni: alcune con un'esperienza basata totalmente sulla voce, come nell'idea iniziale, ma con gravi lacune di sicurezza, e altre con una sicurezza conforme agli standard ma con la necessità di utilizzare strumenti di supporto per garantirla.

Per terminare, verrà analizzato l'impiego di meccanismi di autenticazione basati su biometriche vocali al fine di proporre una soluzione con una sicurezza accettabile e conforme all'idea iniziale.

La struttura del presente documento prevede una prima parte in cui viene definito nel dettaglio cos'è e come funziona un assistente digitale. Viene analizzato l'utilizzo odierno di questi prodotti, con relativi vantaggi e svantaggi, per poi passare ad un'analisi più attenta dello stato dell'arte delle componenti principali che caratterizzano quest'ultimi.

Successivamente il tema della discussione vira verso quello che è il campo di applicazione oggetto di questo lavoro, ovvero il settore bancario. Verrà effettuata un'analisi dei canali ad oggi disponibili al cliente nell'interazione con questo mondo, evidenziando come molti degli svantaggi segnalati dagli utilizzatori potrebbero essere risolti tramite l'impiego di queste tecnologie. A seguire verranno anche menzionati e commentati i prodotti già commercializzati da alcuni grandi gruppi bancari, evidenziando le criticità degli stessi.

Conclude il lavoro di tesi la descrizione delle analisi e del progetto condotti in collaborazione con Iriscube Reply. In quest'ultima parte verranno evidenziati i limiti ad oggi presenti nelle architetture realizzabili e le principali tecniche individuate per porvi rimedio.

Seguirà infine il paragrafo relativo alle conclusioni, qui verranno riassunti i risultati ottenuti dal lavoro e verrà definita un'idea per un possibile sviluppo futuro del lavoro.

II. ASSISTENTI DIGITALI E DISPOSITIVI SMART

1. Cosa sono, utilizzi e benefici

Un assistente digitale, anche chiamato assistente intelligente o assistente virtuale, è un software capace di comprendere comandi in linguaggio naturale ed eseguire determinati compiti per conto dell'utente.
Gli assistenti digitali sono pensati per mimare il più fedelmente possibile la comunicazione umana, ed è proprio grazie a questa peculiarità che presentano un elevata semplicità d'uso.
Ad oggi gli assistenti digitali sono molto diffusi e integrati in quasi ogni tipo di prodotto tecnologico che già conosciamo, quali smartphones, computer, dispositivi indossabili, altoparlanti intelligenti, ecc. Diffusione soprattutto dovuta al fatto, che non vengono richieste alte capacità computazionali ai dispositivi ospitanti.
Questi dispositivi, presentano infatti delle differenze con le prime versioni di prodotti che facevano leva sulla voce. Quest'ultimi, erano prodotti molto semplici che permettevano all'utente di pronunciare un insieme di comandi estremamente ridotti e rigidi. Operavano principalmente con parole chiave, più che una vera e propria interpretazione della richiesta.
Queste limitazioni rendevano l'interazione decisamente poco fluida, non davano perciò l'impressione di avere una vera conversazione. Ad oggi, grazie anche ai recenti sviluppi tecnologici negli ambiti legati all'elaborazione in cloud e all'intelligenza artificiale, questi dispositivi presentano interfacce di conversazione decisamente migliorate.
Come già citato, quelli che oggi conosciamo come assistenti digitali, non richiedono capacità hardware di alto livello, nonostante l'interazione con l'utente sia

estremamente avanzata, questo grazie alla loro architettura software. La totalità dell'elaborazione dei dati avviene in cloud e non nel singolo dispositivo, che sarà invece responsabile solo delle attività strettamente a contatto con l'utente, quali registrazione della traccia audio contenente la richiesta e riproduzione della traccia audio di risposta.

Tra le principali tecnologie sfruttate da questi dispositivi, che sono coinvolte nell'elaborazione della richiesta, abbiamo: gli algoritmi di riconoscimento vocale, utilizzati per trascrivere testualmente il comando dell'utente; quelli di comprensione del linguaggio naturale (NLU), che permettono di trasformare la stringa di testo contenente la richiesta in un'azione comprensibile dal componente che seguirà la catena di manipolazione dei dati; e quelli di sintesi vocale, per rendere la risposta sonora.

Nella prossima sezione verranno analizzati in dettaglio tutti i componenti che si occupano di rendere possibile l'emulazione di una conversazione, andando ad osservare come il dato viene gestito dalla presa in carico della richiesta alla risposta e analizzando lo stato dell'arte per le tecniche principali.

Ad oggi, a pochi anni dalla loro nascita, questi dispositivi sono già fortemente utilizzati, probabilmente grazie all'estrema familiarità che un'interfaccia conversazionale offre, rispetto alle barriere che possono avere dispositivi come computer e smartphones.

Dal 2017 al 2022 l'utilizzo di questi dispositivi ha visto una crescita senza sosta, solo negli Stati Uniti gli adulti che interagiscono con gli assistenti digitali almeno una volta al mese sono incrementati da 79.9 milioni di utenti nel 2017 a 123.5 nel 2022, segnando un incremento di oltre il 50%. Si può notare come la tendenza di utilizzo sia quasi strettamente crescente se ipotizziamo di eliminare il rumore generato dagli utilizzi del 2020 e 2021. È ragionevole supporre infatti che in quest'ultimi a causa delle restrizioni causate dalla pandemia di COVID-19 vi sia stato un utilizzo superiore alla media, essendo molte famiglie bloccate in casa. (Figura 1)

Questo costante incremento nell'utilizzo non sembra intenzionato ad arrestarsi neanche nei prossimi anni, almeno secondo gli ultimi rapporti di Insider Intelligence Inc. [2],

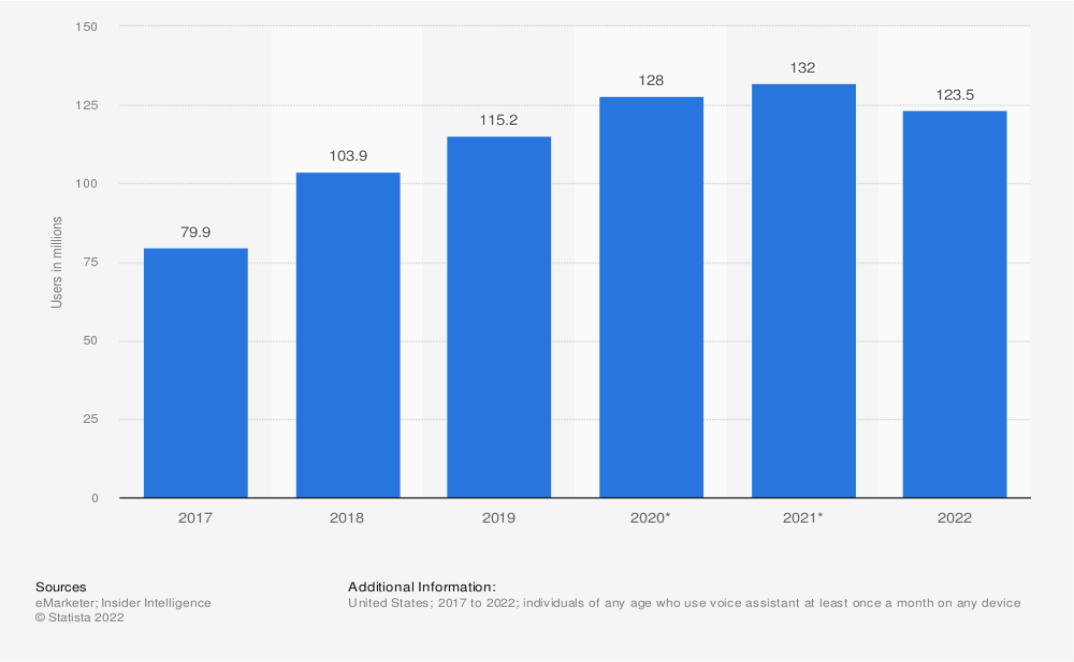

Figura 1: Numero di utilizzatori di assistenti vocali negli Stati Uniti dal 2017 al 2022. Fonte [41]

che sostiene come secondo alcune stime ci si aspetta che circa metà della popolazione adulta statunitense userà queste tecnologie entro il 2025.

Ovviamente non sono solo gli Stati Uniti gli unici interessati a questo settore, ne vediamo un utilizzo indiscriminato un po' in tutto il mondo come sostengono anche i dati sull'assistente cinese DuerOS di Baidu, in cui si nota una crescita ancora più marcata nel numero di interazioni mensili effettuate dagli utenti dal 2018 al 2021, passando da circa 400 milioni di interazioni mensili nella metà del 2018, fino ad arrivare a 6,6 miliardi all'inizio del 2021 [3].

È sicuramente una tecnologia che non si può ignorare, anche se ad oggi nonostante gli alti utilizzi potremmo definire un po' acerba. Andando infatti ad analizzare più nello specifico le interazioni degli utenti con tali dispositivi possiamo facilmente notare come la maggior parte di esse riguardi azioni semplici, tra le principali: domande veloci e aggiornamenti su notizie varie, per gli assistenti su dispositivi casalinghi; e l'avvio di chiamate e ricerche web per quelli su smartphones.

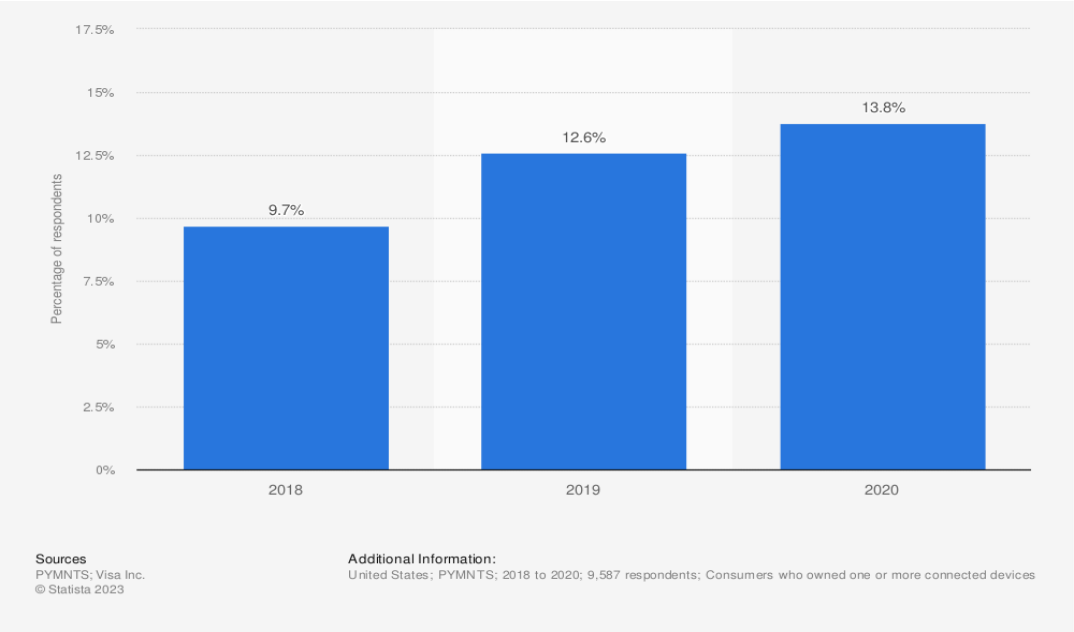

Figura 2: Percentuale di utenti ad utilizzare assistenti vocali per effettuare acquisti negli Stati Uniti dal 2018 al 2020. Fonte [42]

Si nota una significativa discrepanza tra le percentuali di utilizzo di queste funzionalità e quelle riguardanti azioni come l'acquisto di beni online o il controllo di altri dispositivi in casa, sebbene negli anni anch'esse abbiano riscontrato una tendenza positiva (Figura 2).

Le motivazioni dietro la suddetta divergenza sembrano essere prevalentemente due, la preoccupazione per la sicurezza delle operazioni e la scarsa trasparenza, da parte dei distributori di questi dispositivi, sul trattamento dei dati raccolti [1]. Preoccupazioni che sembrano essere più che fondate se si va ad analizzare la struttura di questi dispositivi, in quanto si possono riscontrare diversi punti di attacco. Anche questo è un tema che verrà trattato nello specifico nei capitoli successivi.

Nonostante alcuni scetticismi la tendenza ad utilizzare assistenti intelligenti sembra essere in costante aumento.

2. Struttura e funzionamento

Gli assistenti digitali hanno una struttura complessa, sono composti infatti da diverse componenti, che si occupano di elaborare la richiesta in diversi passaggi. Quasi la totalità dell'elaborazione non avviene nel dispositivo, bensì in cloud, lasciando al primo solo l'onere di registrare il comando dell'utente e riprodurre la risposta del sistema.

Il componente hardware che ospita l'assistente richiede solo un microfono, un altoparlante e una connessione ad Internet.

Questi dispositivi operano tramite delle parole chiave, è infatti tramite la pronuncia di una o più parole predefinite che il dispositivo si risveglia iniziando ad ascoltare e registrare la conversazione.

Una volta registrata, la richiesta viene elaborata in diverse fasi: in primo luogo deve essere inviata dal dispositivo in questione al cloud del distributore dello stesso, qui tramite algoritmi di riconoscimento vocale, la traccia viene trascritta in modo da poter essere utilizzata come input in algoritmi di comprensione del linguaggio naturale (NLU). Una volta identificata l'intenzione dell'utente, la gestione di quest'ultima viene delegata ad un altro componente chiamato gestore del dialogo.

Il gestore del dialogo è il componente che si occuperà di elaborare effettivamente il comando e generare la risposta. Questa fase non è detto che avvenga, similmente alle due precedenti, nel cloud del distributore del dispositivo. Bisogna precisare infatti, che nel caso in cui l'utente non stia interagendo con applicazioni native dell'assistente ma con applicazioni di terze parti, anche il gestore del dialogo sarà ospitato su server di terze parti, e quindi non sotto il controllo del distributore. Anche le applicazioni native potrebbero non essere gestite dallo stesso cloud che si occupa delle prime due operazioni, tuttavia, si tratterebbe in ogni caso di server sotto il controllo del distributore. È importante sottolineare questo aspetto, in quanto le applicazioni di terze parti rappresentano la maggioranza di quelle disponibili, e il fatto che dati utente possano trafficare su server in cui non si è a conoscenza dell'uso che ne viene fatto

rappresenta un primo rischio sia di sicurezza che di privacy per gli utenti, che spesso non ne sono a conoscenza.

L'elaborazione del comando utente, da parte del gestore del dialogo, non è detto che avvenga in un unico blocco. Possiamo infatti spezzare ulteriormente quest'ultimo in diverse componenti, a seconda della complessità dell'azione richiesta. Alcune delle componenti in cui è comunemente suddiviso sono: il servizio che prenderà in carico la richiesta, più comunemente noto come "Fulfillment Service", un database, logica di back-end e un servizio di autenticazione.

Tra quelli precedentemente citati solo il Fulfillment Service è obbligatorio per la creazione di un applicazioni eseguibile su un assistente digitale.

Quest'ultimo è infatti il responsabile della gestione della richiesta e dell'eventuale comunicazione con gli altri componenti, al fine di ottenere informazioni utili alla generazione della risposta desiderata.

Nel capitolo 4 avverrà una trattazione più dettagliata dei componenti coinvolti in questo blocco, in quanto con quest'ultimi uno sviluppatore ha la possibilità di interagire per manipolare le richieste al fine di popolarle con dati personalizzati.

Una volta generata la risposta, questa ritorna al cloud del distributore dell'assistente, che si occuperà di generare la traccia audio della stessa tramite algoritmi di sintesi vocale. La traccia audio di risposta tornerà poi al dispositivo con cui l'utente sta interagendo per essere riprodotta.

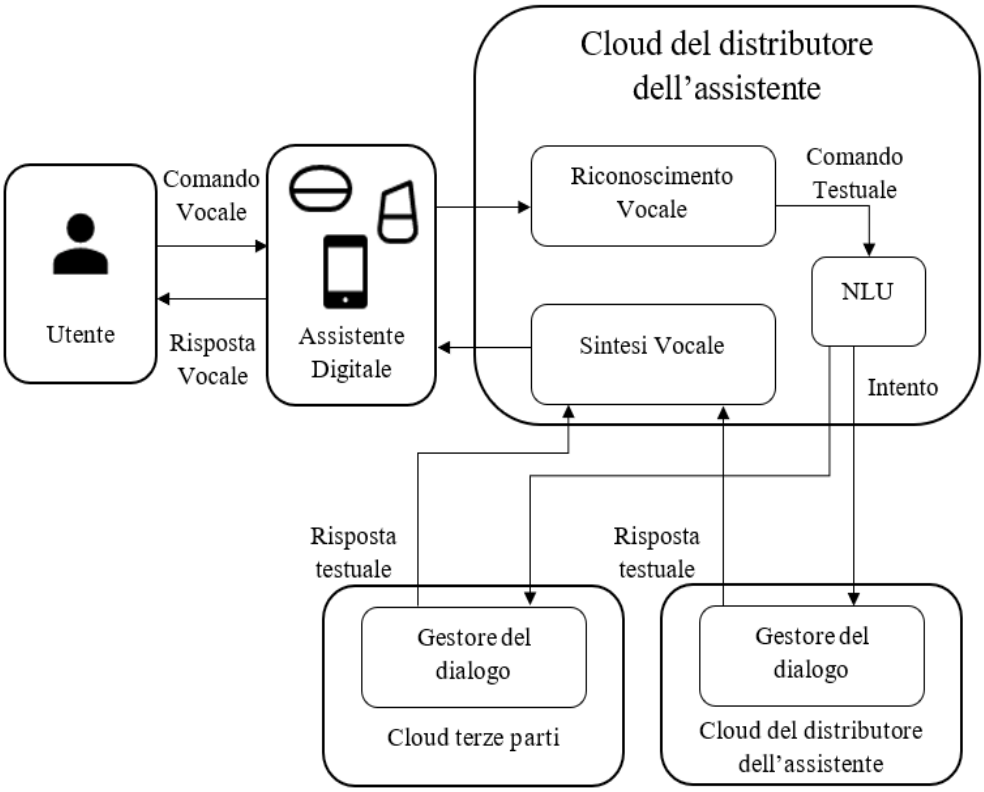

Figura 3: Visualizzazione grafica dell'elaborazione di un comando fornito ad un assistente digitale.

Si può vedere uno schema grafico volto a riassumere e semplificare tutta l'interazione precedentemente descritta. (Figura 3)

Due tra i componenti più importanti nella catena di elaborazione dei dati di un assistente digitale sono il riconoscimento vocale delle richieste (Speech-to-Text o STT) e la sintesi vocale della risposta (Text-to-Speech o TTS).

Esistono ovviamente diverse tecniche per eseguire tali procedure. In questa sezione si vuole dare una panoramica del funzionamento di questi sistemi, analizzando le tecniche considerate ad oggi lo stato dell'arte.

2.a. Speech-to-Text

Il riconoscimento vocale è l'abilità di un programma di identificare parole in un linguaggio parlato e tradurle in un formato comprensibile da una macchina. [4] [5]
Le diverse tecniche di riconoscimento vocale vengono classificate sulla base di tre parametri:

- Oratore: alcuni modelli vengono strutturati per riconoscere specifici oratori, per altri invece è indifferente;
- Espressioni riconosciute: in alcuni casi potrebbero essere in grado di riconoscere singole parole o espressioni, e non frasi complesse;
- Vocabolario: maggiore è il vocabolario a disposizione del modello, maggiori saranno la sua complessità e precisione.

Un generico modello di riconoscimento vocale si compone dei seguenti passaggi: elaborazione del segnale analogico, estrazione delle caratteristiche del segnale e classificazione di quest'ultimo (Figura 4).
L'elaborazione del segnale analogico è un passaggio preliminare all'effettiva analisi. Esso serve a digitalizzare la traccia audio e a ripulirla da suoni indesiderati, come può essere il rumore di sottofondo, al fine di migliorarne la qualità.
Successivamente, con il segnale processato e ripulito, si procede all'estrazione delle caratteristiche, che è uno dei passaggi più importanti. Durante questa elaborazione la forma d'onda in ingresso viene codificata in un insieme di parametri noti come *features*. L'obbiettivo di questo passaggio è quello di ottenere dei vettori di features quanto più compatti possibile, senza però perdere informazioni rilevanti dell'onda originale.
Le features estratte saranno l'input dell'ultimo step, quello di classificazione, insieme ai modelli acustici e ai modelli linguistici.
Il modello acustico [6] è fondamentale nel processo di classificazione, in esso infatti sono contenute le associazioni tra i segnali audio e i fonemi. I fonemi sono quelle unità

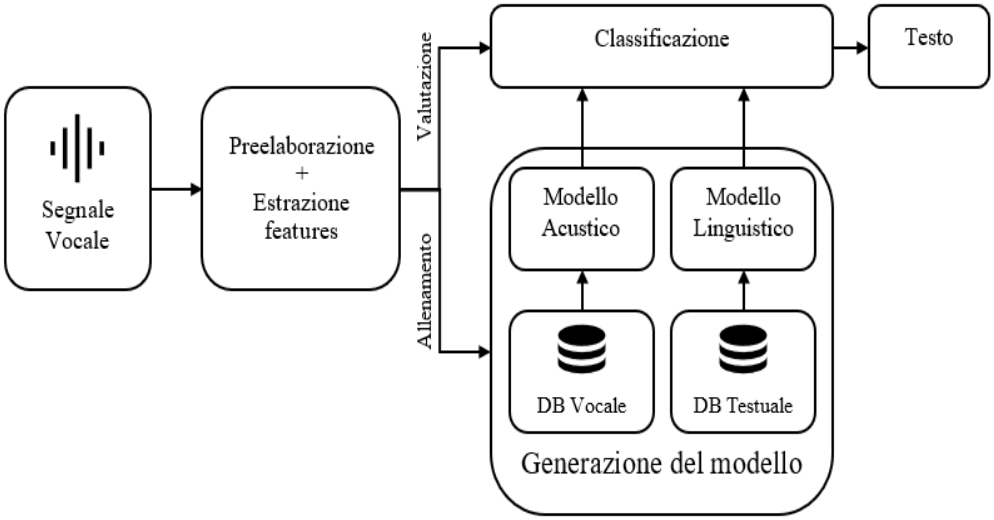

Figura 4: Architettura di un generico sistema di riconoscimento vocale

linguistiche dotate di valore distintivo, che possono perciò produrre variazioni di significato quando scambiate.

Per creare questi modelli si fa uso di software che prendendo come input delle registrazioni audio e le corrispondenti trascrizioni testuali al fine di individuare statisticamente quali suoni sono responsabili nella creazione di specifiche parole.

Il modello linguistico invece rappresenta la probabilità che una parola si trovi dopo una certa sequenza. Da un punto di vista statistico viene assegnata una probabilità alla sequenza di parole in questione ($P(parola_1, \dots, parola_n)$) dove valori prossimi allo zero indicano appunto scarse probabilità che la sequenza analizzata sia valida per il linguaggio considerato [7].

Vi sono diverse possibili tecniche per ognuno dei passaggi evidenziati, e in particolare gli algoritmi che prendono il nome di Speech-to-Text, rappresentano una particolare concatenazione di queste tecniche, a seconda dell'algoritmo scelto.

Si citano di seguito le principali tecniche di Speech-to-Text note in letteratura.

Modello di Markov Nascosto (HMM)

Il modello di Markov nascosto è uno dei più utilizzati nell'ambito del riconoscimento vocale. Nasce negli anni '70 e da allora ha visto sempre più applicazioni in quest'ambito, fino ad oggi in cui viene ancora usato come componente ausiliario di soluzioni basate su reti neurali.

Il modello in questione gestisce il sistema su cui viene applicato come un processo di Markov.

Si definisce processo di Markov un processo stocastico in cui la probabilità di transitare allo stato successivo del sistema dipende solo dallo stato attuale, e non dalla storia di stati che ci hanno condotto ad esso. Quest'ultima è nota come proprietà di Markov o condizione di assenza di memoria.

Quando lo spazio degli stati di un processo markoviano è discreto, esso prende il nome di catena di Markov.

Un modello di Markov nascosto altro non è che una catena di Markov in cui è osservabile l'evento generato da uno stato, ma non lo stato stesso.

Le assunzioni statistiche effettuate da questo modello si rivelano valide nell'ambito in questione in quanto un segnale vocale può essere visto come un segnale stazionario a tratti. La voce, infatti, per periodi di tempo sufficientemente brevi (nell'ordine di qualche millisecondo), può essere approssimata come un processo stazionario e quindi assimilata ad uno stato del modello di Markov.

Si analizza di seguito il funzionamento di questo classificatore [8].

Come già citato in precedenza il segnale vocale viene convertito in un vettore di features $Y = y_1, \dots, y_n$ che possono essere viste come le caratteristiche acustiche del tratto in questione.

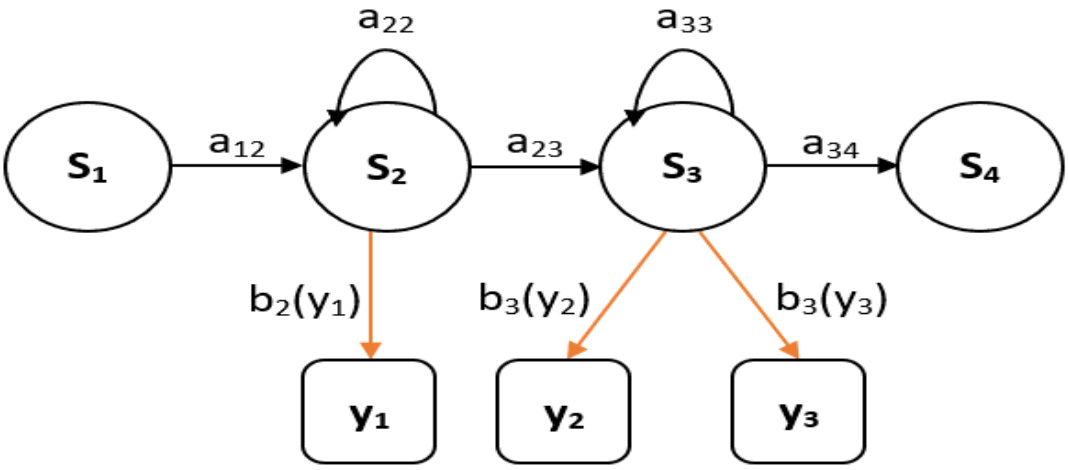

Figura 5: Schema di un HMM per la classificazione di fonemi

Lo scopo del classificatore è individuare la sequenza di parole $W = w_1, \dots, w_m$ che con maggior probabilità ha generato Y

$$W^* = arg\,max\{P(Y|W)P(W)\}$$

Per ricollegarsi ai termini usati nel paragrafo precedente $P(Y|W)$ rappresenta il modello acustico, e $P(W)$ il modello linguistico.

Ai fini della classificazione, le singole parole vengono gestite come la concatenazione di fonemi $q_{1:k}^{(w)} = q_1, \dots, q_k$ e considerando che una parola potrebbe essere pronunciata in più modi, possiamo vedere la probabilità della realizzazione Y data la sequenza di parole W, come la somma delle probabilità delle varie pronunce

$$P(Y|W) = \sum_{Q} P(Y|Q)P(Q|W)$$

Dove Q è una particolare sequenza di pronunce

$$P(Q|W) = \prod_{m} P\left(q^{(w_m)}|w_m\right)$$

Per ottenere il modello di Markov corrispondente alla sequenza Q, vengono concatenati i modelli che rappresentano i singoli fonemi q_k.

Ogni modello è caratterizzato da un insieme di stati. Ad ogni iterazione, il modello transita da uno stato s_i ad uno stato s_j con una certa probabilità a_{ij} e produce come uscita un vettore di features y_k tramite la distribuzione $b_j(y_k)$ associata allo stato.

Normalmente si utilizza una distribuzione gaussiana, però si può rifinire ulteriormente il modello considerando un insieme di gaussiane per modellare distribuzioni generiche.

Durante l'allenamento di questi modelli, si utilizza un dataset contenente campioni vocali di test con le relative trascrizioni, al fine di stimare i parametri a_{ij} e $b_j(y)$ che modellano al meglio la sequenza di campioni considerata.

Il modello risultante è quello che viene definito modello piatto o mono-fonema. Con modello piatto ci si riferisce ad un modello che rappresenta le frasi come una concatenazione di modelli di fonemi indipendenti tra di loro.

Per catturare le relazioni che sussistono tra i vari tratti vocali, vengono utilizzati modelli che non rappresentano un solo fonema, bensì il tris contenente anche il suo predecessore e il suo successore, portando ovviamente il totale teorico dei modelli rappresentabili a N^3, dove N è il numero dei fonemi del linguaggio.

Per ridurre il numero dei tris da quelli teoricamente rappresentabili, viene associato ad ogni stato del modello mono-fonema un albero decisionale, in cui, a seconda dei dati utilizzati durante l'allenamento, vengono definite alcune regole che determinano l'insieme dei fonemi validi ad essere i successori di quello in questione.

Una volta che il modello è stato allenato, il problema di trovare la sequenza W^* dato il vettore Y, può essere risolta cercando la sequenza di stati più probabile che produce come uscita i vettori $y_1, \dots, y_n$.

Nonostante lo scopo di un HMM sia quello di trovare la sequenza più probabile, nella pratica si cerca di selezionare le N migliori soluzioni, solitamente tra 100 e 1000.

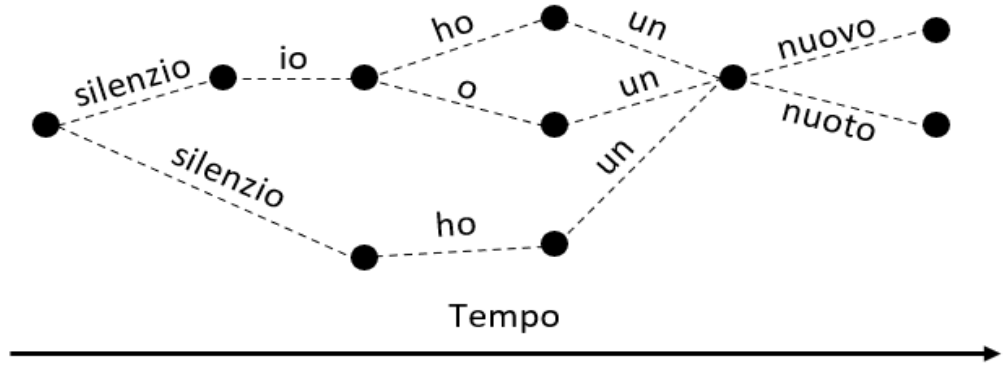

Figura 6: Reticolo di parole generato da un HMM

Il miglior approccio per immagazzinare e rappresentare il legame tra queste soluzioni è tramite un reticolo di parole. In questo reticolo ogni nodo rappresenta un instante temporale e ogni arco la parola ipotizzata.
Si preferisce questo approccio in quanto più flessibile. In caso ci fosse la necessità di rivalutare la decisione, è possibile farlo in modo dinamico, senza dover ricalcolare da zero la soluzione. Inoltre, può essere usata come input in altri sistemi di riconoscimento, permettendo di usare un HMM come componente ausiliario per migliorare le decisioni di altri modelli.

Reti neurali

Tra gli anni ’80 e ’90 le reti neurali emergono come soluzione ai problemi di riconoscimento vocale.
I primi modelli di queste reti, sebbene molto capaci nel riconoscimento di parole isolate, non erano in grado di classificare con successo attività in cui vi era un flusso vocale continuo o comunque complesso, ciò a causa della loro limitata capacità nella gestione delle dipendenze temporali.
A causa di questa limitazione, inizialmente il loro impiego era quello di tecniche di preelaborazione del segnale, prima di arrivare ad un HMM. Successivamente, con le

più recenti versioni di reti profonde (Deep Neural Networks), in particolare con implementazioni come le Reti Neurali Ricorrenti (RNN) e i Transformer, queste soluzioni hanno trovato sempre più spazio, divenendo gli attuali standard nel mercato. Le principali soluzioni che fanno uso di queste tecnologie si dividono in ibride ed end-to-end.

Le soluzioni ibride sono quelle che prevedono l'utilizzo della rete neurale solo come classificatore. In quest'ultime vengono utilizzate altre tecniche, come ad esempio un HMM, per generare il modello acustico e il modello linguistico del linguaggio. La rete neurale riceverà come input un reticolo contenente le ipotesi generate per il tratto vocale analizzato e avrà il compito di identificare la più probabile.

A differenza di quest'ultime, nelle soluzioni end-to-end tutti i componenti vengono stimati dalla rete neurale stessa, creando una relazione diretta tra ingresso e uscita.

Non necessitando di componenti esterne, queste soluzioni semplificano il processo di allenamento e di utilizzo. Per dare un'idea, un normale modello linguistico richiesto da un sistema di tipo HMM ha un peso di svariati gigabytes, rendendone praticamente impossibile l'impego su dispositivi mobili o indossabili. È a causa di ciò che tutti i moderni sistemi di riconoscimento vocale commerciali (Google, Amazon, …) sono implementati in cloud, con il conseguente rovescio della medaglia, ovvero la necessità di avere una connettività di rete per poterne usufruire.

Le due reti prima citate, ovvero RNN e Transformer, hanno trovato spazio nel settore in entrambe le modalità. Sono svariate le soluzioni proposte negli anni che vedono queste reti come protagoniste.

Le reti Transformer, in particolare negli ultimi anni, hanno ottenuto le prestazioni migliori rispetto alle RNNs quando si lavora su intere sequenze di dati. Le RNNs sono ancora preferite per elaborazioni puntuali con ritardi minimi.

Di seguito si procederà ad analizzare la struttura di queste due soluzioni e il loro funzionamento applicato al riconoscimento vocale.

Reti neurali ricorrenti

Una RNN è un particolare tipo di rete neurale profonda che si contraddistingue per la sua memoria. Queste reti nascono dall'idea che in alcune applicazioni vi sia una dipendenza temporale dei dati, e quindi, che dati elaborati in istanti di tempo precedenti possano influire sull'elaborazione dei dati attuali.

Se si parla di registrazioni vocali ovviamente quest'assunzione sussiste. La probabilità di aspettarsi in un certo istante una parola aumenta o diminuisce sulla base delle parole elaborate in precedenza [9].

Considerata la sequenza di campioni vocali $x = x_1, \dots, x_T$, una RNN calcola due sequenze: quella di vettori nascosti $h = h_1, \dots, h_T$ e quella d'uscita $y = y_1, \dots, y_T$.

Il calcolo di queste sequenze avviene ricorsivamente. Il vettore nascosto ottenuto nell'istante di tempo precedente (h_{t-1}), influenza il risultato di quello all'istante corrente (h_t).

Nel calcolo di h_t, l'entrata corrente e lo stato precedente vengono pesati diversamente al fine di produrre il nuovo stato della rete. I pesi attribuiti a quest'ultimi sono definiti tramite alcune matrici opportunamente valorizzate durante la fase di allenamento (rispettivamente W_{xh} *e* W_{hh})

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

Una volta valorizzato lo stato corrente si procede al calcolo dell'uscita corrispondente utilizzando una versione pesata di quest'ultimo tramite W_{hy}, un'altra matrice di pesi definita in allenamento

$$y_t = W_{hy}h_t + b_y$$

Si procederà ricorsivamente in questo modo fino al termine della sequenza d'ingresso.

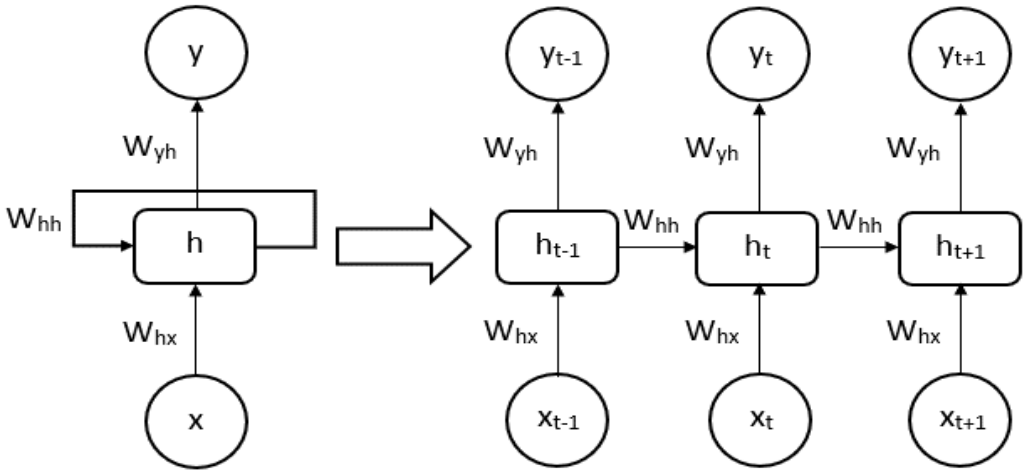

Figura 7: Rappresentazione grafica di una RNN standard, la variante a destra mostra il funzionamento in funzione del tempo

Le varianti più comunemente utilizzate di RNNs sono le reti Long Short Term Memory (LSTM). Queste reti eliminano il problema che le RNNs hanno con le dipendenze a lungo termine. Come si può intuire dalla descrizione matematica precedente, nelle RNNs lo stato corrente viene fortemente influenzato da quello che è successo nel recente passato. È altamente probabile, infatti, che informazioni di lungo termine vengano perse.

Per ovviare a questo problema, nelle LSTM vengono introdotte specifiche funzioni, anche chiamate cancelli (dall'inglese "gate"). Questi cancelli hanno il compito di bilanciare opportunamente gli ingressi per trattenere più informazione utile possibile nel tempo e compongono quella che viene definita la cella della rete. Essi sono:

- Input gate
- Forget gate
- Output gate

L'input gate seleziona dall'ingresso i valori più utili. Ovvero la parte di input che dovrebbe essere usata per cambiare lo stato della rete

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$

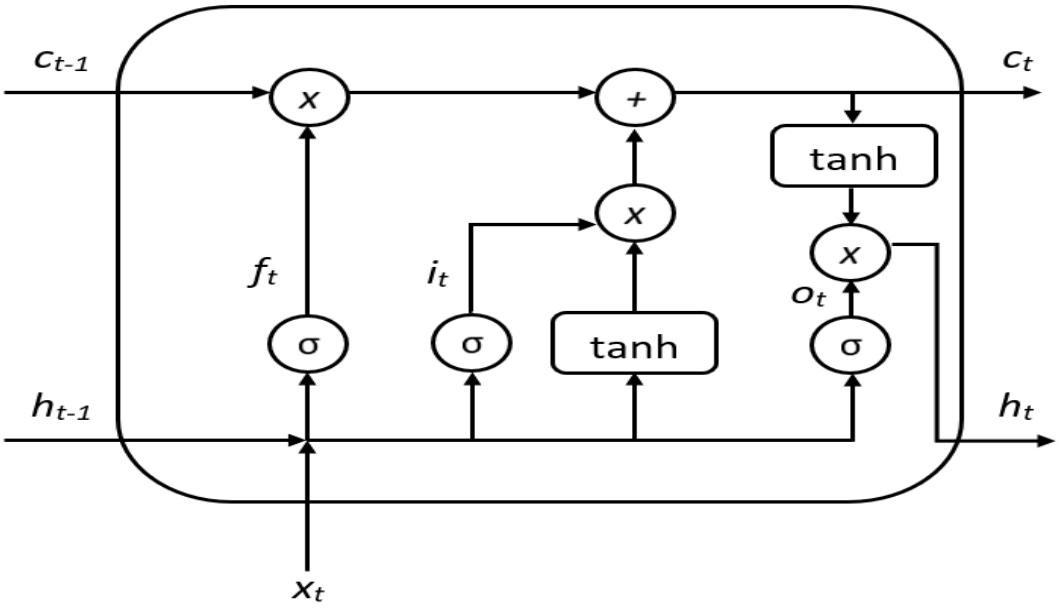

Figura 8: Rappresentazione grafica della cella di una rete LSTM

Il forget gate, come intuibile dal nome, determina quale parte dell'attuale stato della cella deve essere dimenticata, o parzialmente dimenticata. L'uscita della funzione sigmoidea oscilla tra 0 e 1. Ci basterà quindi moltiplicare poi il vettore f_t per lo stato della cella per determinare cosa dimenticare e cosa no

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$

L'output gate, infine, sulla base dello stato della cella, dell'ingresso e dello stato nascosto precedente determina il nuovo stato nascosto. Il nuovo stato verrà sia propagato nella rete per le elaborazioni successive, sia usato per il calcolo dell'uscita

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$

Ognuno dei tre cancelli, oltre ai termini visti in precedenza, riceve in ingresso il vettore che descrive lo stato della cella. Nel caso di input e forget gate si parla dello stato precedente, per l'output gate dello stato corrente.

Per calcolare lo stato corrente della cella, come già citato, vengono usati i vettori forniti dall'input e dal forget gate

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c)$$

Una volta calcolato il nuovo stato interno della cella, e aver deciso come soppesarlo agli input correnti attraverso l'output gate, possiamo calcolare lo stato nascosto della rete all'istante corrente

$$h_t = o_t \tanh(c_t)$$

Il problema principale di queste tipologie di rete (RNN e varianti) è che non possono essere usate direttamente sulla sequenza dei dati da riconoscere, almeno per come sono definite nella loro forma standard.

La funzione di stato di quest'ultime è definita separatamente per ogni punto dell'ingresso, con la conseguenza che le uscite nei vari istanti temporali non possono essere assimilate alle etichette che cerchiamo (parole o fonemi), salvo una particolare segmentazione degli ingressi.

A causa di questa grande limitazione, per anni questi modelli sono stati usati nella loro forma ibrida. Vi era la necessità di avere uno strato che antecedesse la rete e segmentasse a dovere gli ingressi, e un altro che elaborasse le uscite per fornire un'interpretazione di esse coerente con le etichette cercate.

Come già citato l'approccio più usato era quello in combinazione con un HMM.

L'approccio ibrido, oltre a non sfruttare il pieno potenziale delle RNN, andava ad ereditare i difetti di un sistema HMM, principalmente per le assunzioni statistiche sugli ingressi.

Al fine di risolvere i suddetti problemi, sono nati nuovi approcci che permettevano l'allenamento di queste reti come fossero un unico blocco, rimuovendo la necessità del sistema ibrido (sistemi end-to-end).

Senza dubbio l'approccio più usato negli attuali sistemi di riconoscimento vocale basati su RNN è il CTC (*Connectionist Temporal Classification*) [10] [11].

L'idea è quella di interpretare le uscite della rete come una distribuzione di probabilità su tutte le possibili etichette disponibili data una certa sequenza di ingressi.
Lo scopo sarà quindi quello di allenare il modello su campioni noti di sequenze, al fine che la distribuzione di probabilità in uscita dalla rete sia quanto più simile possibile a quella desiderata.
Da un punto di vista più formale, considerato X l'insieme delle possibili sequenze di ingresso e Z l'insieme delle possibili sequenze di etichette, il nostro scopo sarà allenare il classificatore $h : X \rightarrow Z$ su un insieme noto di coppie di sequenze (x, z), al fine di minimizzare l'errore tra le etichette predette e quelle reali. L'insieme delle coppie (x, z) verrà definito come S in seguito.
Come misura dell'errore commesso viene utilizzato il LER (*Label Error Rate*)

$$LER(h, S) = \frac{1}{Z} \sum_{(x,z) \in S} ED(h(x))$$

Il LER definisce l'errore commesso in funzione della media, sul numero di etichette, dell'Edit Distance (ED). Dove con Edit Distance ci riferiamo al numero minimo di operazioni necessarie per cambiare l'etichetta predetta da h in quella reale.
Ciò che rimane da definire per allenare una rete CTC, è come trasformare le uscite di quest'ultima in una distribuzione di probabilità sulla sequenza di etichette.
Si procede quindi a definire l'uscita della rete come un vettore di probabilità con tante celle quante sono le etichette disponibili, più alcune extra.
I valori delle prime celle sono interpretati come la probabilità di osservare la corrispondete etichetta al tempo considerato. I valori delle celle extra sono invece la probabilità di osservare la mancanza di un'etichetta in quell'istante.
Da un punto di vista più formale, considerata x la nostra sequenza di ingresso lunga T e w il vettore rappresentante i pesi della nostra rete, definiamo l'uscita della rete come

$y = N_w(x)$, dove $N_w(x)$ rappresenta la funzione che mappa gli ingressi x alle uscite y considerati i pesi w.

Il valore y_k^t è interpretato come la probabilità di osservare l'etichetta k nell'istante t, permettendoci quindi di definire la distribuzione di probabilità delle etichette sull'intera sequenza x come

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t$$

Dove π rappresenta un elemento dell'insieme L'^T. Con L' definito come l'insieme di tutte le etichette con l'aggiunta dell'elemento rappresentante l'assenza di etichetta.

L'ultimo passaggio rimasto è quello di creare la mappa tra l'insieme espanso di etichette e quello originale, per rimuovere le ripetizioni e le etichette vuote

$$B : L'^T \rightarrow L^{\leq T}$$

Si può così definire la probabilità di una certa sequenza di etichette $l \in L$ come la somma delle probabilità di tutte le sequenze di L' ad essa corrispondenti

$$p(l|x) = \sum_{\pi \in B^{-1}(l)} p(\pi|x)$$

Lo scopo del classificatore sarà quindi quello di scegliere la sequenza di etichette tale da massimizzare il termine $p(l|x)$.

Le RNNs di questo tipo, operando su intere sequenze di ingressi e non sui singoli campioni, possono essere facilmente estese per sfruttare anche le informazioni che ci possono fornire gli elementi in istanti di tempo futuri.

Nel caso in cui una rete neurale ricorrente utilizzi sia le informazioni passate che quelle future per l'uscita in un determinato istante, essa prende il nome di Bidirectional Recurrent Neural Network (BRNN).

Nella versione standard delle RNNs, quindi non CTC, questa tecnica non è troppo usata, in quanto sarebbe necessario ascoltare tutta la sequenza di ingresso prima di iniziare l'elaborazione. Nel caso delle reti CTC, la sequenza di ingresso è già stata segmentata sulla base di particolari tratti nello spettro del segnale, quindi ci possiamo permettere di elaborarla in entrambe le direzioni senza rinunciare al tempo di reazione che contraddistingue questa tipologia di reti.
Nelle reti BRNN per ogni istante di tempo, vengono definiti due stati nascosti, uno che conterrà le informazioni passate e uno quelle future

$$h_t^{\rightarrow} = H(W_{xh^{\rightarrow}} x_t + W_{h^{\rightarrow}h^{\rightarrow}} h_{t-1}^{\rightarrow} + b_{h^{\rightarrow}})$$

$$h_t^{\leftarrow} = H(W_{xh^{\leftarrow}} x_t + W_{h^{\leftarrow}h^{\leftarrow}} h_{t-1}^{\leftarrow} + b_{h^{\leftarrow}})$$

Entrambe le informazioni dello stato poi concorreranno alla definizione dell'uscita

$$y_t = W_{h^{\rightarrow}y} h_t^{\rightarrow} + W_{h^{\leftarrow}y} h_t^{\leftarrow} + b_y$$

Un esempio reale di una rete CTC BRNN lo si può trovare nella funzione di dettatura vocale della Gboard, la tastiera ufficiale di Google [12].
In questo prodotto infatti è implementata un tipo di rete che prende il nome di RNN Transducer [13]. Quest'ultima, altro non è che la composizione dell'appena vista CTC BRNN, che prende il nome di *"rete di trascrizione"* e di un'altra RNN standard chiamata *"rete predittiva"*.
La rete predittiva è una LSTM che cerca di prevedere la prossima etichetta della sequenza basandosi sulle precedenti. È usata per raffinare le uscite della rete di trascrizione, che come già visto, si occupa di classificare gli ingressi sulla base delle loro caratteristiche acustiche.
Si può vedere la rete di trascrizione come la responsabile del modello acustico del classificatore e la rete predittiva come la responsabile di quello linguistico.

Reti Transformer

Le reti transformer nascono nel 2017 da un paper pubblicato da Google *"Attention is all you need"* [14]. Il loro obbiettivo è quello di riconoscere legami nei dati d'ingresso senza ricorrere ad operazioni in sequenza sugli stessi (come le RNN).

L'idea alla basa di esse, e intorno a cui ruota tutta la loro architettura, è quella della *"self-attention"*.

Quest'ultima è la responsabile di individuare le relazioni nella sequenza di ingresso. Data la sequenza di ingresso $x_1, x_2, \dots, x_n$ genererà l'uscita $y_1, y_2, \dots, y_n$ corrispondente, in cui il vettore y_i rappresenta quanto l'i-esimo elemento sia decisivo nella sequenza in questione. La definizione di cosa sia decisivo e cosa no, dipende da come il modello viene allenato.

Nonostante possa apparire come qualcosa di complesso, l'operazione eseguita in uno strato di self-attention è un semplice prodotto scalare. Per quanto semplice, non è banale comprendere il perché quest'operazione funzioni in un contesto simile. Per spiegarlo con un esempio, supponiamo il seguente problema: vogliamo consigliare degli alimenti ad un utente sulla base dei suoi gusti alimentari.

La soluzione del problema sta proprio nel definire il legame tra un utente e un alimento. Per farlo possiamo pensare di codificare ogni alimento sulla base delle sue caratteristiche. Se consideriamo una mela, il vettore corrispondente avrà un valore alto nell'elemento che indica la dolcezza, nullo in quello che indica la piccantezza, ecc.

A questo punto se l'utente viene codificato come un vettore rappresentante le sue preferenze alimentari, il prodotto scalare di quest'ultimo, con i vettori rappresentanti i vari alimenti, ci fornirà una serie di punteggi che indicano la compatibilità dell'utente con l'alimento in questione.

Da un punto di vista matematico, le preferenze dell'utente faranno da peso alla corrispondente caratteristica dell'alimento. Se il segno di una delle caratteristiche combacia tra il vettore dell'utente e quello dell'alimento, essa contribuirà positivamente al punteggio.

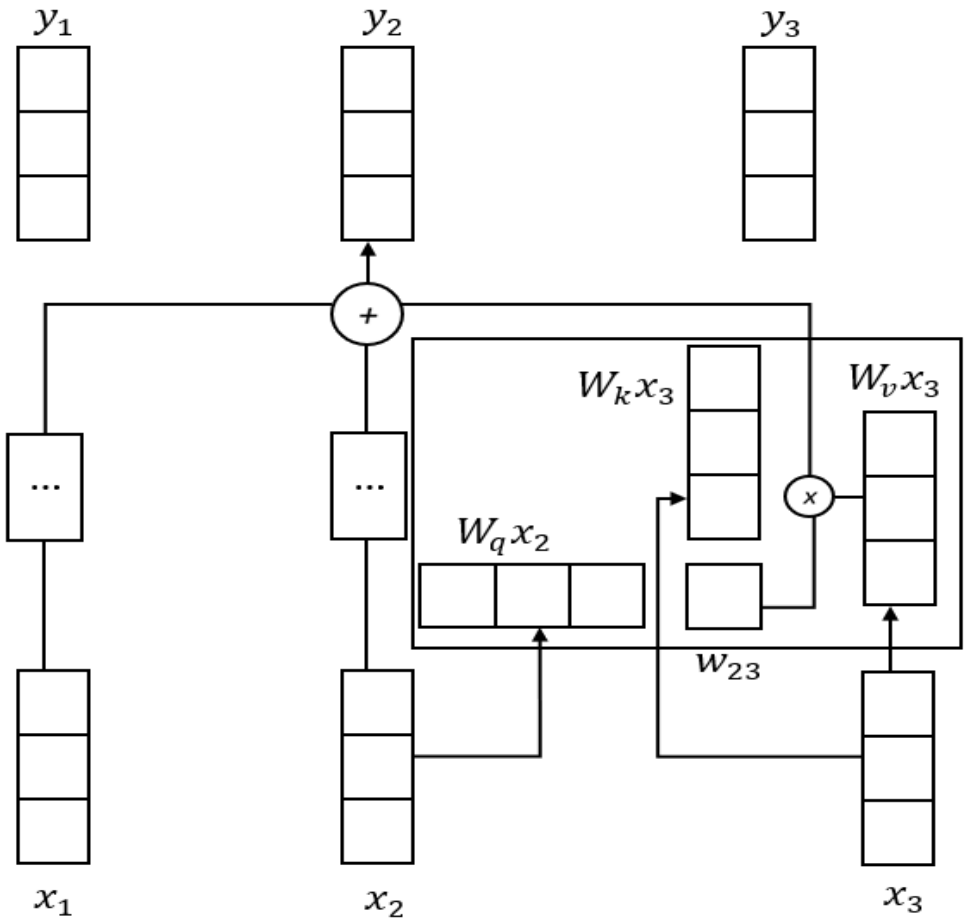

Figura 9: Rappresentazione grafica del funzionamento di uno strato di self-attention che utilizza le matrici di query, key e value

Ovviamente, il passaggio non banale in un problema di questo tipo è proprio l'annotazione delle caratteristiche degli alimenti e dell'utente. Quello che accade nella pratica quando si modella un problema come questo, è che la scelta delle caratteristiche degli ingressi verrà definita durante l'allenamento, le caratteristiche diverranno quindi i parametri del modello stesso.

Lo strato di self-attention utilizza gli ingressi in tre modi differenti. I tre ruoli che ogni ingresso ricopre sono:

- Ogni ingresso viene confrontato con tutti gli altri per stabilire i pesi che definiranno la sua uscita y_i

- Ogni ingresso viene confrontato con tutti gli altri per stabilire i pesi che definiranno l'uscita del j-esimo vettore y_j
- Dopo aver stabilito i pesi, ogni ingresso viene usato nella somma pesata che definirà tutti i vettori di uscita

Questi tre utilizzi in letteratura sono noti con nomi specifici, rispettivamente: query, key, value.

Il modello sarà responsabile di stimare le tre matrici W_q, W_k, W_v che consentano la trasformazione degli ingressi nei tre vettori utili a ricoprire i suddetti ruoli.

$$q_i = W_q x_i \qquad k_i = W_k x_i \qquad v_i = W_v x_i$$

Dopo aver calcolato i vettori query, key e value, lo strato di self-attention ha tutto il necessario per calcolare le sue uscite.

Come già detto il vettore query del i-esimo elemento, viene moltiplicato con tutti gli altri ingressi per definire i pesi dell'i-esima uscita. Ovviamente in questo calcolo tutti gli elementi che non siano x_i parteciperanno tramite il loro vettore key (ruolo 2)

$$w'_{ij} = q_i^T k_j$$

Considerato che il prodotto scalare restituisce un valore in un intervallo illimitato, il vettore di pesi w'_i viene rielaborato tramite la funzione softmax. Ognuno dei pesi w'_{ij} viene ricalcolato come

$$w_{ij} = \frac{\exp w'_{ij}}{\sum_j \exp w'_{ij}}$$

Questo fa sì che ogni peso sia contenuto nell'intervallo [0,1] e che la somma degli stessi sia unitaria.

Definiti tutti i pesi resta solo il terzo dei tre ruoli menzionati. Si calcoleranno i vettori di uscita tramite una somma pesata degli ingressi, dove il vettore dei pesi per l'i-esimo

elemento sarà quello appena calcolato e il contributo degli altri ingressi sarà quello definito dal corrispettivo vettore value

$$y_i = \sum_j w_{ij} v_j$$

Nella versione dello strato di self-attention definita in [14], il calcolo dei pesi non avviene tramite semplice prodotto scalare, bensì tramite una versione scalata di esso. Questo in quanto la funzione di softmax è sensibile ad input particolarmente grandi, con conseguente degrado delle prestazioni e rallentamento della fase di allenamento. Per evitare questi input, il prodotto scalare dei pesi viene diviso per la radice della dimensione degli ingressi

$$w'_{ij} = \frac{q_i^T k_j}{\sqrt{k}}$$

Il motivo principale per cui si preferisce il prodotto scalare tra le diverse operazioni disponibili, risiede nelle maggiori prestazioni. Ognuno dei calcoli visti finora infatti, invece di essere effettuato per ogni vettore singolarmente può facilmente essere espresso come un unico prodotto tra matrici.

Allo stato attuale, lo strato di self-attention definito finora data una serie di parole in ingresso stabilisce quale siano le più rilevanti nei confronti dell'intera sequenza. Spesso una parola assume ruoli diversi nei confronti dei suoi vicini. Se si pensa alla frase "Marco ha baciato Sara", il verbo *ha baciato* è legato diversamente agli altri componenti della frase. *Marco* sta compiendo l'azione descritta dal verbo, *Sara* la riceve. Allo stato attuale, usando un solo strato di self-attention si stanno sommando insieme queste due informazioni.

In un'implementazione moderna, per dare più flessibilità al modello si utilizzano più strati di self-attention in parallelo, ottenendo quella che viene definita *attenzione multi-testa.* Ogni strato avrà le sue matrici per definire i ruoli di query, key e value.

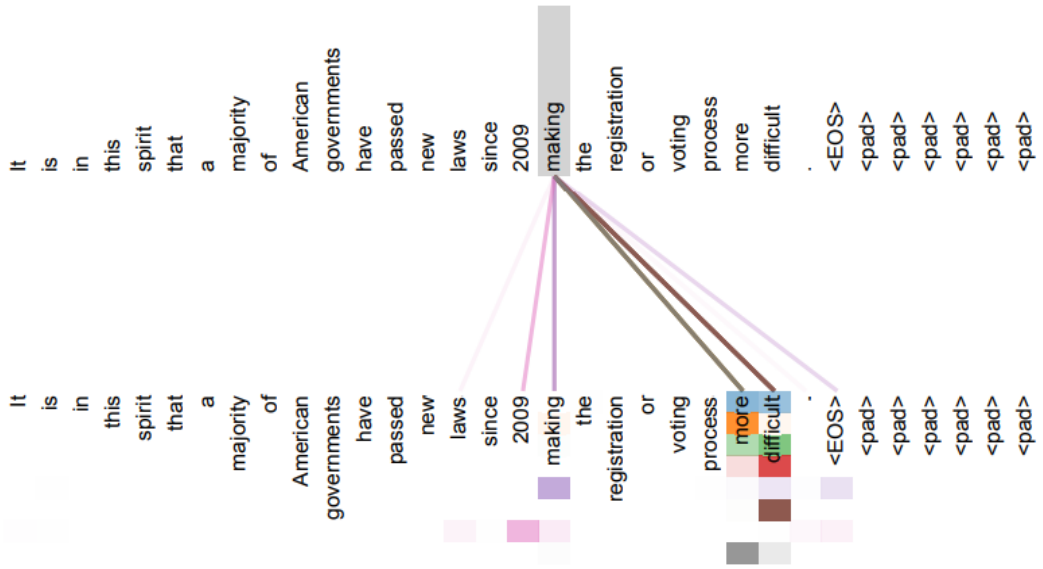

Figura 10: Visuallizazione grafica di come le teste di un Transformer identifichino le relazioni a distanza nella seuqenza. Si può notare come l'uscita per il verbo "making" ha identificato forti relazioni con "laws", "2009", "more" e "difficult", andando a delineare le informazioni importanti che lo riguardavano. Ogni colore nel disegno rappresenta l'uscita di una delle teste del modello. Fonte [14].

Sebbene in una rete Transformer lo strato di self-attention sia probabilmente il più importante, non è l'unico a comporne l'architettura.

L'architettura di queste reti si può dividere in due grandi blocchi, un blocco di *codificatori* e uno di *decodificatori*.

I codificatori sono una serie di blocchi identici, che si susseguono in sequenza. L'entrata di un codificatore corrisponde all'uscita del precedente.

La loro struttura si compone di uno strato di self-attention, descritto precedentemente, e una piccola rete neurale di tipo *Feed Forward.* L'unica parte della rete in cui gli ingressi hanno modo di condizionarsi a vicenda è lo strato di self-attention, nella rete di Feed Forward ogni vettore di ingresso segue il proprio percorso, permettendo l'esecuzione in parallelo di questo passaggio.

Come detto ogni codificatore utilizza come ingresso l'uscita del precedente. Il primo della serie invece, riceve gli ingressi originali opportunamente elaborati. Ad ogni

vettore di ingresso viene anche aggiunta un'informazione di posizione. Questa sarà utile al modello per stabilire l'ordine delle parole nella sequenza e la distanza tra esse. Una volta completata la serie di elaborazioni del blocco di codificatori, l'uscita dell'ultimo viene usata come ingresso per tutti i decodificatori.

La struttura dei decodificatori è molto simile a quella dei codificatori con la differenza che tra lo strato di self-attention e la rete feed forward vi è uno strato di attenzione che aiuta il decodificatore a selezionare le informazioni rilevanti dell'ingresso.

La fase di decodifica avviene in un modo similare ad una RNN. Infatti, il blocco di decodificatori emetterà un'uscita alla volta, ognuna legata ad uno degli ingressi. L'uscita in un certo istante viene poi fornita nuovamente in input al primo decodificatore, insieme alle uscite dei codificatori per calcolare la successiva uscita della sequenza.

Nell'ambito del riconoscimento vocale, così come per le RNNs, questi modelli hanno trovato spazio sia con applicazioni ibride che end-to-end.

Tra le soluzioni ibride troviamo ad esempio quella proposta da Facebook [15]. Questa soluzione prevede di avere un classificatore in cui il ruolo di codificatore acustico viene svolto dal Transformer.

Il Transformer codifica la sequenza audio di ingresso in dei vettori contenenti informazioni di alto livello. Le nuove informazioni vengono usate per definire un HMM che, combinato con informazioni derivanti da un modello linguistico, definirà il reticolo di parole con le varie ipotesi.

Come si può notare l'elaborazione del dato è similare ai modelli ibridi descritti in precedenza, la novità sta nell'uso preliminare della rete Transformer.

La struttura del Transformer descritto in questo paper è praticamente quella standard. Le uniche variazioni effettuate riguardano le informazioni di posizione. Nella definizione originale di un Transformer [14] le informazioni di posizione usate sono assolute, perché pensato per applicazioni di NLP (*Natural Language Processing*). Nell'elaborazione di segnali vocali è stato ritenuto più significativo focalizzarsi sulle informazioni di posizione relative. La soluzione proposta prevede di usare una piccola rete di tipo CNN (*Convolutional Neural Network*) prima del blocco di codificatori per

codificare implicitamente le informazioni di posizione relative della sequenza nei vettori d'ingresso.

Nonostante le soluzioni ibride siano preferite per la loro flessibilità, non sfruttano il pieno potenziale di questi modelli. Spostiamo quindi l'attenzione sulle soluzioni end-to-end. Tra queste, come per le RNNs, abbiamo l'integrazione dei Transformer con la tecnica di *Connectionist Temporal Classification* [16] e quella che prende il nome di Tranformer Transducer [17].

L'integrazione della *Connectionist Temporal Classification* nella struttura del Transformer avviene nella fase di decodifica. Le uscite dei codificatori vengono usate contemporaneamente sia come ingressi per i decodificatori che per la parte di CTC.

Quest'ultima crea degli allineamenti espliciti tra le caratteristiche dei vettori di ingresso e la trascrizione letterale dei caratteri. Unendo opportunamente queste informazioni a quelle dei decodificatori si migliora la convergenza della rete.

La struttura proposta in [16] prevede che ogni codificatore calcoli la sua uscita come

$$X_i' = X_i + SelfAttention(X_i)$$

$$X_{i+1} = X_i' + FeedForward(X_i')$$

Dove i è l'indice del codificatore e X_i l'ingresso dello stesso.

L'ingresso del primo codificatore è, come nel caso standard, rappresentato dai vettori di ingresso con l'aggiunta delle informazioni posizionali.

I decodificatori, ricevuta l'uscita dell'ultimo codificatore (X_c), modelleranno la probabilità di vedere la sequenza d'uscita Y dato l'ingresso X_c. Dove Y rappresenta la sequenza di caratteri riconosciuti $Y[1], \dots, Y[u]$.

$$p_{dec}(Y|X_c) = \prod_u p(Y[u]|Y[1:u-1], X_c)$$

Il blocco di CTC invece, cerca di trovare un allineamento esplicito tra i singoli elementi del vettore X_c e i caratteri dell'uscita.

$$p_{ctc}(Y|X_c) = \prod_t C[t, \pi[t]]$$

Dove $C[t, \pi[t]]$ è la probabilità che il carattere $\pi[t]$ sia allineato all'elemento t del vettore X_c. Per il calcolo di C vengono definiti la matrice di pesi W^{ctc} e il vettore di costanti b^{ctc} stimati dal modello durante l'allenamento

$$C = softmax(X_c W^{ctc} + b^{ctc})$$

L'uscita finale del modello è data da una somma pesata dei risultati del blocco di decodificatori e di quello CTC, a cui viene ulteriormente aggiunta la conoscenza di un modello linguistico

$$\hat{Y} = \arg max \{\lambda \log p_{dec}(Y|X_c) + (1-\lambda) p_{ctc}(Y|X_c) + \gamma \log p_{ml}(Y)\}$$

Dove λ e γ sono dei parametri extra da definire opportunamente, che indicano rispettivamente il peso del CTC e quello del modello linguistico.

La seconda soluzione end-to-end che sfrutta la rete Transformer per applicazioni di riconoscimento vocale in tempo reale è quella che prende il nome di Transformer Transducer.

Esattamente come per un RNN-T (Recurrent Neural Network Transducer), si cerca di creare una corrispondenza tra una sequenza di vettori audio di ingresso e le etichette di uscita, per intervalli temporali fissati. Dove nelle etichette di uscita è anche presente l'etichetta nulla, per indicare l'assenza di corrispondenze per il tratto analizzato.

È un processo simile a quello che avviene in una rete CTC, con la differenza che la distribuzione dell'uscita è anche condizionata dalle etichette precedenti.

La definizione formale delle uscite del modello sarà quindi

$$P(z|x) = \prod_i P(z_i|x, t_i, Labels(z_{1:(i-1)}))$$

Dove z rappresenta la sequenza di etichette di uscita e $Labels(z_{1:(i-1)})$ rappresenta la sequenza delle precedenti etichette non nulle.

Ciò che differenzia le reti transducer dalle altre, è che la definizione della distribuzione di uscita del modello tiene conto sia del vettore di ingresso corrente sia della sequenza di uscite precedenti per la definizione dell'uscita corrente.

Si identificano in queste reti due blocchi caratteristici che svolgono le mansioni menzionate, e sono:

- Codificatore audio: il blocco che, data la sequenza di vettori audio in ingresso, identifica l'etichetta più probabile in un certo istante, sulla base delle caratteristiche acustiche dell'ingresso corrente e dei precedenti;
- Codificatore delle etichette: il blocco che, predice su base statistica, l'etichetta all'istante corrente basandosi sulla sequenza di etichette predette negli istanti di tempo precedenti.

L'uscita del modello tiene conto delle decisioni dei due blocchi pesate in modo opportuno

$$Joint = FeedForward\left(CodificatoreAudio_{t_i}(x)\right) + FeedForward(CodificatoreEtichette(Labels(z_{1:(i-1)})))$$

$$P\left(z_i \middle| x, t_i, Labels\left(z_{1:(i-1)}\right)\right) = Softmax(FeedForward(tanh\,(Joint))))$$

In una rete RNN-T la parametrizzazione di $P(z|x)$ avviene tramite due blocchi di reti LSTM. Nel lavoro proposto da Google [17], si è invece deciso di rimpiazzare

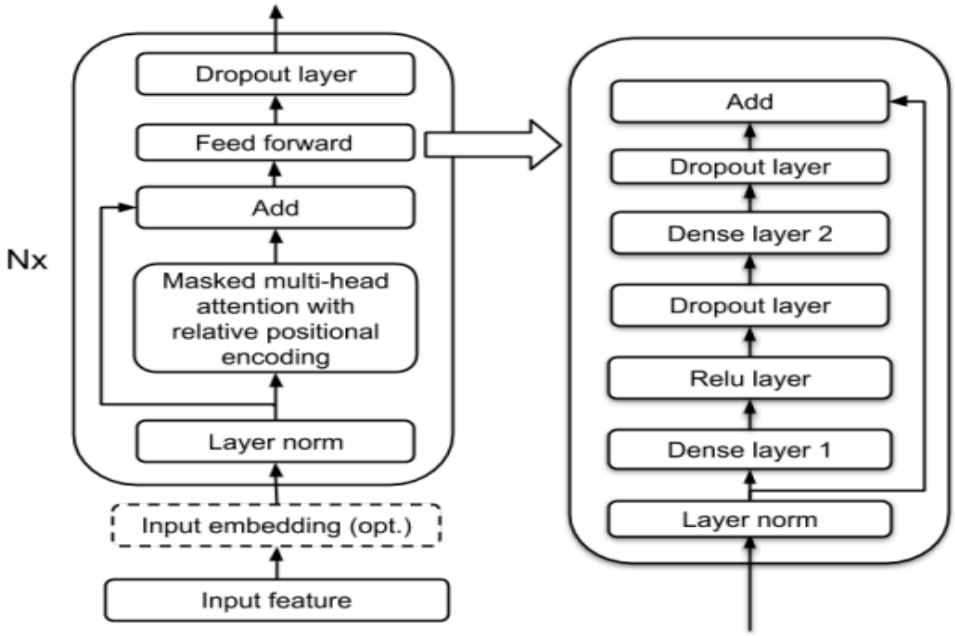

Figura 11: Architettura della rete Transformer usata in [17] come codificatore audio e codificatore delle etichette.

quest'ultime con delle reti Transformer, da qui il nome Transformer Transducer (T-T). Come per le reti CTC, l'introduzione dell'etichetta nulla rende necessario un passaggio extra per definire la probabilità della vera sequenza di etichette. Quest'ultima sarà la somma delle probabilità di tutte le sequenze predette che possono essere assimilate ad essa a seguito della rimozione dell'etichetta nulla.

Calcolare questo termine in modo esaustivo, tenendo conto di tutte le combinazioni, sarebbe un problema non trattabile nella pratica. Per farvi fronte, in [13], si rappresenta la distribuzione d'uscita come un reticolo, e si definisce $\alpha(t, u)$ come la probabilità di avere i primi u elementi come uscita al tempo t.

Movimenti orizzontali nel reticolo rappresentano la probabilità di emettere un'etichetta nulla nel passaggio da t a $t + 1$, movimenti verticali invece, indicano la probabilità di avere l'elemento $u + 1$ come uscita.

Così facendo possiamo definire la funzione da minimizzare nel nostro modello come

$$loss = -\sum_{i} \log P(y_i|x_i) = -\sum_{i} \alpha(T_i, U_i)$$

Dove T_i rappresenta la lunghezza della sequenza di ingresso e U_i la sequenza di etichette d'uscita per l'i-esimo elemento.

2.b. Text-to-Speech

Il Text-to-Speech è il processo in cui un input testuale viene analizzato, elaborato e compreso, al fine di poterlo convertire in un audio digitale che successivamente verrà riprodotto [5].
Un sistema di TTS solitamente si compone di due parti, front-end e back-end [18].
Nella parte di front-end avviene la normalizzazione del testo, la sua trascrizione fonetica e la divisione in unità prosodiche.
La normalizzazione, anche detta tokenization, è la pratica per cui il testo grezzo contenente numeri, simboli e abbreviazioni, viene convertito in parole scritte per esteso.
La trascrizione fonetica o text-to-phoneme, si occuperà poi di assegnare i rispettivi fonemi alle varie parole e successivamente il testo verrà rielaborato in proposizioni, frasi e periodi.
La trascrizione fonetica unita alle informazioni relative a intonazione e durata di ogni fonema (generate dall'analisi prosodica), definiscono la rappresentazione linguistica del testo e corrisponde all'output del front-end.

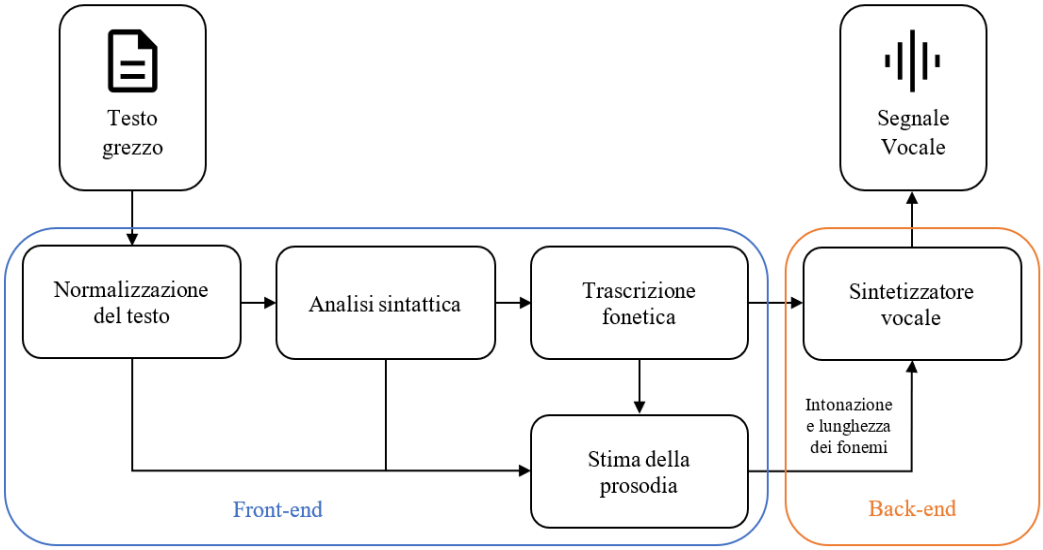

Figura 12: Architettura di un generico sistema di sintesi vocale

Il back-end riceverà come input la rappresentazione linguistica e avrà il compito di convertirla in suono, per questo di solito è chiamato sintetizzatore.
Il sintetizzatore è la parte più importante di un sistema TTS, in quanto definisce la naturalezza e l'intelligibilità dell'uscita. Dove la naturalezza rappresenta quanto l'uscita sia simile ad una vera voce umana, e l'intelligibilità quanto sia comprensibile. Le due tipologie di sintesi più usate sono, quella basata su concatenazione e quella per formanti, ognuna con i suoi pro e contro.

Sintesi basata su concatenazione

In questa tecnica la voce sintetizzata viene formata tramite la concatenazione di pezzi di registrazioni più piccole. Il suono generato con questo approccio è il più naturale tra i vari sintetizzatori; tuttavia, non è infrequente la possibilità di notare sovrapposizioni o piccoli errori in alcune concatenazioni.

Le sintesi per concatenazione si differenziano sulla base della dimensione dei tratti concatenati. I tre tipi principali sono: quella per campioni, quella per difoni e quella per applicazioni specifiche.
Nella sintesi per campioni, grandi database di voci registrate vengono frazionati in uno o più dei seguenti tratti: sillabe, morfemi, parole, frasi o interi periodi. Durante la composizione vengono poi utilizzati i tratti a disposizione con eventuali piccoli aggiustamenti per livellare le transizioni tra gli stessi. Questa tecnica produce ovviamente i risultati con la naturalezza maggiore a discapito delle dimensioni del database contenente i campioni; infatti, a maggiori prestazioni corrisponderanno database più grandi, anche di diversi gigabyte.
La seconda tecnica, la sintesi per difoni, è quella con i risultati peggiori e si distingue solo per le dimensioni minime del database. Questa tecnica sfrutta per l'appunto una combinazione di difoni per creare l'uscita. Un difono è una coppia di suoni fonetici adiacenti in una sequenza verbale, si può vedere come la transizione tra due suoni. Tenendo conto che il numero di difoni per un dato linguaggio è basso (qualche migliaio) le dimensioni del database in queste tecniche sono abbastanza ridotte.
La sintesi per applicazioni specifiche invece si basa sull'unione di parole o intere frasi preregistrate. È usata in contesti con un dominio di frasi limitato ed appare ovviamente molto naturale. Viene usata in diversi ambiti commerciali, in quanto estremamente semplice da implementare.

Sintesi per formanti

La sintesi per formanti non usa campioni di voce umana per sintetizzare l'uscita, ma lo fa basandosi su un modello acustico. In questa tecnica è un software che genera e modula le forme d'onda, basandosi sulle formanti della voce umana.
Per formante, nella scienza del linguaggio, si fa riferimento al massimo locale dello spettro formato da una risonanza acustica del tratto vocale umano.
Questi sistemi producono un'uscita spesso metallica e poco naturale, tuttavia sono caratterizzati da un elevata intelligibilità, anche ad elevate velocità di riproduzione e

dimensioni ridotte del database, rendendoli ottimi prodotti per sistemi con memoria e capacità limitate.

2.c. Natural Language Understanding (NLU)

L'ultimo componente fondamentale nell'architettura di un assistente digitale è il modulo di comprensione del linguaggio naturale, nonché anche il più complesso tra quelli visti.

La comprensione del linguaggio naturale infatti è la capacità di una macchina di comprendere un determinato testo, ed è considerato uno dei problemi IA-completo [19].

Con l'espressione IA-completo si fa riferimento a quella serie di problemi la cui risoluzione definitiva richiede la creazione di un'intelligenza artificiale forte o anche detta generale (AGI), ovvero un agente capace di comprendere e imparare ogni attività intellettuale al pari di un'intelligenza umana.

Un sistema di NLU generalmente ha bisogno di un database contenente il vocabolario del linguaggio in esame, una descrizione delle sue regole grammaticali e un sistema per analizzare la sintassi del testo, questo indipendentemente dall'approccio utilizzato.

Per analizzare la sintassi il sistema avrà bisogno di regole sintattiche da seguire, e tendenzialmente la capacità di comprensione di tutto il modello di NLU dipende dalla qualità di quest'ultime.

Lo scopo di questa analisi è di spezzare la frase e categorizzare i vari tratti secondo uno schema interno. All'interno di questo schema si possono identificare due macrocategorie: lo scopo e le entità.

In un sistema di NLU il riconoscimento dello scopo è l'azione primaria, e consiste nell'identificare il sentimento dell'utente e il suo obbiettivo. In parallelo avviene anche il riconoscimento delle entità, termine con cui ci si riferisce alle informazioni rilevanti presenti nel messaggio che vanno a specificare una direzione precisa nello scopo dell'utente.

Considerando la frase “Vorrei acquistare un biglietto per il treno per Milano di domani”, un esempio di scomposizione ed etichettatura che un modello di NLU potrebbe fare è:

- Acquistare un biglietto per il treno [obbiettivo]
- Milano [destinazione]
- Domani [data]

Una volta che i dati sono stati etichettati, al sistema non resta che comunicare l’azione e le relative informazioni aggiunte ai successivi moduli di competenza, a patto che ve ne sia almeno uno in grado di soddisfare tale richiesta.
La procedura di allenamento con la quale vengono definite le regole sintattiche crea un collegamento tra un’azione e delle frasi campione. Un sistema come quello dell’esempio precedente, che permette di acquistare un biglietto del treno, sarà stato allenato con frasi come: “Acquisto biglietto”, “Voglio comprare un biglietto”, “Devo prendere un treno”, “Vorrei un biglietto del treno”, ecc.
Grazie a questi campioni, il modello creerà un’associazione tra le frasi d’ingresso con una sintassi simile e l’azione specificata, ovvero avviare un’operazione di acquisto.
Attenzione però, in questo momento il suddetto modello sarebbe capace di comprendere solo l’obbiettivo primario, lo scopo per l’appunto. Non è infatti stato allenato per distinguere le diverse entità che possono presentarsi all’interno della frase, finirebbe per ignorarle.
Per far ciò servirebbe innanzitutto dichiarare le varie entità possibili e poi definire altre frasi campioni che insegnino al modello in quali contesti si possano presentare. Sempre in riferimento all’esempio, potremmo dire che le entità del modello, semplificando il problema, sono:

- stazione di partenza;
- stazione di destinazione;
- giorno.

Dovremo a questo punto ridefinire altri campioni del tipo: “Voglio acquistare un biglietto da [Stazione di partenza] a [Stazione di destinazione]”, “Biglietto treno per [Stazione di destinazione]”, “Acquisto biglietto da [Stazione di partenza] a [Stazione di destinazione] per [Giorno], “Biglietto per [Stazione di destinazione] per [Giorno]”, ecc.

Grazie a questo nuovo elenco di campioni il sistema sarà in grado di ricollegare le informazioni che si presenteranno in frasi sintatticamente simili alle entità dichiarate.

L’unione di tutti gli scopi definiti in un sistema, assieme alle varie entità e a tutte le frasi campione è detto modello linguistico.

La procedura di allenamento trattata in precedenza prende il nome di *fine tuning*. Quest’ultima prevede di modificare la struttura di un modello generico, ad esempio aggiungendo un nuovo strato alla fine di quest’ultimo, per ottenerne uno specifico per l’applicazione in questione. Sarà a questo punto sufficiente fornire i dati in ingresso necessari a stimare i pesi dello strato aggiunto per ottenere il nuovo modello.

Non si può pensare infatti di allenare ad ogni modifica un modello di NLP da zero, essendo la procedura estremamente dispendiosa sia in termini di tempo che computazionali. La vera operazione di allenamento, che prende il nome di *preallenamento,* viene effettuata una sola volta, e definisce il comportamento generale della rete.

Uno dei modelli generici più noti nel settore è BERT (*Bidirectional Encoder Representations from Transformers*). BERT è il modello proposto da Google nel 2019 [20], e che ancora ad oggi ricopre un ruolo centrale in tutti i prodotti Google che lavorano in un contesto di comprensione del linguaggio naturale, a partire dal motore di ricerca proprietario della compagnia stessa.

Usando BERT come base è possibile creare, tramite fine tuning, diversi modelli specifici, come ad esempio modelli per il riconoscimento delle entità (*Named Entity Recognition*), come nell’esempio precedente, oppure modelli in grado di comprendere le relazioni tra frasi e ad esempio rispondere a delle domande poste dall’utente (*Question Answering Task*).

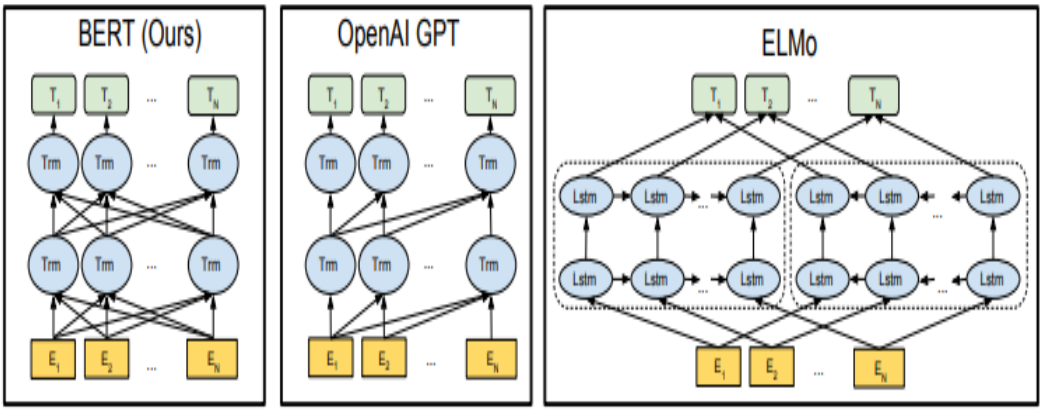

Figura 13: Tre modelli architetturali a confronto. BERT usa dei Transformer bidirezionali. OpenAI GPT usa Transformer che sfruttano solo il contesto passato. ELMo concatena due blocchi separati di LSTMs uno che sfrutta solo il contesto passato della sequenza di ingresso e uno solo quello futuro. Fonte [20]

L'architettura di BERT è estremamente simile a quella di un Transformer standard. La differenza principale con le altre versioni di queste reti, tuttavia, risiede nella procedura di allenamento di quest'ultima.

Al fine di creare un modello così flessibile, nella fase di preallenamento di BERT sono state usate due strategie fondamentali: MLM (*Masked Language Model*) e NSP (*Next Sentence Prediction*).

Al contrario di altri modelli noti, come *ELMo* [21] o *OpenAI GPT*, che limitano gli strati di self-attention del modello ad elaborare solo il contesto passato, BERT riesce a sfruttarne le piene potenzialità proprio grazie al Masked Language Model.

Normalmente modelli di questo tipo vengono allenati usando solo il contesto sinistro o destro dato un certo elemento della sequenza, in quanto, senza MLM, usare il contesto completo permetterebbe di sfruttare informazioni legate ad altri elementi della sequenza, che a loro volta però vedono la parola che stiamo cercando di classificare, rendendo il problema mal formattato.

Per ovviare a ciò, nel MLM, si sceglie una certa percentuale di elementi dell'ingresso da mascherare. Il modello cercherà di classificare solo quest'ultimi, e non ogni elemento della sequenza d'ingresso.

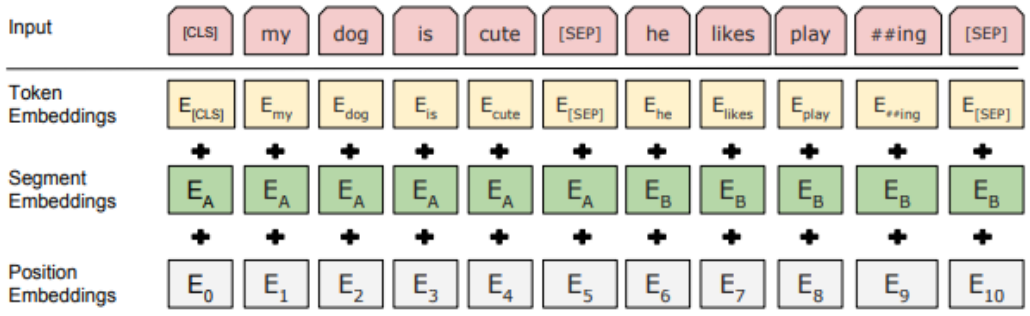

Figura 14: Rappresentazione grafica degli ingressi in una rete BERT. Gli ingressi del modello sono generati sommando la versione codificata del vettore di ingresso alle informazioni extra di: frase di appartenenza e posizione nella stessa. Fonte [20]

Il rovescio della medaglia di questa tecnica sta nel fatto che stiamo creando un disallineamento tra i dati ricevuti durante il preallenamento e quelli durante la fase di fine tuning, in quanto in quest'ultima non esisterà l'elemento mascherato.

Per ridurre questa discrepanza, durante il preallenamento non tutti gli elementi scelti come mascherati lo saranno veramente. Quest'ultimi hanno l'80% di probabilità di essere sostituiti con l'etichetta [MASK], e quindi mascherati, il 10% di essere sostituiti con un elemento casuale e un altro 10% di rimanere invariati.

Definito l'insieme di elementi mascherati, che, come detto, sono gli unici da classificare, avremo risolto il problema sollevato in partenza che ci impediva di utilizzare le informazioni derivanti dall'intera sequenza nella classificazione di un particolare elemento.

A questo punto, una volta incontrato un elemento mascherato, la rete userà il vettore d'uscita dell'ultimo codificatore per classificare quest'ultimo.

La seconda tecnica usata nel preallenamento di BERT si rivela particolarmente utile in applicazioni di riconoscimento della lingua (*Natural Language Inference*) e QA (*Question Answering*).

Lo scopo dietro l'utilizzo del NSP è rendere il modello consapevole delle relazioni tra frasi successive. L'operazione effettuata è molto semplice. Durante il preallenamento il modello riceve in ingresso una coppia di frasi, in cui solo la metà delle volte la

seconda è la vera succeditrice della prima. Viene chiesto al modello di identificare quando questo legame sia vero o meno.

Nel preallenamento del modello BERT, queste due tecniche, MLM e NSP, vengono allenate insieme. L'obbiettivo del modello sarà quindi minimizzare la perdita combinata di entrambe.

Per far sì che il modello possa essere allenato contemporaneamente con frasi singole (MLM) e coppie di frasi (NSP), i dati in ingresso devono essere opportunamente formattati. Un elemento speciale chiamato [CLS], viene inserito all'inizio di ogni frase, nel caso di due frasi nei dati d'ingresso viene anche aggiunto l'elemento [SEP] come separatore tra le due. Inoltre, ad ogni elemento della sequenza vengono aggiunte due informazioni: la frase di appartenenza e la posizione nella stessa.

III. L'INTERAZIONE BANCA - CLIENTE

1. L'interazione attuale e i benefici degli assistenti digitali

Le banche rappresentano da secoli i pilastri del nostro sistema finanziario mettendo a disposizione dei propri clienti tutta una serie di strumenti atti alla propria gestione patrimoniale.

Proprio il rapporto con quest'ultimi è evoluto in modo significativo negli anni, adattandosi con il tempo alle nuove frontiere tecnologiche e alle azioni regolamentari arrivate con esse.

Tra le motivazioni che hanno spinto le banche ad innovare e migliorare l'interazione con i clienti, oltre alla mera soddisfazione degli stessi, che ovviamente si ripercuote positivamente sulla reputazione e la fidelizzazione nei confronti dell'istituto, si trovano anche quelle economiche. È interesse di ogni banca migliorare e automatizzare le interazioni con la massa. Se la gestione delle stesse interazioni richiede meno risorse, si sta abbassando quello che è il costo medio di ognuna di esse, ed è esattamente quello che ogni istituto bancario ha sempre fatto finora.

Inizialmente l'esperienza di interazione tra banca e cliente era totalmente fisica. Ogni utente necessitava di recarsi in una sede della stessa per ogni tipo di operazione, semplice o complessa.

I primi cambi a tale struttura ci furono alla fine degli anni '60 con l'introduzione degli ATM (Automated Teller Machines), dispositivi ancora ad oggi molto utilizzati, anche se la tendenza negli anni del loro utilizzo ne indicherebbe un lento abbandono, probabilmente a causa dei sempre più limitati usi del contante. Grazie ad essi, operazioni più semplici, come il prelievo o il deposito di contanti, sono state automatizzate, permettendo ai clienti di servirsene senza la necessità di visitare la filiale e attendere in fila di essere serviti.

Con l'avvento dell'era del digitale, gli istituti bancari hanno affrontato quella che senza dubbio è stata la rivoluzione più grande su questo tema.

Iniziano a nascere una vasta serie di canali di interazione con l'utente al passo con la tecnologia. Prendono vita quelli che oggi sono noti come "Online Banking" e "Mobile Banking". Ogni tipo di operazione, dall'apertura del conto corrente all'esecuzione di un bonifico è stata integrata in questi canali.

Sono proprio questi i prodotti con cui la maggior parte degli utenti interagisce e sono considerati l'attuale standard nell'accesso ai servizi finanziari.

Anche altri canali come e-mail, chatbot e social media sono molto utilizzati nell'interazione moderna, quest'ultimi tuttavia, ricoprono prevalentemente ruoli di supporto. È abbastanza usuale, infatti, che un utente utilizzi uno di essi per ricevere assistenza immediata per la risoluzione di un problema, e rappresentano per gli istituti bancari un valido supporto nella gestione delle richieste più frequenti con tempi brevi e bassi costi.

Avere una gamma di prodotti e canali di comunicazione così vasta non è solo un vantaggio per istituzioni di questo tipo.

A causa del continuo rilascio di nuovi servizi sempre più accessibili, e soprattutto raggiungibili tramite reti pubbliche, si vanno ad incrementare notevolmente tutti i rischi legati alla sicurezza e alla privacy dei clienti.

Questo è sicuramente uno tra i temi più rilevanti in un contesto come quello bancario. Una vulnerabilità che porti ad una fuoriuscita di dati sensibili (Data Breach), o peggio alla perdita dei fondi degli utenti sarebbe disastrosa per l'istituto in questione. Anche supponendo un danno relativamente piccolo e riparabile dal punto di vista economico, l'effetto che avrebbe sull'immagine e la reputazione della banca stessa sarebbe comunque ingente e probabilmente difficile da riparare.

Come risultato di ciò, le varie istituzioni bancarie investono pesantemente in misure di sicurezza per farvi fronte, e devono essere coscienziose di fronte all'implementazione di nuove tecnologie, ed evitare che la voglia di essere tra i pionieri di una certa innovazione si trasformi in un disastro.

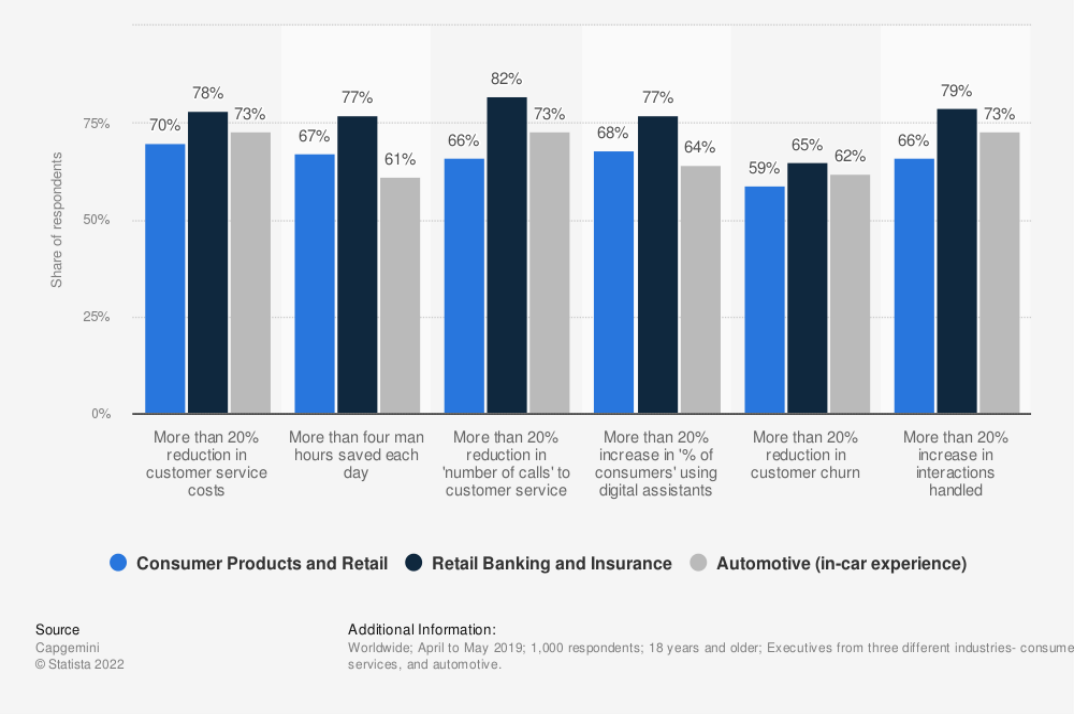

Figura 15: Percentuale di compagnie che riscontrano i benefici menzionati in figura dall'adozione degli assistenti digitali raggruppate per settore. 1. Riduzione di oltre il 20% del costo del servizio clienti, 2. Oltre quattro ore persona risparmiate ogni giorno, 3. Riduzione di oltre il 20% delle chiamate all'assistenza clienti, 4. Incremento di oltre il 20% nella percentuale di utenti che utilizzano gli assistenti digitali, 5. Riduzione di oltre il 20% del tasso di abbandono dei clienti, 6. Incremento di oltre il 20% delle interazioni gestite. Fonte [43]

Proprio in questo contesto si vanno ad inserire gli assistenti digitali. Questa tecnologia desta sempre più interesse tra i vari istituti bancari e sposa perfettamente le argomentazioni precedentemente citate.

Gli assistenti digitali rappresentano, nel settore bancario, il futuro dell'interazione con il cliente sotto diversi aspetti. Grazie a quest'ultimi gli utenti gioverebbero di un'esperienza sempre più personalizzata, veloce e soprattutto comoda, in quanto l'unico mezzo di interazione sarebbe la voce.

Dal lavoro in [22] si può notare che le motivazioni che spingono gli utenti ad usare attualmente l'Online Banking ruotano prevalentemente intorno all'accessibilità e all'efficienza dell'interazione, permettendo di risparmiare tempo rispetto ai metodi convenzionali. Attraverso l'Online Banking è facile usare in autonomia la maggior parte delle operazioni offerte da un'istituzione bancaria. Unendo questa peculiarità alla

velocità con cui si eseguono le operazioni e all'immediato ritorno visivo dell'esito nei confronti degli utenti, è facile comprendere come mai rappresenti ad oggi il prodotto preferito, per interagire con questo settore.

Nonostante sia il preferito per la maggior parte degli utenti, non lo è per tutti. Si può notare una certa correlazione tra gli utilizzatori dei servizi di Online Banking e utenti con giornate molto impegnate e che soprattutto si sentono a loro agio con prodotti tecnologici. Questa categoria di utilizzatori, infatti, nonostante abbia preoccupazioni in merito ad alcuni aspetti legati a questi servizi, utilizza costantemente quest'ultimi per ogni operazione considerata quotidiana o comunque poco importante e che implica un piccolo rischio. Si continua a preferire l'interazione fisica in una sede della banca per operazioni considerate più complesse, come ad esempio la richiesta di un mutuo, anche da questa categoria di utilizzatori.

Entrando più nello specifico di quelle che sono le preoccupazioni e i problemi riscontranti dagli utenti in questi servizi, si notato il tema della sicurezza e quello della poca personalizzazione.

Tra quest'ultimi due la sicurezza di tutto il sistema è ciò che preoccupa principalmente gli utilizzatori e tiene lontano chi, spesso a causa della poca affinità con il mondo tecnologico, non si è ancora avvicinato a questa tipologia di interazione.

Sempre in [22] infatti, si può notare come coloro che non utilizzano questo tipo di servizi e prodotti, giustificano il loro comportamento sulla base del fatto che, oltre alle sopra citate preoccupazioni per la sicurezza, non si sentono a loro agio con il mondo tecnologico o peccano di una conoscenza dello stesso. Per quest'ultimi l'interazione preferita è quella di recarsi fisicamente in una sede dell'istituto.

Sia dagli utilizzatori dell'Online Banking che dai non, la possibilità di creare un legame di conoscenza e fiducia con il personale della banca è il fattore decisivo nella scelta di utilizzo di questa modalità.

Gli utilizzatori abituali dei servizi di Online Banking, solo nel momento in cui il vantaggio di avere una persona fidata come consigliera supera gli svantaggi di poca convenienza e tempo perso, dati dall'interazione fisica, optano per quest'ultima.

Tabella 1: Caratteristiche personali imputate dell'inutilizzo di sevizi di Online Banking. Fonte [22]

	Utilizzatori regolari di servizi bancari online	Non utilizzatori di servizi bancari online
Evitare le nuove tecnologie (%)	11	60
Insicurezza e mancanza di privacy (%)	56	50
Mancanza di conoscenza e poca comodità (%)	22	60
Depersonalizzazione / Problemi sociali (%)	11	50
Poco coinvolgimento finanziario (%)	22	20
Età media tra gli intervistati	42	56
Percentuale di diplomati	77	40

Tabella 2: Valutazione dei servizi di Online Banking da parte dei gruppi di utilizzatori e non. Fonte [22]

	Percentuale di utilizzatori regolari di servizi bancari online	Percentuale di non utilizzatori di servizi bancari online
Vantaggi servizi bancari online		
Accessibilità	100	20
Risparmio di tempo	100	20
Semplicità di utilizzo	100	38
Gestione dei feed-back	78	25
Funzionalità adeguate	67	20
Qualità delle informazioni	56	30
Convenienza	44	20
Autonomia	56	30
Facilità di personalizzazione	0	0
Svantaggi servizi bancari online		
Preoccupazioni sulla sicurezza	78	50
Mancanza di funzionalità	67	20
Mancanza di qualità di informazione	78	0
Problemi tecnici	22	0
Mancanza di personalizzazione	33	40
Processi complessi	22	10
Mancanza di interazione col personale d'ufficio	11	0
Nessun valore aggiunto ai canali già usati	0	30
Mancanza di conoscenza	44	70

Tenendo in considerazione quanto detto finora, è facile comprendere come gli assistenti digitali possano rappresentare uno strumento cruciale in quella che sarà l'interazione con i servizi bancari negli anni a venire.

Questi strumenti potrebbero portare l'interazione di chi già usa servizi bancari online ad un nuovo livello e allo stesso tempo essere un ponte per le generazioni più adulte, che, come visto in precedenza, a causa della loro poca familiarità con i prodotti tecnologici, evitano questi servizi.

Per la prima categoria di utenti che è stata definita, quelli che hanno familiarità con questi servizi, gli assistenti digitali rappresenterebbero un ulteriore miglioramento dei vantaggi che quest'ultimi hanno identificato. Prendendo ad esempio i tre vantaggi principali dell'online banking, ovvero accessibilità, risparmio di tempo e semplicità d'uso, si può notare come, con l'utilizzo di questi dispositivi, non si fa altro che incrementarne l'efficacia. La possibilità di usare gli assistenti ovunque, anche mentre si è impegnati, con il solo ausilio della nostra stessa voce, è il fattore più incisivo in questa tecnologia.

Tutte le funzionalità ad oggi disponibili nei comuni prodotti di Online Banking possono essere implementate in questi dispositivi, da semplici operazioni informative date dalla consultazione del conto all'esecuzione di vere e proprie transazioni.

Inoltre, questi dispositivi ad oggi hanno spesso nomi propri e voci molto simili a quella umana. Queste caratteristiche, unite alla capacità di essere disponibili 24/7 a supporto del cliente, possono anche risolvere il difetto di depersonalizzazione riscontrato da molti utenti nei riguardi dell'Online Banking. È abbastanza comune, infatti, che gli istituti che per primi si stanno approcciando a queste tecnologie, mettano a disposizione dei propri clienti servizi di assistenza o addirittura consigli di spesa che il dispositivo propone sulla base delle abitudini e lo storico di transazioni, dando l'impressione di avere un consigliere personale che conosce la propria situazione finanziaria, esattamente come accade nell'interazione fisica.

Analizzando la seconda categoria di clienti, la scarsa familiarità con la tecnologia e la depersonalizzazione dell'esperienza bancaria rappresentano sicuramente gli scogli principali.

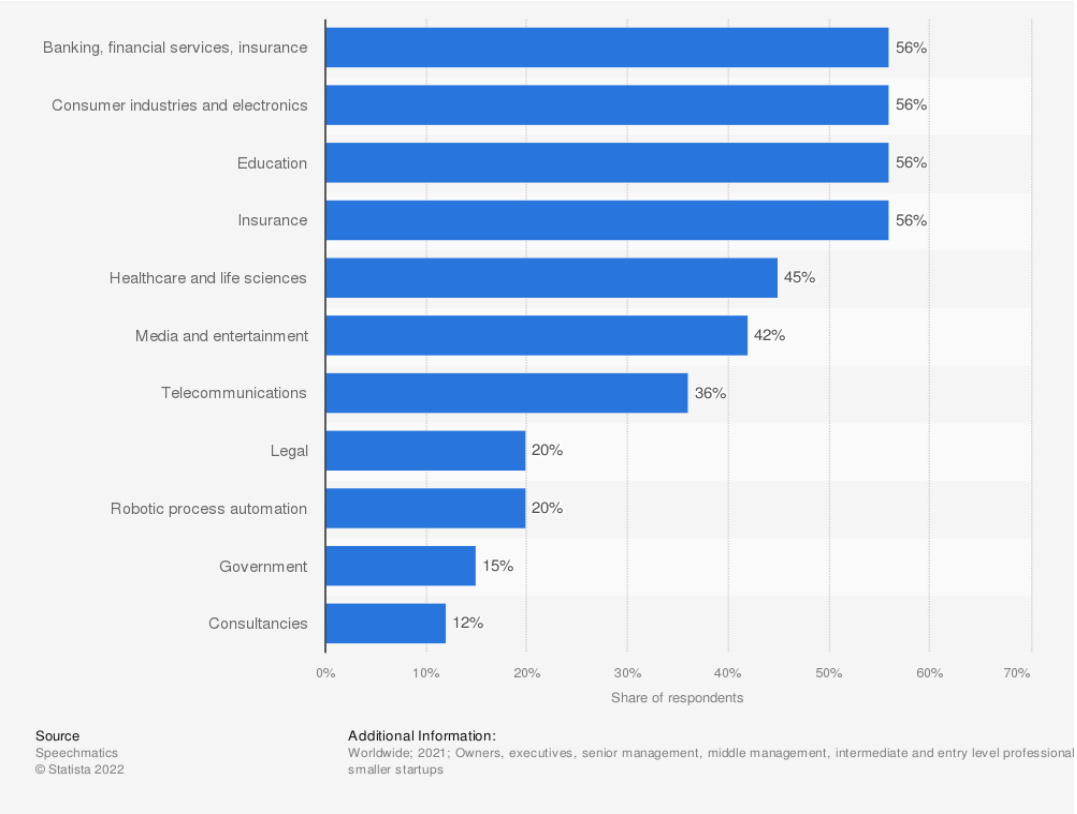

Figura 16: Industrie che incrementeranno gli usi e le applicazioni delle tecnologie vocali nel giro di 3-5 anni (dal 2021). Fonte [44]

Soprattutto per i più anziani tra quest'ultimi, gli assistenti digitali rappresenterebbero un punto d'ingresso per i servizi bancari online andando a sopperire proprio alle mancanze precedenti.

Dal lavoro di Margot Hanley e Shiri Azenkot [23], si nota come l'utilizzo di questi dispositivi sia in rapido aumento tra gli adulti con più 65 anni.

Gli assistenti digitali sono pensati per essere socievoli, con comportamenti ed espressioni molto simili a quelli umani, aggiungendo inoltre che l'interazione con quest'ultimi avviene tramite voce, è facile comprendere il perché.

In questi prodotti non si trovano le consuete barriere che contraddistinguono gli altri dispositivi tecnologici, permettendo anche a chi non ha familiarità con questo mondo di interagirci.

2. Analisi delle soluzioni dei principali istituti bancari

Molti istituti bancari, viste le premesse, hanno già iniziato a sperimentare e commercializzare delle soluzioni per sfruttare le potenzialità di questi dispositivi.
Si citeranno e analizzeranno di seguito alcune tra le soluzioni commercializzate da banche internazionali.
Purtroppo, come si vedrà meglio nel quarto capitolo, nonostante le brillanti premesse, le attuali tecnologie sfruttabili non sono sufficienti alla completa realizzazione delle idee finora citate dietro questi dispositivi. Sarà necessario optare per prodotti con una struttura simile a quelli che verranno analizzati ora, che sono soggetti ad un compromesso per quello che riguarda la sicurezza e l'esperienza utente.
È possibile dividere in due gruppi gli istituti bancari che hanno rilasciato un prodotto di questa tipologia. Un primo gruppo che ha provato ad integrare le funzionalità finanziarie desiderate in assistenti digitali già commercializzati (Amazon Alexa, Google Assistant, …), andando a creare delle applicazioni per quest'ultimi. Questa soluzione delega la maggior parte dell'elaborazione vocale al fornitore del prodotto (tutte le fasi di elaborazione viste nel capitolo precedente: riconoscimento vocale, comprensione della richiesta e sintesi vocale), permettendo agli sviluppatori di concentrarsi solo sulla parte relativa alla gestione delle funzionalità che si desiderano implementare. Il rovescio della medaglia di questa soluzione, come intuibile, è che non si ha il pieno controllo della catena di elaborazione del dato, e come tale, in alcune circostanze, si è vincolati a cercare un compromesso tra l'idea di realizzazione iniziale e quella effettivamente realizzabile (si analizzerà nel dettaglio questa problematica nel prossimo capitolo, principalmente i compromessi legati alla sicurezza). Un secondo gruppo di istituti, invece, ha optato per la creazione di un proprio assistente digitale, distribuito come applicazione sui principali mercati per smartphones (Google Play e AppStore). Anche in questo caso, tuttavia, non mancano i compromessi. Con una soluzione di questo tipo si risolve il problema della libertà di gestione dei dati, ma si perde il vantaggio di accessibilità che l'utilizzo di dispositivi noti e già commercializzati rappresentava. Come già detto, infatti, questi assistenti personalizzati

nascono come applicazioni per smartphones, e ciò limita molto le possibilità di utilizzo degli stessi se paragonati ad esempio ad un assistente come Alexa disponibile anche su dispositivi indossabili, auto, smart speaker, ecc.

La prima compagnia bancaria che ha creato un prodotto per l'accesso ai servizi finanziari come applicazione per gli assistenti digitali già commercializzati è Capital One.

Capital One è una delle più grandi banche ad operare negli Stati Uniti, e nel 2017 ha rilasciato una sua applicazione (anche note come *skill*) per integrare l'assistente di Amazon, Alexa, con i propri servizi [24].

Tramite questa applicazione è possibile, con il solo uso della voce, interagire con il proprio conto e gestire le proprie carte.

Sono state abilitate la maggior parte delle operazioni che un utente effettua con il proprio conto, come ad esempio:

- Controllare il bilancio del proprio conto e delle proprie carte;
- Chiedere informazioni sulle ultime spese;
- Iniziare un pagamento tramite carta di credito/debito;
- Autorizzare il pagamento delle rate di un prestito in sospeso;
- Chiedere informazioni sui prestiti in corso per il proprio conto.

Come detto in precedenza, tuttavia, abilitare un'applicazione di questo tipo, seppur apparentemente molto utile nella vita quotidiana, espone il correntista ad un grosso rischio. Nell'applicazione distribuita da Capital One non vi è alcuna misura forte di sicurezza (questa non è una mancanza di quest'ultima, si analizzerà nel prossimo capitolo come non sia possibile averla a causa di vincoli architetturali).

Come si può leggere chiaramente nel contratto di termini e condizioni distribuito dalla compagnia [24], l'applicazione non usa la voce dell'utente per verificare che lo stesso abbia il permesso di accedere al conto e ai vari servizi. Ogni comunicazione dalla skill a Capital One, viene trattata da quest'ultima come autorizzata dall'utente. Sarà quindi responsabilità dell'utente gestire e controllare tutte le comunicazioni in uscita dall'applicazione.

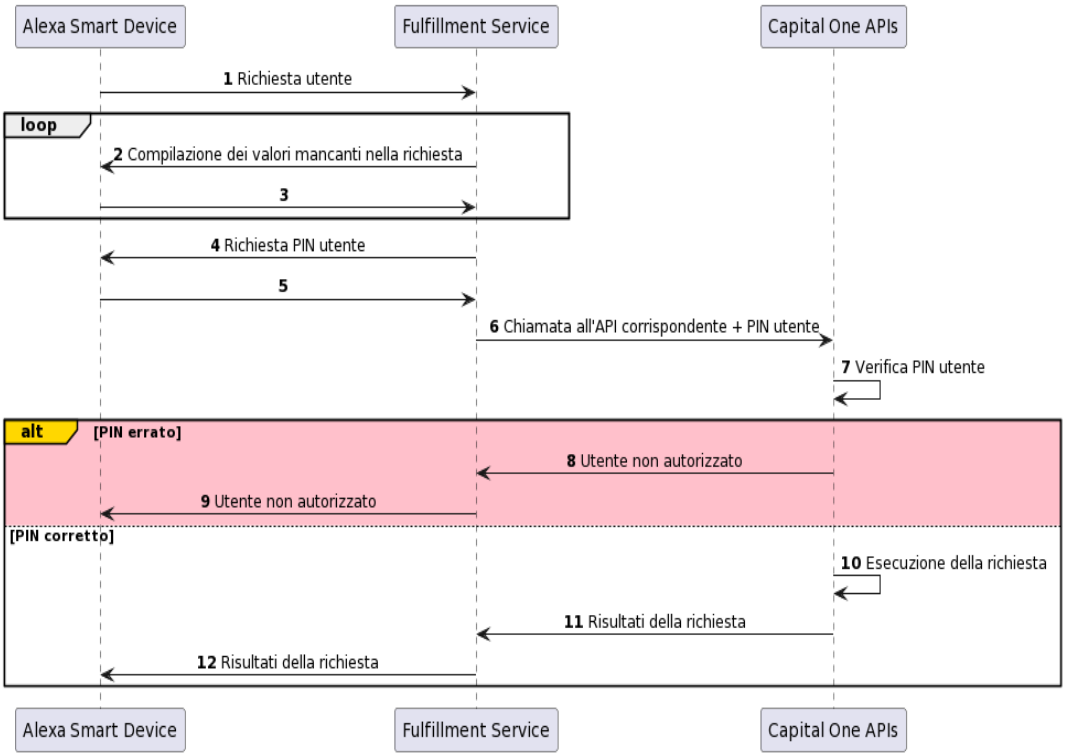

Figura 17: Rappresentazione grafica semplificata dell'interazione con l'applicazione per Alexa distribuita da Capital One. La rappresentazione è stata semplificata escludendo gli elementi relativi alla connessione tra il dispositivo Amazon e l'account utente Capital One, l'interazione completa di una soluzione di questo tipo sarà disponibile nel quarto capitolo.

Chiunque ascolti l'utente interagirci o abbia accesso alla skill, potrà ottenere informazioni relative al conto e le carte ad essa collegati, e avrà la stessa capacità di eseguire operazioni al pari dell'effettivo proprietario.

L'unico accorgimento di sicurezza utilizzato da Capital One è la presenza di un PIN numerico richiesto nel momento di interazione con l'applicazione. È abbastanza chiaro che sia una misura poco efficace sotto diversi punti di vista. Il PIN può essere ascoltato da qualcuno durante un'interazione legittima e poi replicato. Se poi si considera che lo storico delle conversazioni con una skill viene salvato in chiaro anche sull'account Amazon del proprietario e che inoltre transita in chiaro sui vari server durante l'elaborazione, si arriva direttamente alla conclusione che un sistema di sicurezza di questo tipo è come se non ci fosse.

Il secondo tipo di soluzioni attualmente distribuite prevede la creazione di un proprio assistente, proprio per arginare le limitazioni di sicurezza che la soluzione vista in precedenza comporta.

Di questo secondo gruppo, il prodotto di maggior successo è *Erica* [25], l'assistente sviluppato da Bank of America, la seconda banca più importante degli Stati Uniti che fornisce circa il 10% di tutti i depositi bancari americani.

Questo prodotto, a differenza di quello analizzato in precedenza, può essere visto più come una scorciatoia per interagire con la propria applicazione bancaria, piuttosto che un vero e proprio assistente.

Per interagire con Erica, è necessario avere l'applicazione per dispositivi mobili di Bank of America, autenticarsi nella stessa e aprire l'assistente. È abbastanza chiaro che in questo schema di interazione la sicurezza è implicita, in quanto per interagire con l'assistente devo innanzitutto autenticarmi nell'applicazione, scaricando l'onere di mantenere un alto livello di sicurezza alla comune verifica effettuata da quest'ultima.

Il problema legato a queste soluzioni è che si sono andati a perdere molti dei vantaggi legati all'usabilità e accessibilità.

L'usabilità è stata compromessa, rispetto all'idea originale, in quanto è stato reintrodotto l'utilizzo dello smartphone. Non è più un prodotto con cui interagire utilizzando solo la voce. Inoltre, optando per lo sviluppo di un assistente personalizzato, come già detto, è stata esclusa una vasta gamma di dispositivi, che ovviamente riduce l'accesso all'assistente. Vi è poi un fattore legato alla accessibilità linguistica. Sviluppare da zero un assistente impone anche la necessità di gestire singolarmente i vari linguaggi supportati. Se si prende Erica, ad esempio, supporta solo un'interazione in lingua inglese, al contrario dell'assistente di Capital One che, basandosi su Alexa, può facilmente gestire tutti i linguaggi con cui la stessa è compatibile.

È facile comprendere attraverso questi due prodotti, come sia necessario un forte compromesso tra usabilità e sicurezza. Nonostante ciò, entrambi hanno riscosso e continuano a riscuotere un buon successo, e lo si nota dall'utilizzo crescente che i rispettivi correntisti fanno degli stessi.

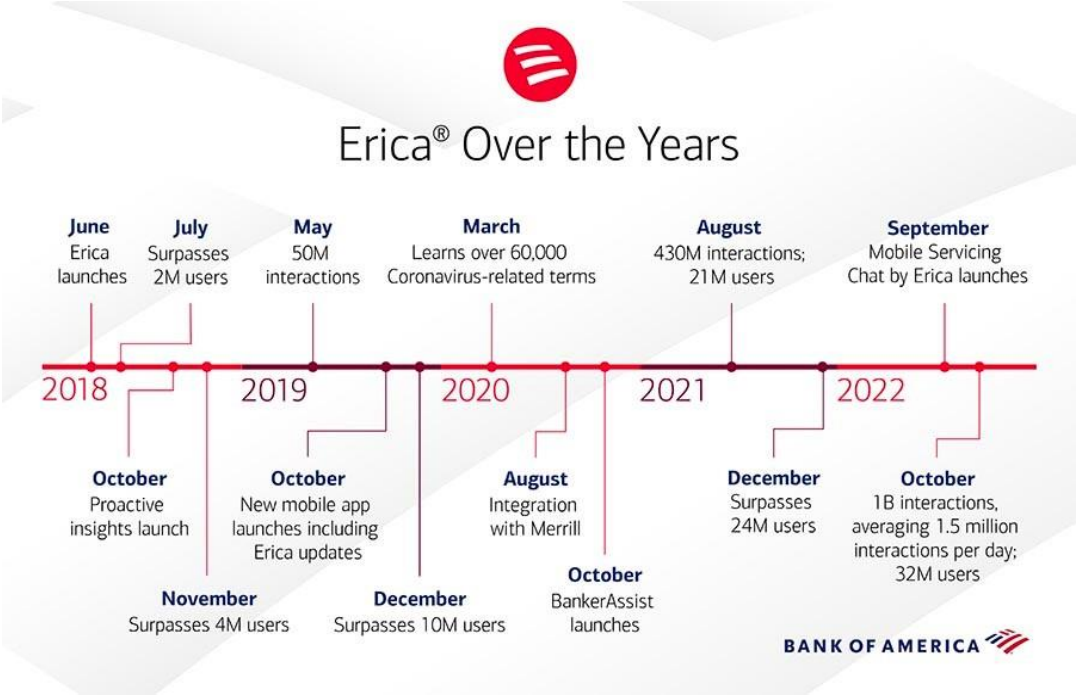

Figura 18: Principali traguardi nel numero di interazioni di Erica, l'assistente di Bank of America dal lancio alla fine del 2022. Fonte [26]

Si può notare in un articolo pubblicato da Bank of America alla fine del 2022 [26], di come il suo assistente, Erica, contasse dall'uscita circa un miliardo di interazioni, con una media attuale di circa 1.5 milioni di interazioni al giorno.

IV. ANALISI E RISULTATI DEL PROGETTO

1. Panoramica

Questo lavoro di tesi è stato proposto dalla società Iriscube Reply e svolto in collaborazione con essa.

L'obbiettivo del lavoro di tesi è stato quello di analizzare le opportunità date dalle principali piattaforme cloud (Google e Amazon) per integrare i servizi offerti da un istituto bancario con gli assistenti digitali conversazionali (Google Home, Amazon Echo), per permettere di eseguire semplici operazioni informative e dispositive.

La tesi si è composta di una parte di ricerca e analisi delle principali soluzioni disponibili sul mercato e dei relativi aspetti architetturali e tecnologici.

Sono state selezionate le tecnologie più adatte a quello che era il tema e si sono evidenziate quelle che erano le criticità all'interno delle stesse.

In una seconda parte si è poi proceduto alla realizzazione di un prototipo dimostrativo delle potenzialità di cui questi dispositivi sono capaci nell'attuale condizione tecnologica.

Nei capitoli precedenti è stata già fornita un'anticipazione dei risultati dell'analisi fatta, evidenziando i limiti a cui questi dispositivi devono sottostare. A causa di quest'ultimi il prototipo proposto non rispecchia quelle che erano le idee iniziali di una soluzione totalmente basata su un'interazione vocale con l'utente.

Per far fronte al problema evidenziato sono state proposte alcune possibili varianti di quest'idea con un grado di sicurezza accettabile.

E' stato infine proposto e analizzato, solo dal punto di vista teorico, l'uso di tecnologie basate su biometriche vocali come possibile soluzione. Quest'ultima soluzione è stata solo teorizzata e non implementata a causa di un limite che verrà specificato nel paragrafo dedicato. Per riassumere il problema, questa soluzione deve essere

implementata dal distributore del dispositivo, non è nelle capacità dello sviluppatore integrare questo strato di sicurezza nella catena di elaborazione.

Come supporto su cui sviluppare il prototipo sono stati presi in considerazione i prodotti dei principali attori di mercato.

Dopo aver analizzato l'offerta di ognuno di essi, è stato scelto l'assistente di proprietà di Amazon, Alexa.

A seguito dell'analisi, la scelta di quest'ultimo è stata abbastanza forzata, ma non ha influito particolarmente sulla realizzazione del prototipo e sulle considerazioni che in seguito sono nate.

La struttura architetturale e tecnologica degli assistenti offerti dai vari produttori è estremamente simile, non inficiando appunto sull'esito del lavoro. L'unica differenza degna di nota riguarda l'aspetto di sviluppo su quest'ultimi. Alcuni assistenti, infatti, prevedono che lo sviluppo di un'applicazione per quest'ultimi avvenga tramite dei software offerti dal produttore stesso, in altri casi lo sviluppatore si occupa di programmare il comportamento della skill tramite codice.

Considerato che non vi sono particolari ragioni tecnologiche per selezionare un assistente in particolare, la scelta è avvenuta considerando la familiarità con gli strumenti necessari allo sviluppo e riducendo l'insieme di prodotti disponibili a quelli più utilizzati dagli utenti.

Considerando che Google e Amazon, con i relativi assistenti, si dividono la maggior parte delle quote di mercato [n], sono stati selezionati come i due migliori candidati.

Di quest'ultimi si è poi scoperto che Google, a partite da giugno 2023, rimuoverà dai propri assistenti la possibilità di avere conversazioni personalizzate con l'utente, limitando gli assistenti della compagnia ad operazioni quali controllare dispositivi compatibili in casa e azioni di base degli stessi. Si può leggere di più riguardo questa decisione in questo documento che Google stessa chiama "*Tramonto delle azioni conversazionali*" [27].

Le motivazioni che hanno spinto Google ad abbandonare queste funzionalità sembrerebbero legate alla possibilità che la compagnia opti ad una reinterpretazione

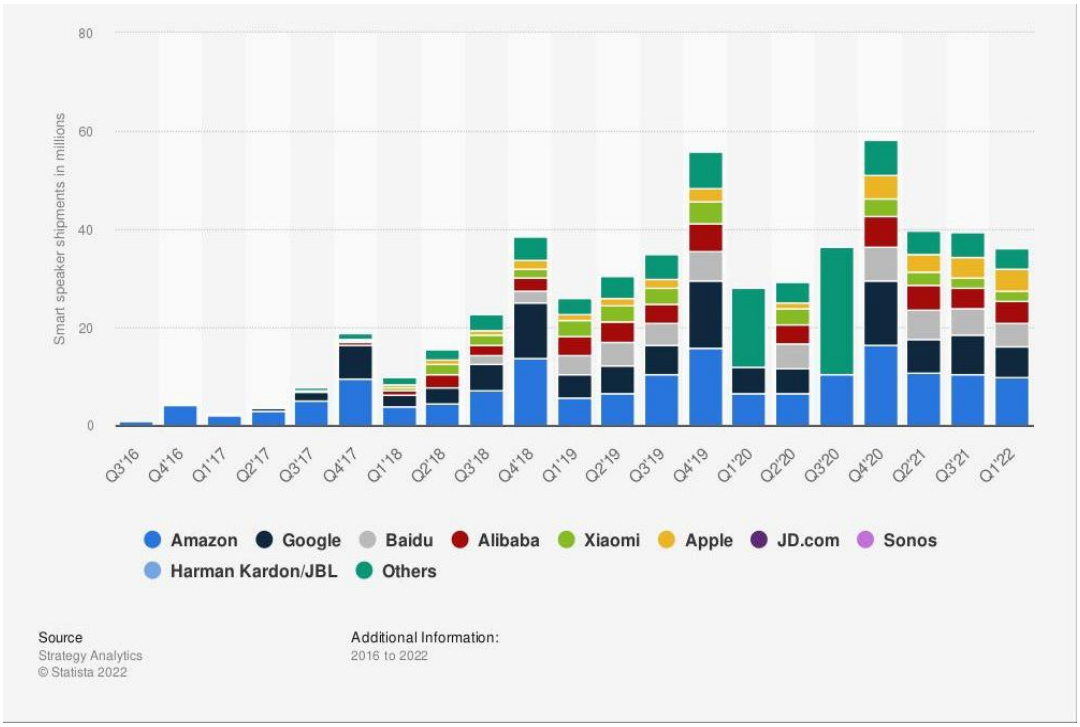

Figura 19: Quantità di smart speaker spediti nel mondo, raggruppati per venditore, dal 2016 al 2022. Fonte [45]

del ruolo degli assistenti che attualmente distribuisce, orientandoli ad una visione Android-centrica.

A causa di questa limitazione siamo stati costretti a scartare quest'ultimo come candidato, optando di conseguenza per la soluzione offerta da Amazon.

Nei paragrafi successivi ogni riferimento a tecnologie ed architetture è in relazione all'assistente Amazon Alexa. Come detto in precedenza, la scelta di un altro assistente come supporto per lo sviluppo del prototipo avrebbe portato alle stesse considerazioni finali legate alla sicurezza.

2. Architettura e logica di funzionamento

Quando si sviluppa un'applicazione (o skill) sfruttando l'assistente Alexa, ci si deve focalizzare unicamente nella gestione delle richieste da e per quest'ultima.
La gestione della registrazione audio contenente la richiesta dell'utente viene delegata al distributore del dispositivo (Amazon in questo caso).
Una volta che il dispositivo in questione (telefono, smart speaker, …) ha registrato la richiesta, quest'ultima verrà inviata a server appositi per la sua elaborazione. L'elaborazione in questione si compone dei passaggi analizzati nel dettaglio nei capitoli precedenti. La traccia viene trascritta tramite algoritmi di riconoscimento vocale e successivamente elaborata tramite algoritmi di NLU per riconoscere l'intento dell'utente ed eventuali informazioni aggiuntive nella richiesta.
La rete neurale che si occupa di quest'ultimo passaggio, viene rifinita (*fine tuning*) con informazioni inserite dallo sviluppatore. Come spiegato nel paragrafo relativo alle tecniche di NLU, lo sviluppatore inserirà frasi di esempio che si aspetta l'utente possa usare durante l'interazione con la sua skill. Ogni insieme di frasi, viene associata ad un componente logico chiamato *intento (Intent)*, rappresentante per l'appunto una particolare intenzione dell'utente.
Ad ogni intento vengono poi associati, se presenti, dei valori che ci si aspetta di dover ricevere per completare correttamente l'azione.
Se supponiamo di ricreare il lancio di una moneta con una skill, una delle frasi d'esempio sarà "Voglio giocare a testa o croce". Il valore aggiuntivo che l'intento si aspetta di ricevere, invece, potrebbe essere l'esito della scommessa e quindi "testa" o "croce". In un esempio come il precedente l'utente potrà attivare l'intento in questione con una delle frasi semplici, senza la presenza del valore aggiuntivo, facendo sì che sia l'applicazione ad accorgersi della mancanza e chiedergli di colmarla, oppure con una frase più complessa, che indichi oltre alla richiesta anche l'esito, dicendo ad esempio "Lancia una moneta, secondo me è testa" (supponendo che la skill sia stata allenata per gestirla).

La rete neurale e le vari componenti relative alla comprensione delle richieste utente sono trasparenti allo sviluppatore. Quest'ultimo, dopo aver creato la relazione logica tra le varie frasi e le azioni corrispondenti, comunica con l'utente tramite un'interazione ad alto livello.

Dal punto di vista dello sviluppatore il dispositivo con cui l'utente sta interagendo invia alla sua skill una serie di richieste in formato JSON e si aspetta di ricevere risposte adeguate nel medesimo formato.

Ogni traccia audio registrata dal dispositivo, dopo essere stata elaborata dall'algoritmo di NLU, viene associata ad uno degli intenti definiti dall'utente nella sua skill, o ad alcuni intenti di sistema (ad esempio le richieste di chiusura, di aiuto, ecc.), indicando se presenti anche i valori aggiuntivi riconosciuti.

Quest'informazione viene poi codificata in un oggetto JSON ed inviata al servizio che si occuperà di gestirla, anche detto servizio di f*ulfillment*. Quest'ultimo è l'unico componente che interagisce con il dispositivo utente. Può essere considerato la frontiera di tutto il blocco di servizi di backend.

Ad alto livello, l'interazione consiste in uno scambio di messaggi tra il dispositivo utente e il servizio di fulfillment, considerando però che quest'ultimo interagirà a sua volta con altri servizi legati alla skill in uso.

```
{
    "version": "1.0",
    "session": {
        "new": true,
        "sessionId": "amzn1.echo-api.session.99fefe80-d55f-428d-96ec-72ad89a86a52",
        "application": {
            "applicationId": "amzn1.ask.skill.6626d9d6-a9c0-4c7b-aea6-dbbe88e45df9"
        },
        "attributes": {},
        "user": {
            "userId": "amzn1.ask.account.AFJEIRJVDNDEAXINQXOWRF7KHQKQT44CW2YK2LNURT...",
            "accessToken": "eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCIsImtpZCI6Ildpdk9HdVN..."
        }
    },
    "request": {
        "type": "IntentRequest",
        "requestId": "amzn1.echo-api.request.49ac6136-040b-4539-9f7e-376cce429bfc",
        "locale": "it-IT",
        "timestamp": "2023-03-10T17:50:30Z",
        "intent": {
            "name": "CreditRefillIntent",
            "confirmationStatus": "NONE",
            "slots": {
                "PhoneNumber": {
                    "name": "PhoneNumber",
                    "confirmationStatus": "NONE"
                },
                "opCode": {
                    "name": "opCode",
                    "confirmationStatus": "NONE"
                },
                "Amount": {
                    "name": "Amount",
                    "confirmationStatus": "NONE"
                },
                "Name": {
                    "name": "Name",
                    "confirmationStatus": "NONE"
                }
            }
        },
        "dialogState": "STARTED"
    }
}
```


Figura 20: Oggetto JSON di esempio inviato dall'assistente Alexa al servizio di fulfillment nel momento in cui il comando richiesto dall'utente sia stato abbinato ad uno degli intenti gestibili dalla skill ricevente. In questo esempio in particolare, è stato chiesto al servizio di gestire l'intento di nome "CreditRefillIntent" e come si può notare non è stato valorizzato nella richiesta nessuno dei valori aggiuntivi (slots).

Durante l'interazione, il dispositivo utente invia al servizio di fulfillment una notifica, tramite i pacchetti citati in precedenza, per ogni cambiamento di stato che lo riguarda. Tra le tipologie di richieste inviate, le principali sono:

- LaunchRequest: inviata quando l'utente invoca l'applicazione in questione senza nessun comando specifico;

- IntentRequest: inviata quando l'utente usa un comando che può essere collegato ad un intento gestito dall'applicazione ricevente;
- AudioPlayer.*: inviata dal dispositivo alla skill per notificare lo stato attuale della riproduzione audio;
- SessionEndedRequest: inviata quando si notifica la chiusura dell'attuale sessione della skill in esecuzione. Può avvenire in seguito ad un'esplicita richiesta dell'utente, ad una risposta non gestibile o al verificarsi di un errore.

Di quest'ultime la IntentRequest ricopre un ruolo particolarmente importante per il servizio di fulfillment. Essa conterrà le informazioni relative all'intento riconosciuto dal servizio di NLU, permettendo alla skill ricevente di rispondere nella modalità consona (Figura 20).

Un'applicazione creata per l'assistente Amazon Alexa, prevede che lo sviluppatore crei il servizio di fulfillment, scrivendo esplicitamente il codice per gestire le richieste in ingresso e provvedere le risposte in uscita, al contrario di altri assistenti in cui quest'operazione avviene solitamente tramite interfacce grafiche apposite.

Il vantaggio di un approccio come quello di Alexa sta nel fatto che, se lo sviluppatore ha competenze di programmazione adeguate, può facilmente comunicare con altri servizi per rendere l'esperienza di dialogo più interattiva.

Nella realizzazione del progetto legato al lavoro di tesi, questa possibilità si rivela decisamente utile. Il prototipo in questione, infatti, dovrà sia avere accesso ad un database locale per salvare le preferenze espresse dall'utente durante l'utilizzo, che comunicare con servizi esterni. Nel particolare i due servizi principali saranno quello incaricato di gestire il collegamento tra i due account dell'utente (l'account Amazon collegato al dispositivo e l'account bancario), e quello che riceverà le operazioni bancarie richieste dall'utente al fine di eseguirle (Business Logic Server). Trattandosi di un prototipo il servizio che esegue le operazioni dell'utente, normalmente rappresentato da un server della banca in questione, è stato sostituito con un server di prova di cui sono state esposti pubblicamente dei punti di accesso.

Quest'ultimo è stato realizzato utilizzando Java come linguaggio di programmazione, e in particolare sfruttando il framework open source SpringBoot. Il ruolo principale di

questo servizio era quello di fornire un punto d'accesso ai servizi offerti (API endpoints).

Il tassello mancante in questo schema di interazione è proprio il collegamento tra l'utente che interagisce con il dispositivo Alexa e l'utente a cui va associata la transazione nel sistema della banca. La procedura che risolve questa mancanza prende il nome di *account linking*.

L'account linking è la procedura che permetterà al dispositivo Alexa con cui l'utente sta interagendo, di comunicare nella richiesta alcune informazioni utili al suo riconoscimento in un sistema di terzi.

Il protocollo attraverso il quale avviene l'account linking è OAuth.

OAuth è un protocollo che consente l'emissione di un token d'accesso da parte di un server autorizzativo ad un client di terze parti, se ovviamente a monte vi è stata l'autorizzazione del proprietario della risorsa di cui si richiede l'accesso.

La sua versione OAuth 2.0 (RFC 6749 [28]) è considerata ad oggi lo standard per concedere il servizio di delega dell'accesso.

Grazie a questo protocollo, un client di terze parti può accedere a specifiche risorse protette di un utente senza la necessità che quest'ultimo comunichi al client le credenziali di accesso al servizio che le protegge.

Gli attori coinvolti nell'interazione del protocollo sono:

- Utente: il proprietario delle risorse di cui si richiede l'accesso. Nel prototipo relativo al corrente lavoro, l'utente è il proprietario del conto con cui si vuole interagire;
- Client: l'applicazione che richiede l'accesso alle risorse. Nel prototipo corrisponde al dispositivo Alexa con cui l'utente sta interagendo;
- Fornitore del servizio: il server che contiene e protegge le risorse dell'utente. Nel prototipo è il servizio della banca capace di accedere al conto in questione e associarvi delle operazioni (Business Logic Server);
- Server autorizzativo: il server attraverso il quale l'utente può approvare o negare il permesso di accesso alle risorse. Sarà quest'ultimo a rilasciare al client un token di accesso valido da presentare al fornitore del servizio.

In un'applicazione Alexa, l'utente è tenuto a completare la procedura di account linking nel momento in cui abilita l'app in questione (abilitare una skill è il processo che consente all'utente di farne uso all'interno del suo assistente). Al momento dell'abilitazione, se la skill richiede l'account linking, il dispositivo con cui l'utente interagisce comunicherà con il server autorizzativo l'intenzione di iniziare la procedura. Quest'ultimo reindirizzerà il dispositivo ad una pagina in cui l'utente avrà la possibilità di autenticarsi. Le credenziali inserite dall'utente vengono poi comunicate al server del fornitore del servizio per essere verificate. Se richiesto, in questa procedura possono anche essere integrati altri fattori di autenticazione oltre alle semplici credenziali.

Una volta confermata l'identità dell'utente, il server autorizzativo si occuperà di rilasciare al client un codice di autorizzazione attraverso il quale quest'ultimo sarà in grado, quando necessario, di richiedere un token di accesso valido o aggiornarne uno preesistente (questo token prende il nome di JWT).

Il token in questione (JSON Web Token, JWT) altro non è che un oggetto JSON contenente le informazioni necessarie a identificarne il proprietario e i permessi ad esso concessi in un dato servizio.

Il token si compone di tre parti, l'intestazione, il carico e la firma.

L'intestazione contiene solitamente informazioni relative al tipo di token (in questo caso JWT) e il tipo di algoritmo usato per la firma dello stesso.

Il carico, invece, è la parte del token contenente i veri dati. Qui troviamo l'identificativo dell'utente proprietario del token, la data di emissione e scadenza, i servizi a cui il token è destinato e cosa permette di fare all'interno degli stessi, ecc.

Per assicurarsi che quest'ultimi non subiscano manomissioni da parte di attori malevoli, viene aggiunta la terza parte, quella contente la firma delle informazioni precedenti combinate ad un segreto (in questo caso la chiave privata del server autorizzativo).

Un esempio di firma utilizzando l'algoritmo HMAC SHA256 è:

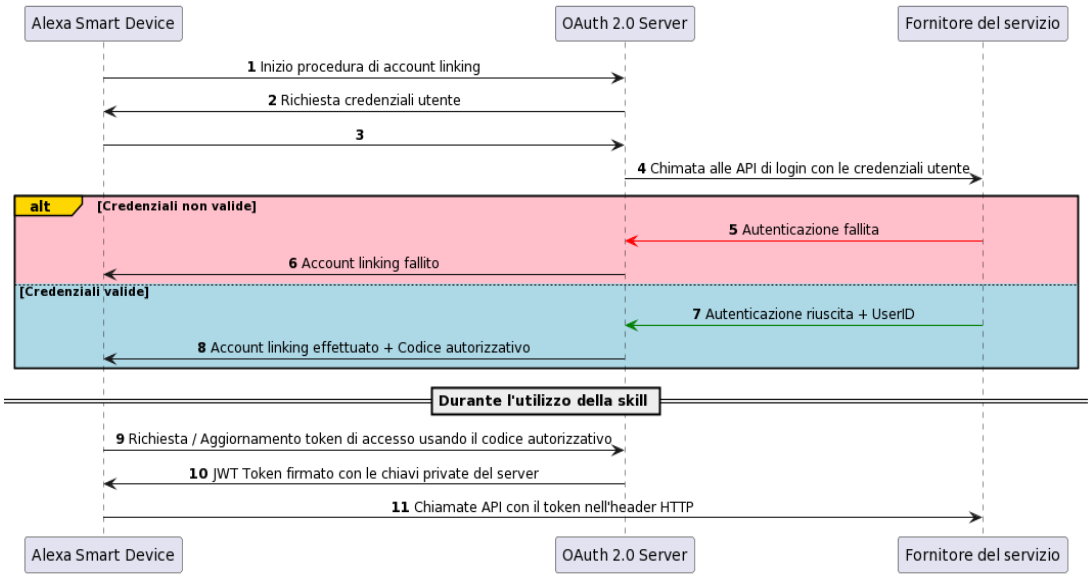

Figura 21: Rappresentazione tramite PlantUml della procedura di account linking effettuata nel prototipo.

firma = HMACSHA256 (CodificaBase64 (intestazione) + "." + CodificaBase64 (carico), segreto)

Ognuna delle precedenti parti viene poi codificata tramite il sistema Base64, ottenendo tre stringhe. La concatenazione di quest'ultime costituisce il JWT completo che verrà poi usato nelle comunicazioni tra le parti.

Da questo momento ogni richiesta effettuata dal client conterrà all'interno dell'oggetto JSON, scambiato con il servizio di fulfillment, il token d'accesso. Nel momento in cui il servizio di fulfillment andrà ad inoltrare la richiesta dell'utente al fornitore del servizio (Business Logic Server), aggiungerà nell'header del pacchetto HTTP il token di accesso, permettendo a quest'ultimo di validarlo e successivamente riconoscere l'utente a cui associare la richiesta.

È disponibile una schematizzazione dell'interazione descritta nella figura 21.

Prendendo come esempio l'oggetto JSON inviato nella richiesta in figura 20, si può notare come le informazioni relative al token di accesso sono contenute nel campo *accessToken* sotto il gruppo che identifica i dati della sessione.

In quel campo si trovano le tre stringhe codificate in Base64 che sono state menzionate precedentemente. A scopo esplicativo si decodifica di seguito il JWT contenuto nell'esempio.

Intestazione:


```
{
  "alg": "RS256",
  "typ": "JWT",
  "kid": "WivOGuSCJezVlEY1uCBgY"
}
```


Carico:


```
{
  "iss": "https://dev-d-fersino-reply.eu.auth0.com/",
  "sub": "auth0|1",
  "aud": [
    "alexa-bank-skill-api",
    "https://dev-d-fersino-reply.eu.auth0.com/userinfo"
  ],
  "iat": 1678290408,
  "exp": 1678376808,
  "azp": "4lfserO6bgnp0jZWy826cHxPktckKzu7",
  "scope": "openid profile read:user read:credit-refill write:credit-refill read:bank-movements write:bank-movements offline_access"
}
```


Ognuna delle informazioni contenuta in questo oggetto JSON non è modificabile, e ciò viene garantito dalla firma posta in aggiunta. Grazie ad esso, nel momento in cui il servizio di fulfillment comunica al servizio preposto l'intenzione di effettuare una particolare operazione, quest'ultimo può verificare la bontà della richiesta.

Tra i campi più rilevanti abbiamo:

- *iss*: indica il servizio che ha rilasciato il JWT, e quindi colui da contattare per verificarne l'integrità;

- *sub*: indica l'utente proprietario del JWT;
- *iat ed exp*: indicano rispettivamente la data in cui il JWT è stato emesso e quella in cui scade;
- *aud* e *scope*: indicano rispettivamente a che servizi è rivolto il JWT (audience) e quali sono le sue capacità operative all'interno di essi (*scope*).

Quella descritta finora è l'architettura completa di una skill sviluppata su Alexa.
Nel prototipo creato, oltre alle componenti precedentemente descritte, è stata aggiunta anche l'interazione con un database locale da parte del servizio di fulfillment.
Questo database, essendo direttamente raggiungibile senza interrogare servizi esterni, ed essendo legato all'account Amazon dell'utente ha permesso di migliorare in alcune parti l'esperienza dell'utente nell'interazione con la skill.
Nel caso specifico, è stato usato per il salvataggio di preferenze utente legate al flusso della conversazione e per creare delle scorciatoie nell'interazione.
Uno schema riassuntivo dell'architettura completa del prototipo è disponibile in figura 21.

2.a. Limiti e rischi

Nello schema evidenziato finora è facile notare come non vi sia una vera e propria distinzione tra gli utenti che interagiscono con un dispositivo.
Una volta conclusa la procedura di account linking viene definito un collegamento tra l'account Amazon dell'utente e il conto bancario a cui sta concedendo l'accesso. Da quel momento in avanti ogni interazione con i dispositivi Alexa in cui è collegato tale account, dal punto di vista del business logic server, è come fosse stata richiesta dall'utente proprietario del conto.

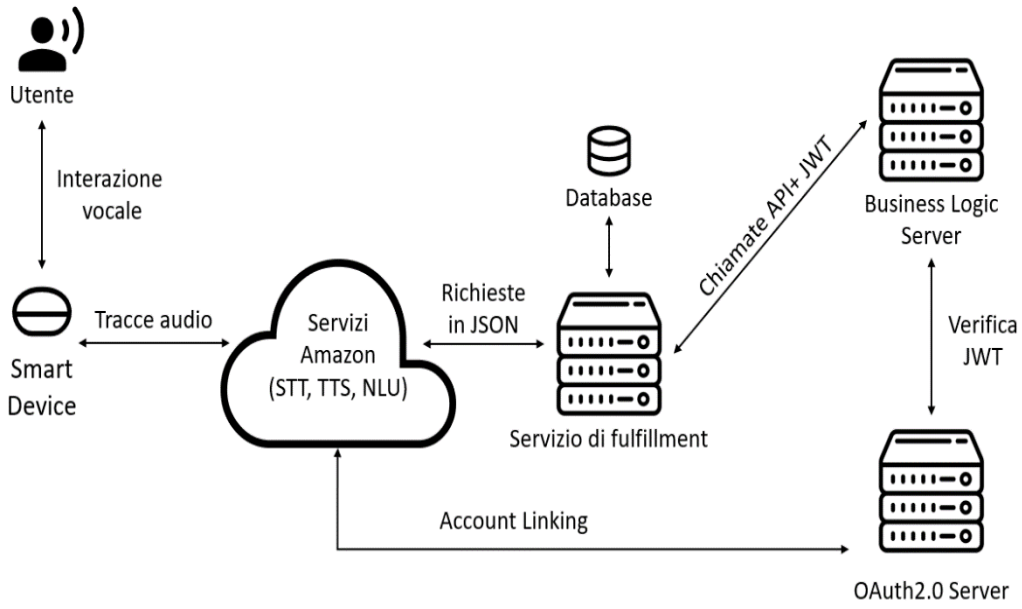

Figura 22: Rapprensentazione grafica dell'architettura della skill creata come prototipo per l'assistente Amazon Alexa.

La nota precedente può apparire poco rilevante se si suppone che il dispositivo in questione venga tenuto in casa, e quindi in un ambiente che, tendenzialmente, definiamo sicuro.

La supposizione di per sé non è errata, tuttavia in un settore come quello in analisi non si può ignorare una lacuna di questo tipo. In un'architettura come quella descritta, troviamo due gravi problemi strutturali di sicurezza. Il primo è il problema descritto in precedenza, chiunque abbia la capacità di interagire con il dispositivo, sarà in grado di richiedere l'esecuzione di transazioni al pari dell'effettivo proprietario del conto, anche senza la sua approvazione. Si vuole far notare inoltre, che adottando misure blande di sicurezza, come l'utilizzo di un codice o una password, per autorizzare le operazioni si attenua leggermente il problema ma senza risolverlo. Quest'ultimi dovranno essere pronunciati a voce e quindi facilmente udibili da chi si ha intorno. Anche prestando molta attenzione ai momenti con cui si interagisce con la skill, tutte le conversazioni

tra l'utente e quest'ultima vengono salvate in chiaro sull'account dello stesso, dando ad un attaccante un ulteriore punto di attacco.
Il secondo problema riguarda invece i token di accesso (JWT). Permettere al business logic server di prendere per legittime e autenticate tutte le richieste in arrivo solo perché affiancate da un JWT valido è molto rischioso.
In un'architettura come quella descritta non vi è alcun controllo aggiuntivo sulle richieste ricevute dal server della banca, e non può neanche essere introdotto senza rinunciare all'esperienza ricercata.
Un'attaccante che fosse in grado di intercettare il token di accesso di un utente potrà usarlo per commissionare operazioni fingendosi il proprietario del conto anche senza la necessità di interagire con il dispositivo Alexa dello stesso, almeno fino alla scadenza del token stesso.
Si vuole far notare inoltre che tutti i JWT rilasciati in un'architettura come questa, come si può notare anche in figura 22, sono gestiti e immagazzinati nei server della compagnia proprietaria dell'assistente (Amazon in questo esempio).
Un'attaccante sufficientemente capace potrebbe quindi attaccare i server in questione allo scopo di sottrarre tutti i JWT in possesso degli stessi. L'attacco in questione è decisamente complesso, e probabilmente sono pochi i soggetti in grado di bucare un sistema di sicurezza come quello di Amazon, tuttavia, considerando il vasto impatto che un attacco di questo tipo avrebbe, non può comunque essere ignorato.

2.b. Varianti proposte

Al fine di mitigare i pericoli rilevati durante l'analisi del primo prototipo sono state analizzate due varianti dell'idea originale che garantissero un livello di sicurezza adeguato.

Ovviamente, come citato in precedenza, l'introduzione di altre componenti al fine di raggiungere la sicurezza desiderata, modifica anche l'interazione dell'utente con il prototipo in questione.

Fino a questo punto il prototipo evolveva con l'obbiettivo di gestire l'interazione con l'utente solo tramite il canale vocale, senza ricorrere all'integrazione di altri dispositivi. Dopo le dovute analisi, e avendo testato in prima persona i limiti dell'architettura precedente, è stato valutato un cambio nell'idea dietro al prodotto, cercando le migliori soluzioni che garantissero un'esperienza quanto più simile alla precedente, ma con una sicurezza adeguata.

La prima soluzione prevede di cambiare l'utilità del prodotto. Quest'ultimo manterrà le sue funzioni per quanto riguarda le operazioni informative (le operazioni che forniscono all'utente informazioni sul conto, quali saldo o operazioni recenti), mentre non permetterà all'utente di effettuare operazioni dispositive (un trasferimento, una ricarica telefonica, ecc.), quest'ultime verranno solo prenotate sul suo profilo e rimarranno pendenti fino all'approvazione.

Si può notare come entrambi i problemi legati all'architettura originale vengono risolti con questa soluzione. Anche nel caso in cui un utente non autorizzato interagisca con il dispositivo collegato al conto, o nel caso vi sia un attacco per il quale il JWT venga rubato, non vi è un reale rischio per il proprietario. Quest'ultimo, come risultato dell'attacco, troverà nella propria applicazione di mobile banking solo una serie di operazioni pendenti che non approverà.

La seconda proposta prevede invece di verificare la transazione in maniera simile a quello che avviene per un acquisto online.

A seguito della chiamata effettuata dal servizio di fulfilment al business logic server, quest'ultimo, a seconda della configurazione, potrà richiedere l'autorizzazione dell'operazione tramite applicazione di mobile banking, oppure tramite l'invio di codice OTP sul numero di telefono associato al proprietario del conto. Nel caso di autorizzazione tramite applicazione la skill deve solo restare in attesa della risposta e notificare l'utente del risultato. Se invece si opta per l'uso di un codice OTP, sarà la skill stessa a ricoprire il ruolo di interfaccia preposta a riceverlo. L'utente comunicherà

il codice alla skill come prosecuzione del discorso fino a quel momento creato e quest'ultima lo inoltrerà al business logic server per la verifica.
Anche attraverso questa soluzione si risolvono i due problemi evidenziati in precedenza. In caso di attacco le operazioni necessitano comunque di un'autorizzazione manuale del proprietario.
Anche se entrambe le varianti definite risolvono i problemi di sicurezza identificati nell'architettura originale, presentano un bel cambiamento in quella che era l'idea alla base dell'interazione di questi prodotti.
Se si riprendono i benefici analizzati nel capitolo III, si può notare come essi vengano rispecchiati solo in parte dall'utilizzo di queste varianti.
Al fine di rispettare le idee definite in partenza e sfruttare tutti i benefici di questi prodotti, si propone di seguito un ulteriore soluzione, solo analizzata ma non implementata in questo lavoro di tesi, che prevede l'introduzione di un'autenticazione basata su biometriche vocali.
Si premette che la soluzione in questione non è stata implementata in quanto il punto di applicazione della stessa è sul dispositivo con cui l'utente interagisce e quindi non alla portata di uno sviluppatore, bensì del produttore dello stesso.
Introducendo un meccanismo di autenticazione basato su biometriche vocali si avrebbe la capacità di riconoscere un oratore sulla base di componenti univoche presenti nella voce.
Se si suppone che il dispositivo con cui l'utente interagisce sia in grado di effettuare un'identificazione di questo tipo, si può notare come i problemi presenti nell'architettura originale verrebbero risolti senza l'ausilio di componenti esterne, e quindi con un'interazione unicamente vocale.
Un servizio di riconoscimento basato su biometriche vocali andrà ad associare ad una traccia audio un valore di confidenza sul fatto che quest'ultima appartenga o meno ad un certo individuo precedentemente registrato. Se supponiamo di impedire la manomissione di tale valore, ad esempio aggiungendoci un hash firmato, possiamo facilmente creare un'architettura in cui quest'ultimo viene inviato insieme al JWT al servizio preposto ad effettuare le operazioni (business logic server). Quest'ultimo deve

solo verificare che entrambi non siano stati manomessi, e che il valore di confidenza ricevuto sia superiore ad una certa soglia, in base all'operazione richiesta.
Nel caso in cui non vi siano state manomissioni la sicurezza dell'architettura nel complesso dipende dalla bontà del riconoscimento performato sulla traccia audio dell'utente.
Viene fornita di seguito una spiegazione sul funzionamento di una tecnologia basata su biometriche vocali ed alcuni utilizzi pratici attualmente presenti.

2.c. Le biometriche vocali

Le tecnologie basate su biometriche vocali verificano l'identità di un oratore basandosi sulla sua voce.
Con biometriche vocali ci si riferisce alle caratteristiche distintive della voce umana, determinate dall'anatomia dell'oratore e dal comportamento che adotta durante una conversazione.
Tra le componenti che influenzano le biometriche vocali abbiamo la forma e la dimensione di bocca e gola, come viene condotto un discorso, ad esempio la velocità che si usa nella parlata, il tono della voce, ecc.
L'idea alla base di queste tecnologie è quella di usare una o più tracce audio di un utente per estrarre le caratteristiche univoche che lo identificano al fine di creare un modello dello stesso (anche detto impronta vocale), che verrà salvato e usato come riferimento.
A seconda della tecnica utilizzata, queste caratteristiche sono identificate in modo diverso. Generalmente si elabora la voce nel dominio della frequenza. Un esempio è analizzare la distribuzione della potenza del segnale vocale distribuita su un intervallo di frequenze (*PSD*).

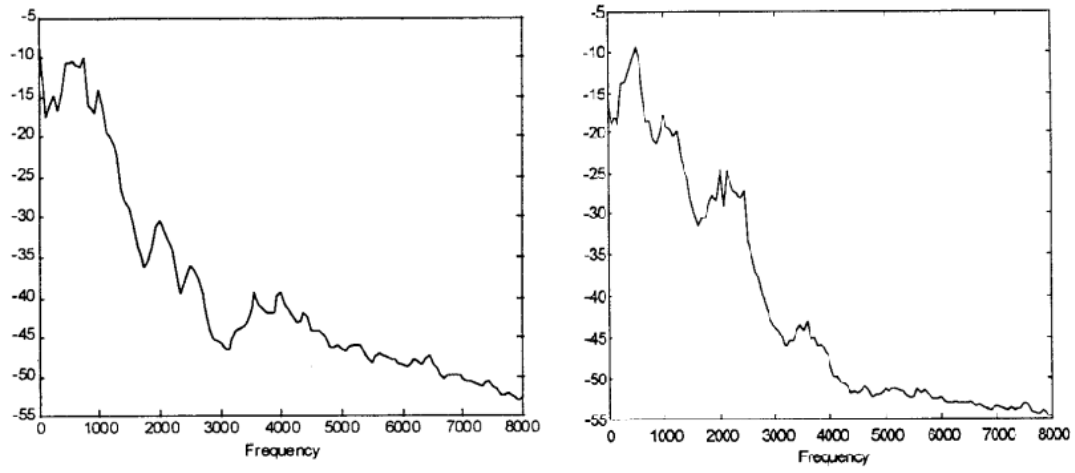

Figura 23: PSD di due oratori differenti presi come campioni. Fonte [47]

Quest'ultima è una tecnica semplice, utile per chiarificare l'obbiettivo di queste elaborazioni. Si può notare in figura 23, come i PSD di due individui diversi, pur presentando un andamento simile, sono facilmente riconoscibili.

Nell'esempio considerato il PSD rappresenta quella che precedentemente è stata definita come impronta vocale.

Nelle interazioni successive la nuova impronta vocale calcolata viene confrontata con quella precedentemente salvata nel sistema, assegnandogli un certo grado di somiglianza. Quest'ultimo, tendenzialmente, è il valore utilizzato da altri componenti del sistema per decretare, sulla base di una certa soglia, se l'oratore corrente può essere considerato lo stesso che ha creato il modello di riferimento o meno.

Nell'implementazione di queste tecniche non è richiesto nessun dispositivo specifico o costoso, è sufficiente un semplice microfono (già presente nella maggior parte dei dispositivi con cui ad oggi si interagisce).

Il riconoscimento basato su biometriche vocali può avvenire sulla base di due stili diversi:

- Riconoscimento dipendente dal testo: il sistema viene allenato con una serie di frasi pronunciate dall'utente, e successivamente si aspetta di ricevere una di queste specifiche frasi per procedere con il riconoscimento. Non è in grado di

riconoscere un oratore tramite una generica interazione vocale. Spesso vengono anche chiamate *password vocali*;

- Riconoscimento indipendente dal testo: il sistema viene allenato con una o più frasi pronunciate dall'utente, per poi riconoscerlo indipendentemente dall'interazione in questione. Un utente può pronunciare una frase qualsiasi, anche non facente parte di quelle salvate, ed essere comunque riconosciuto (a patto che la frase sia sufficientemente lunga per il dato sistema). Sono le soluzioni più usate e su cui vi è lo sviluppo maggiore.

Nonostante ad oggi sia considerata ancora un'autenticazione di nicchia con pochi utilizzi, i risultati ottenuti grazie ai recenti sviluppi tecnologici migliorano costantemente, aumentando le aspettative sul futuro di questa tecnologia.

Il NIST (National Institute of Standards and Technology) propone ogni anno una sfida pubblica allo scopo di migliorare e sviluppare queste tecnologie (Conversational Telephone Speech Speaker Recognition Evaluation, CTS SRE [29]). L'obbiettivo di questa sfida è identificare un particolare oratore durante una data traccia audio.

Dai risultati dai migliori modelli in questa competizione, si può notare come attualmente i valori di EER (Equal Error Rate) si attestino tra i 2 e i 2,5 punti percentuale (Tabella 3). Valori ancora alti se paragonati ad altre tipologie di autenticazioni basate su biometriche, come riconoscimento di impronte digitali e riconoscimento facciale, ma comunque validi in alcuni contesti specifici.

Tabella 3: Risultati ottenuti dai primi 5 classificati nella NIST CTS SRE. Fonte [29]

Posizione nella competizione	Team	Timestamp	EER [%]	MIN_C	ACT_C
1	THUEE	20210921-005400	2.23	0.063	0.076
2	STC	20210826-020705	2.48	0.074	0.079
3	TEAM-MGQG-50	20210930-003746	2.01	0.078	0.085
4	I4U	20210528-015904	2.53	0.077	0.094
5	TEAM-CDPE-28	20210616-024452	2.55	0.087	0.101

Ad oggi uno degli utilizzi maggiori di questa tecnologia è quello di mezzo di riconoscimento di un oratore in un call center.

Un esempio di questo utilizzo è il caso di Citigroup, società di servizi finanziari americana, che ha implementato un riconoscimento basato su biometriche vocali per gestire le operazioni bancarie richieste dai propri clienti tramite chiamate telefoniche (il software in questione è chiamato NICE) [30].

Grazie a questo servizio, il call center, autentica l'utente già dall'inizio della telefonata riconoscendolo dopo pochi secondi di conversazione. Questo ha permesso a Citigroup di migliorare il servizio di call center rimuovendo le tediose domande che venivano utilizzate in precedenza per il medesimo scopo (Qual è il nome del tuo primo animale domestico? Squadra di calcio preferita? ...).

Un ulteriore vantaggio ottenuto grazie a NICE è il riconoscimento immediato dei truffatori. Come intuibile oltre a utilizzare un modello salvato per riconoscere un utente legittimo, si può anche riconoscere un attore malevolo.

Nel caso specifico di NICE, viene usata una lista nera di truffatori, ottenuta analizzando le precedenti chiamate di frode. Una telefonata viene subito etichettata come pericolosa se il chiamante viene assimilato ad uno in lista nera e gestita di conseguenza.
Anche Amazon, ha un servizio simile basato su biometriche vocali chiamato Amazon Connect Voice ID [31], purtroppo il suo utilizzo è ancora limitato ai call center e non è stato implementato nell'assistente digitale della stessa compagnia.

V. CONCLUSIONI

L'idea dalla quale è nato il lavoro di tesi era quella di migliorare l'attuale interazione tra un cliente e il sistema bancario attraverso l'ausilio di tecnologie vocali, in particolare gli assistenti digitali.

A seguito delle analisi fatte in merito ai benefici che se ne potrebbero ottenere e dopo aver valutato le posizioni intraprese da altri istituti bancari di fama internazionale, si è deciso di analizzare a fondo l'architettura tecnologica alla base dei dispositivi ad oggi commercializzati, e comprendere quale fosse il percorso di integrazione migliore per sfruttare le potenzialità di questi dispositivi, senza rinunciare ai requisiti di sicurezza necessari in un settore come quello in esame.

Come si è potuto notare dalle varie analisi effettuate nei capitoli precedenti, se si vuole integrare gli assistenti digitali già commercializzati (Amazon Alexa, Google Home, …) ai servizi offerti dal mondo bancario sfruttando la mera interazione vocale con quest'ultimi, non si può raggiungere un livello di sicurezza soddisfacente per il settore. La motivazione principale di ciò è data dal fatto che non vi è negli stessi un meccanismo di autenticazione dell'utente che vi interagisce, costringendo chi crea applicazioni che hanno questi dispositivi a supporto, a ricorrere all'utilizzo di dispositivi esterni per raggiungere lo scopo, venendo meno alla richiesta di ottenere un'interazione unicamente vocale.

Dall'analisi effettuata le soluzioni architetturali che arginano il problema prevedono quasi sempre l'utilizzo dello smartphone in almeno un momento dell'interazione totale. Nonostante il suo utilizzo nelle soluzioni proposte sia marginale e ristretto a semplici operazioni della durata di pochi secondi, va comunque a minare l'idea iniziale e parte dei benefici che si ottenevano da essa. Nei capitoli precedenti è stato evidenziato come gli utenti più anziani, o comunque coloro che sono poco a loro agio con il mondo tecnologico, avessero potuto beneficiare da queste tecnologie, soprattutto se integrate

con i servizi del settore bancario. Beneficio che si perde nelle attuali risoluzioni del problema.

Si possono riassumere i risultati ottenuti in questo lavoro dicendo che l'attuale evoluzione tecnologica degli assistenti digitali non ha ancora raggiunto il livello di maturità tale da permettere l'integrazione di quest'ultimi in settori con un elevato livello di sicurezza, in cui il riconoscimento e l'autenticazione dell'utente ricoprono un ruolo cruciale.

Il problema di sicurezza principale delle architetture analizzate risiede proprio nella mancanza di un meccanismo di autenticazione dell'utente disponibile nativamente. È necessario che quest'ultimo sia nativo in quanto la traccia audio dell'utente, come visto nei capitoli precedenti, è gestita nella prima fase dell'elaborazione e quindi non disponibile ad uno sviluppatore esterno.

È noto che l'autenticazione di un utente può avvenire sulla base di ciò che possiede, ciò che conosce o ciò che è. Senza la possibilità di analizzare la traccia audio viene meno la terza opzione, inoltre, essendo che l'interazione avviene tramite il canale vocale, e quindi in chiaro, perde di significato anche autenticare l'utente sulla base di ciò che conosce (esempio di Capital One), forzando l'utilizzo dell'unica opzione rimanente (usare uno smartphone nel processo di autenticazione come soluzioni analizzate nel quarto capitolo).

Come sostenuto più volte, il problema di quest'autenticazione rispetto alle altre è che mina l'idea originale di un'interazione unicamente vocale, con i benefici che ne derivavano. Tuttavia, al momento è l'unica su cui si può ripiegare.

Lavori futuri

Per concludere il lavoro di tesi sono stati identificati i meccanismi di autenticazione basati su biometriche vocali come dei candidati validi per far fronte alla serie di problemi riscontrati.

Come evidenziato anche nell'analisi effettuata, nell'attuale stato di avanzamento tecnologico, non possono essere considerati al pari di altri meccanismi di

autenticazione più noti (riconoscimento facciale o dell'impronta digitale), anche se rappresentano dei vantaggi di utilizzo rispetto a quest'ultimi.

Grazie all'applicazione di questi sistemi agli assistenti digitali, si potrebbero sbloccare le vere potenzialità degli stessi, permettendone l'integrazione con diversi settori ad oggi tagliati fuori per le elevate richieste di sicurezza.

Il settore bancario rappresenta probabilmente uno dei settori che beneficerebbe di più da tali tecnologie.

Con il maturare delle stesse, potrebbe essere interessante l'analisi di un'architettura di interazione in cui la traccia audio dell'utente viene analizzata e autenticata tramite l'ausilio di un modello vocale precedentemente creato, per poi comunicare ai componenti che seguiranno un grado di affidabilità del riconoscimento.

Il punteggio ricevuto, magari accompagnato anche da un hash crittografico che ne certifichi l'autenticità, potrebbe poi essere usato nella comunicazione con i sistemi di backend per ottenere uno schema di interazione sicuro e trasparente all'utente.

VI. Bibliografia

[1] M. McCaffrey, P. Hayes, M. Hobbs e J. Wagner, «Consumer Intelligence Seri Prepare for the voice revolution,» *PwC Consumer Intelligence Series,* 2018.

[2] V. Petrock, «Voice Assistant and Smart Speaker Users 2020,» Insider Intelliger Inc., 11 2020. [Online]. Availab https://www.insiderintelligence.com/content/voice-assistant-and-smart-speaker-users-2020.

[3] Baidu, «Number of monthly voice queries of the Chinese voice assistant DuerOS Baidu from 2nd quarter 2018 to 1st quarter 2021 (in billions),» 2021. [Onlin Available: https://www.statista.com/statistics/1080006/baidu-dueros-voi assistant-monthly-query-number/. [Consultato il giorno 20 01 2023].

[4] Wikipedia, «Riconoscimento vocale,» [Online]. Availab https://it.wikipedia.org/wiki/Riconoscimento_vocale. [Consultato il giorno 10 2023].

[5] A. Trivedi, N. Pant, P. Shah, S. Sonik e S. Agrawal, «Speech to text and text speech recognition systems-Areview,» *IOSR-JCE,* 2018.

[6] Wikipedia, «Acoustic model,» [Online]. Availab https://en.wikipedia.org//wiki/Acoustic_model. [Consultato il giorno 10 01 2023

[7] Wikipedia, «Language model,» [Online]. Availab https://en.wikipedia.org/wiki/Language_model. [Consultato il giorno 10 01 2023

[8] M. Gales e S. Young, «The application of hidden Markov Models in spee recognition,» *Foundations and Trends in Signal Processing,* vol. 1, 2007.

[9] A. Amberkar, P. Amberkar, G. Deshmukh e P. Dave, «Speech Recognition usi Recurrent Neural Networks,» in *Proceedings of the 2018 International Conferer on Current Trends towards Converging Technologies, ICCTCT 2018*, 2018.

[10] A. Graves, S. Fernández, F. Gomez e J. Schmidhuber, «Connectionist tempo classification: Labelling unsegmented sequence data with recurrent neu networks,» in *ACM International Conference Proceeding Series*, 2006.

[11] A. Graves, A. R. Mohamed e G. Hinton, «Speech recognition with deep recurr neural networks,» in *ICASSP, IEEE International Conference on Acoustics, Spee and Signal Processing - Proceedings*, 2013.

[12] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Ryba A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pund K. C. Sim e T. Bagby, «Streaming End-to-end Speech Recognition for Mob Devices,» in *ICASSP, IEEE International Conference on Acoustics, Speech a Signal Processing - Proceedings*, 2019.

[13] A. Graves, «Sequence Transduction with Recurrent Neural Networks,» 2012.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kai e I. Polosukhin, «Attention is all you need,» in *Advances in Neural Informati Processing Systems*, 2017.

[15] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig e M. L. Seltzer, «Transform Based Acoustic Modeling for Hybrid Speech Recognition,» in *ICASSP, IE International Conference on Acoustics, Speech and Signal Processing Proceedings*, 2020.

[16] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa e T. Nakata «Improving transformer-based end-to-end speech recognition with connectior temporal classification and language model integration,» in *Proceedings of Annual Conference of the International Speech Communication Associati INTERSPEECH*, 2019.

[17] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo e S. Kum «Transformer Transducer: A Streamable Speech Recognition Model w

Transformer Encoders and RNN-T Loss,» in *ICASSP, IEEE Internatio Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020.

[18] Wikipedia, «Sintesi vocale,» [Online]. Availab https://it.wikipedia.org/wiki/Sintesi_vocale. [Consultato il giorno 10 01 2023].

[19] Wikipedia, «Natural-language understanding,» [Online]. Availab https://en.wikipedia.org/wiki/Natural-language_understanding. [Consultato giorno 10 01 2023].

[20] J. Devlin, M. W. Chang, K. Lee e K. Toutanova, «BERT: Pre-training of de bidirectional transformers for language understanding,» *NAACL HLT 2019 - 20 Conference of the North American Chapter of the Association for Computatio Linguistics: Human Language Technologies - Proceedings of the Conference,* v 1, 2019.

[21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee e L. Zettlemoy «Deep contextualized word representations,» 2018.

[22] L. Patrício, R. P. Fisk e J. Falcão Cunha, «Improving satisfaction with bank serv offerings: Measuring the contribution of each delivery channel,» *Managing Serv Quality: An International Journal,* vol. 13, 2003.

[23] M. Hanley e S. Azenkot, «Understanding the Use of Voice Assistants by Ol Adults,» 2021.

[24] Capital One, «Alexa Terms & Conditions,» [Online]. Availab https://www.capitalone.com/digital/alexa/terms/. [Consultato il giorno 1 03 2023

[25] Bank of America, «Erica - Virtual Financial Assistant from Bank of Americ [Online]. Availab https://promotions.bankofamerica.com/digitalbanking/mobilebanking/erica#discl ure-1882890920. [Consultato il giorno 01 03 2023].

[26] Bank of America, «Bank of America's Erica Tops 1 Billion Client Interactions, N Nearly 1.5 Million Per Day,» [Online]. Availab https://newsroom.bankofamerica.com/content/newsroom/press-

releases/2022/10/bank-of-america-s-erica-tops-1-billion-client-interactions--now-.html. [Consultato il giorno 01 03 2023].

[27] Google, «Conversational Actions sunset overview,» [Online]. Availab https://developers.google.com/assistant/ca-sunset. [Consultato il giorno 01 2023].

[28] IETF, «The OAuth 2.0 Authorization Framework,» [Online]. Availab https://www.rfc-editor.org/rfc/rfc6749. [Consultato il giorno 01 03 2023].

[29] NIST, «NIST Speaker Recognition Evaluation (SRE),» [Online]. Availab https://sre.nist.gov/. [Consultato il giorno 01 03 2023].

[30] T. Groenfeldt, «Citi Uses Voice Prints To Authenticate Customers Quickly A Effortlessly,» Forbes, [Online]. Availab https://www.forbes.com/sites/tomgroenfeldt/2016/06/27/citi-uses-voice-prints-to authenticate-customers-quickly-and-effortlessly/#752c76f1109c. [Consultato giorno 01 03 2023].

[31] Amazon, «Use real-time caller authentication with Voice ID,» [Online]. Availab https://docs.aws.amazon.com/connect/latest/adminguide/voice-id.html. [Consult il giorno 01 03 2023].

[32] S. Kodituwakku, «BIOMETRIC AUTHENTICATION: A REVIEW,» *Internatio Journal of Trend in Research and Development,* vol. 2, pp. 113-123, 2015.

[33] Amazon, «Understand Custom Skills | Alexa Skill Kit,» 2022. [Online]. Availab https://developer.amazon.com/it-IT/docs/alexa/custom-skills/understanding-custom-skills.html.

[34] A. Ponticello, M. Fassl e K. Krombholz, «Exploring authentication for securi sensitive tasks on smart home voice assistants,» in *roceedings of the 17th Symposi on Usable Privacy and Security, SOUPS 2021*, 2021.

[35] J. S. Edu, J. M. Such e G. Suarez-Tangil, «Smart Home Personal Assistants: Security and Privacy Review,» *ACM Computing Surveys,* vol. 53, 2021.

[36] N. Abdi, K. M. Noura e J. M. Such, «More than smart speakers: Security and priva perceptions of smart home personal assistants,» in *Proceedings of the 1. Symposium on Usable Privacy and Security, SOUPS 2019*, 2019.

[37] G. Terzopoulos e M. Satratzemi, «Voice assistants and smart speakers in everyc life and in education,» *Informatics in Education,* vol. 19, 2020.

[38] S. Kayte, M. Mundada e J. Gujrathi, «Hidden Markov Model based Spee Synthesis: A Review,» *International Journal of Computer Applications,* vol. 1 2015.

[39] L. Dong, S. Xu e B. Xu, «Speech-transformer: A no-recurrence sequence- sequence model for speech recognition,» in *ICASSP, IEEE International Conferer on Acoustics, Speech and Signal Processing - Proceedings*, 2018.

[40] H. Sak, A. Senior, K. Rao e F. Beaufays, «Fast and accurate recurrent neural netwo acoustic models for speech recognition,» in *Proceedings of the Annual Conferer of the International Speech Communication Association, INTERSPEECH*, 2015.

[41] eMarketer; Insider Intelligence, «Number of voice assistant users in the Uni States from 2017 to 2022 (in millions),» 2022. [Online]. Availab https://www.statista.com/statistics/1029573/us-voice-assistant-users/. [Consultato giorno 20 01 2023].

[42] PYMNTS, «Share of consumers making purchases using voice-activated devi while performing their daily routines in the United States from 2018 to 2020,» 20 [Online]. Available: https://www.statista.com/statistics/1234149/share-of- consumers-making-purchases-using-voice-activated-devices/. [Consultato il gio 20 01 2023].

[43] Capgemini, «Benefits organizations receive from adopting voice assistan worldwide as of 2019, by sector,» 2019. [Online]. Availab https://www.statista.com/statistics/1073798/worldwide-high-benefit-voice- assistant-business/. [Consultato il giorno 20 01 2023].

[44] Speechmatics, «Industries that will increase their use and application of vo technology in the next three to five years worldwide as of 2021,» 2021. [Onlin Available: https://www.statista.com/statistics/1208460/global-voice-technolo future-industries/. [Consultato il giorno 20 01 2023].

[45] Strategy Analytics, «Smart speaker unit shipments worldwide from 3rd quarter 20 to 1st quarter 2022, by vendor (in millions),» 2022. [Online]. Availab https://www.statista.com/statistics/792598/worldwide-smart-speaker-unit-shipment/. [Consultato il giorno 10 03 2023].

[46] Oracle, «Che cos'è un assistente digitale? | Oracle Italia,» [Online]. Availab https://www.oracle.com/it/chatbots/what-is-a-digital-assistant/. [Consultato giorno 10 01 2023].

[47] G. Venayagamoorthy, V. Moonasar e K. Sandrasegaran, «Voice recognition usi neural networks,» in *Proceedings of the 1998 South African Symposium Communications and Signal Processing-COMSIG '98 (Cat. No. 98EX214)*, 1998