OAR Journal

Firenze University Press – USiena PRESS
https://riviste.fupress.net/index.php/oarj

Speech Technology

Phoné: Una iniziativa per la creazione di un dataset per il riconoscimento automatico dell'italiano parlato

Phoné: An Initiative to Develop a Dataset for the Automatic Recognition of Spoken Italian

Gianpaolo Coro[1], Francesco Cutugno[2], Loredana Schettino[3], Emilia Tanda[2,*], Alessandro Vietti[3], Vincenzo Norman Vitale[2]

[1] *Istituto di Scienze e Tecnologie dell'Informazione 'A. Faedo' del Consiglio Nazionale delle Ricerche, Pisa, Italia*
[2] *Centro Interdipartimentale di ricerca Urban/Eco, Università degli Studi di Napoli Federico II, Italia*
[3] *Libera Università di Bolzano, Italia*
E-mail: gianpaolo.coro@isti.cnr.it; francesco.cutugno@unina.it; loredana.schettino@unibz.it; emilia.tanda@unina.it; alessandro.vietti@unibz.it; vincenzonorman.vitale@unina.it
*Corresponding author

Abstract. Large Language Models (LLM) have revolutionised natural language processing and its applications. However, high-performance LLMs require copious data and computing resources for their development and are rarely public. This also concerns Large Acoustic Models (LAM) for processing spoken language. The Phoné initiative seeks to build an open Italian speech dataset to advance Automatic Speech Recognition (ASR) systems and support public research. Spearheaded by institutions in Naples, Pisa, and Bolzano, the project gathers diverse Italian audio sources and applies advanced ASR architectures, including supervised and self-supervised models. This paper details Phoné's dataset creation, ASR model evaluation, and ethical considerations, aiming to democratise access to Italian-language resources and foster innovation in ASR technologies.

Keywords: speech technology, automatic speech recognition, Large Language Models, dataset.

OPEN ACCESS

Citation: Coro, G., Cutugno, F., Schettino, L., Tanda, E., Vietti, A., & Vitale, V. N. (2025). Phoné: Una iniziativa per la creazione di un dataset per il riconoscimento automatico dell'italiano parlato, *Oral Archives Journal*, 1: 89-107. doi: 10.36253/oar-3340

Received: June 15, 2024
Accepted: October 4, 2024
Published: February 20, 2025

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

ORCiD
GC: 0000-0001-7232-191X
FC: 0000-0001-9457-6243
LS: 0000-0002-3788-3754
ET: 0009-0008-6480-301X
AV: 0000-0002-4166-540X
VNV: 0000-0002-0365-8575

Ai soli fini concorsuali, l'attribuzione dei paragrafi è la seguente: Francesco Cutugno per i §§ 1, 3, 4, 4.3; Loredana Schettino per i §§ 1, 3.1; Emilia Tanda per i §§ 1, 2, 3, 4.1, 4.2, 4.3; Norman Vincenzo Vitale per il § 4.2. Tutti gli autori sono responsabili per il § 5.

1. INTRODUZIONE: MODELLI DI LINGUAGGIO NELLO SCRITTO E NEL PARLATO

"*Can machines think*?" (Turing 1950, 433). Con questa domanda si apre l'articolo *Computing machinery and intelligence* del 1950 di Alan Turing, considerato uno dei padri dell'informatica e del ramo di ricerca che ha aperto la strada alla possibilità di sviluppare un sistema di Intelligenza Artificiale (IA). All'interno di questo articolo, il celebre matematico presenta *the imitation game*, comunemente noto come il test di Turing (Turing 1950, 433-4). Lo scopo era quello di valutare se una macchina fosse in grado di manifestare un comportamento intelligente, inteso come un'imitazione del comportamento umano. Al centro di questa discussione vengono poste l'importanza e la complessità del linguaggio, elemento costitutivo dell'esperienza umana e da sempre considerato uno dei tratti distintivi della manifestazione del pensiero con le sue variazioni, sfumature e ambiguità. Nel campo dell'IA, l'idea di creare macchine in grado di comprendere e generare linguaggio è considerata una pietra miliare. I sistemi di elaborazione del linguaggio naturale (*Natural Language Processing,* NLP) permettono di svolgere compiti linguistici (*task*) di diversa natura, fra cui compiti generativi come la traduzione automatica, la generazione di riassunti e la generazione di risposte (per una rassegna in italiano si veda Ježek e Sprugnoli 2023, 165-92). Al fine di svolgere un qualsiasi *task* linguistico nell'ambito dell'NLP, abbiamo bisogno di un input, costituito dal set di dati linguistici che vogliamo analizzare, e di un modello computazionale che ci permette di ottenere in output i dati linguistici analizzati, come, ad esempio, l'analisi sintattica di un testo dato in input, oppure la classificazione delle parti del discorso, o ancora la disambiguazione di alcune parole presenti nel set di dati e così via. Un modello computazionale è centrale nella risoluzione dei *task* e si configura come un potente oggetto di calcolo che è capace di eseguire automaticamente il compito affidatogli nella maniera più efficiente secondo le sue possibilità. Esistono diversi tipi di modelli computazionali che, in base al tipo di architettura di base, possono essere suddivisi in due macrocategorie:

a) i modelli computazionali simbolici – oramai in disuso per alcuni compiti specifici come il riconoscimento del parlato – che si basano su categorie esplicite, ovvero set di regole definite a priori da esperti (per una rassegna in italiano cfr. Nissim e Pannitto 2022, 64);
b) i modelli statistici che, al contrario, si basano su calcoli probabilistici e sull'osservazione di ingenti quantità di dati (Ježek e Sprugnoli 2023).

Nell'ambito dell'NLP, i modelli computazionali di tipo statistico che sono in grado di riconoscere e generare il linguaggio umano, di predire la probabilità di sequenze di parole (*word prediction*) o generare nuovo testo basato su un dato input (Chang et al. 2024, 3) sono definiti modelli del linguaggio (*Language Models*, LM). Per i testi scritti costituiscono la base dell'elaborazione, ma sono utilizzati anche nell'analisi automatica del parlato. Questi modelli possono essere suddivisi a loro volta in due tipologie: i modelli statistici classici e i modelli neurali. I primi sono in grado di prevedere la parola successiva in base al contesto più recente, spesso definito anche 'storia' o 'cotesto locale'. A questa categoria appartengono i LM basati su N-grammi, definiti tali in quanto il cotesto presenta n-1 parole, che permettono di prevedere l'n-sima parola. I modelli neurali più recenti invece si basano sulle *Deep Neural Network* (DNN), strutture create sul modello delle reti neurali umane, in cui il calcolo è distribuito fra i vari nodi organizzati in strati interconnessi (Graves 2014; Goodfellow, Bengio, e Courville 2016).

I principali metodi di apprendimento automatico includono l'apprendimento supervisionato *(supervised*) e l'apprendimento non supervisionato (*unsupervised*). L'apprendimento supervisionato avviene fornendo al modello un *training set* etichettato manualmente, ovvero una parte dei dati linguistici che intendiamo analizzare insieme ai rispettivi *output*, cioè i risultati che vogliamo ottenere dall'analisi. Questi dati fungono da esempio per il modello che, in questo modo, apprende come classificarli correttamente osservando le coppie *input-output* che gli sono state fornite. Ponendo il caso di un compito di etichettatura delle parole di un testo rispetto alle relative parti del discorso, un esempio di coppie, *input-output*, di addestramento ('input', 'OUTPUT') sarebbe: ('allora', 'ADV'), ('Sara', 'PROPN'), ('hai', 'VERB'), ('presente', 'ADJ'), ('i', 'DET'), ('limoni', 'NOUN'). Sulla base di queste osservazioni il modello impara ad associare a dati in *input*, come 'quindi' 'devi' 'andare' 'dalla' 'tua' 'sinistra' 'verso' 'la' 'tua' 'destra', l'*output* stimato: 'ADV', 'AUX', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'DET', 'NOUN'[1]. Al contrario, nell'apprendimento non supervisionato è il modello stesso che partendo dai dati in input (sprovvisti del corrispettivo output) cerca di astrarre e identificare strutture significative per classificare i dati. Nel dettaglio questa tipologia di apprendimento viene utilizzata per predire la parola successiva come facevano gli N-grammi, ma con cotesto di lunghezza non limitata *a priori* a poche parole, superando i limiti di calcolo e complessità implicita nei sistemi precedenti.

Negli ultimi anni hanno assunto una rilevanza sempre maggiore, sia in ambito del *Natural Language Processing* che in ambito industriale, i *Large Language Models* (LLM, Chang et al. 2024), modelli del linguaggio il cui addestramento si basa su rappresentazioni vettoriali di enormi quantità di unità lessicali (nell'ordine dei miliardi). Le architetture allo stato dell'arte su cui si basano gli LLM consistono di moduli computazionali chiamati *Transformers* che sono costituiti da due componenti principali: la prima, l'*encoder,* ha come scopo creare una rappresentazione dei dati forniti; la seconda, il *decoder*, a partire da tali rappresentazioni genera l'output. Tale architettura è la prima basata sul meccanismo di *self-attention*. Nella fase di calcolo il meccanismo di *self-attention* rivela quali parole nella sequenza sono di maggiore interesse, ovvero contribuiscono in misura maggiore a definire e collegare le parole nel contesto linguistico. Gli LLM, negli ultimi anni, hanno stravolto completamente il nostro modo di approcciarci al mondo di internet e dei dispositivi digitali. Tuttavia, sebbene questi modelli basati su calcoli probabilistici sembrino cogliere dinamiche di uso della lingua, non ne comprendono realmente il funzionamento, imparano a svolgere compiti senza aver bisogno di comprenderli, per cui di fatto non riescono a simulare perfettamente la ricchezza e la variabilità dei comportamenti linguistici umani (caratterizzati da elementi non espliciti e lineari quali referenze, impliciti, ambiguità ecc.). I modelli che raggiungono le prestazioni migliori richiedono ingenti risorse per essere prodotti – sia in termini di dati per l'addestramento, sia in termini di risorse di calcolo – e perciò non sono quasi mai di dominio pubblico. Questi modelli, chiamati anche Modelli Generativi (*Generative Large Language Models* – GLLM), utilizzano ancora una volta la tecnica di prevedere la parola successiva dato il cotesto precedente, ma questa volta iterano potenzialmente all'infinito questa capacità, fino a generare un testo di senso compiuto (nella maggior parte dei casi) in risposta ad una richiesta fatta dall'utente (*prompt*).

[1] Esempio di compito di annotazione delle parti del discorso (*POS-tagging*) effettuato mediante l'utilizzo di una libreria Python *open source* per l'elaborazione del linguaggio naturale, ovvero spaCy (Honnibal e Montani 2017, https://spacy.io/).

Un processo assolutamente analogo si sta manifestando anche per i *Large Acoustic Models* (LAM)[2], ossia per sistemi volti a svolgere compiti riguardanti l'uso della lingua parlata: il riconoscimento e la conversione automatica del parlato in testo scritto (*Automatic Speech Recognition*, ASR) e la generazione di una realizzazione fonica a partire da un testo (*Text-to-Speech synthesis*, TTS). La diffusione e il successo di tecnologie del parlato, come gli assistenti vocali comunemente presenti nei dispositivi di telefonia mobile, potrebbero lasciar intendere che quello del riconoscimento e della sintesi siano problemi da poter considerare risolti. A ben vedere ciò non è propriamente vero. Infatti, questi sistemi mirano in prima istanza a 'comprendere' i comandi forniti dagli utenti in domini ristretti ad alcune applicazioni generali (meteo, interrogazione del web eccetera) o a funzioni specifiche del dispositivo (scattare una foto, fare una telefonata, eccetera). La trascrizione automatica del comando vocale impartito può non essere accurata, l'importante è che, in maniera anche vaga, il dispositivo abbia acquisito gli elementi necessari per eseguire il compito richiesto. Diverso, invece, è il caso della dettatura accurata o della trascrizione di un prompt vocale, potenzialmente espresso in un parlato spontaneo, che, eventualmente, attivi un processo gestito da un GLLM. Più in generale, per sistemi ASR e TTS che debbano offrire un alto grado di accuratezza, i materiali utilizzati per l'addestramento e la valutazione non sono dipendenti dal dominio semantico di partenza, che si può evincere dall'argomento trattato nel testo. Inoltre, generalmente è richiesta una quantità di parlato trascritto e non (per tacere poi delle risorse di calcolo necessarie) tale che questi modelli possono essere prodotti solo dalle grandi compagnie private attive nell'ambito degli sviluppi tecnologici. La provenienza dei dati di addestramento raramente è indicata e, per la valutazione delle prestazioni, vengono considerati *benchmark*[3] che non sempre garantiscono l'effettiva generalizzabilità dei risultati. Inoltre, le metriche di valutazione dei modelli acustici sono ancora incentrate sul calcolo di misure standard poco informative come il *Word-Error Rate* (WER) (Morris, Maier, e Green 2004; McCowan et al. 2005; Palmerini e Savy 2014, come vedremo nel paragrafo 3.1).

In questo contesto, approfittando delle opportunità offerte dal progetto *Future Artificial Intelligence Research* (FAIR)[4], all'interno del Piano Nazionale di Ripresa e Resilienza (PNRR)[5], pur senza, al momento, avere accesso diretto ad alcuna forma di finanziamento, un consorzio, formato dall'Università di Napoli Federico II, dal CNR-ISTI di Pisa e dalla Libera

---

[2] *Large Acoustic Model* (LAM) è un termine coniato dagli autori del presente articolo, dopo un controllo dell'uso della stessa denominazione in casi alternativi che non ha portato evidenze. Il richiamo agli LLM è voluto, il termine è stato scelto in alternativa a *Spoken Language Models* (SLM) anch'esso emerso dalle discussioni fra gli autori, ma non selezionato in quanto evocativo di un più complesso sistema di interazione che include aspetti di generazione del linguaggio. In definitiva indicheremo con LAM applicazioni di riconoscimento e sintesi automatica del parlato, mentre in futuro utilizzeremo il termine SLM per riferirci alla concatenazione di un LAM con un sistema interattivo in cui si incontri eventualmente anche una quota generativa, come nel caso di prompt vocali verso un LLM.

[3] Con il termine *benchmark* s'intende una serie di metriche e test standardizzati, il cui fine è fornire una misura delle prestazioni di un modello.

[4] Phonè è un workpackage del *Transversal Project* N.2 (https://fondazione-fair.it/transversal-projects/tp2-vision-language-and-multimodal-challenges/) di FAIR (https://fondazione-fair.it/).

[5] Il Piano di Nazionale di Ripresa e Resilianza (PNRR, https://www.governo.it/sites/governo.it/files/PNRR.pdf) è il documento che ogni Stato membro dell'UE ha dovuto preparare per accedere ai fondi stanziati attraverso il Dispositivo per la ripresa e la resilienza (*Recovery and resilience facility* – RRF) nell'ambito del *Next Generation EU* (NGEU), il programma europeo introdotto per rilanciare l'economia degli stati membri dell'Unione Europea post-pandemia Covid-19 promuovendo investimenti improntati alla sostenibilità e la digitalizzazione.

Università di Bolzano, si pone come obiettivo la raccolta di dati di parlato destinati al pubblico utilizzo da parte di chiunque voglia addestrare (sia in *fine-tuning* che *ex-novo*), testare o estendere l'utilizzo di sistemi di *Automatic Speech Recognition* o di *Text-to-Speech Synthesis* di ultima generazione.

Sebbene il progetto riguardi sia il versante del riconoscimento che quello della sintesi vocale, in questo articolo ci concentreremo sul reperimento e la descrizione della risorsa linguistica destinata allo sviluppo di modelli di riconoscimento automatico del parlato. Questo materiale costituisce una fonte orale per applicazioni di tipo tecnologico, e non solo, di dati raccolti con tecnologie moderne in epoca contemporanea. Dopo una breve descrizione del funzionamento generale dei sistemi di riconoscimento automatico (§2), sarà fornita una panoramica sull'evoluzione nel corso del tempo delle architetture alla base di questi sistemi fino ad arrivare a descrivere quelle che rappresentano lo stato dell'arte (§3) e a discutere delle principali metriche generalmente considerate per la valutazione dei modelli menzionandone le criticità (§3.1). Successivamente verrà presentata l'iniziativa *Phoné*, i suoi obiettivi generali e le attività di raccolta dati e selezione di architetture specifiche per l'addestramento di un modello acustico dell'italiano (§4).

2. SISTEMI DI RICONOSCIMENTO AUTOMATICO DEL PARLATO: DALL'ADDESTRAMENTO ALLA VALUTAZIONE

Il *task* di un sistema di riconoscimento automatico del parlato consiste nel mappare le forme d'onda dei segnali vocali dati in input e restituire in output le stringhe di testo corrispondenti. Più specificamente, un segnale acustico X viene scomposto in una sequenza di osservazioni acustiche [x_1, x_2, ….., x_n] che possono essere sovrapposte o meno, alle quali l'ASR fa corrispondere una sequenza di unità linguistiche (caratteri, fonemi, parole *'subword' units*) Y=[y_1, y_2, ….y_m] che può essere ulteriormente elaborata da un 'modello del linguaggio' che stima la sequenza di parole w_1, …, w_k più probabile con la maggiore probabilità a posteriori P(Y|X) (Malik et al. 2020). Tipicamente, la realizzazione di un sistema di riconoscimento automatico del parlato prevede una fase di raccolta dati, una di addestramento ed infine una fase di valutazione delle prestazioni (Ghai e Singh 2012). La raccolta dati è cruciale per la creazione di un modello con prestazioni adeguate al *task* desiderato e risulta essere determinante per successive fasi. La base dei dati risultante consiste in coppie audio più trascrizione, che consistono tipicamente di molte ore di registrazioni vocali accuratamente trascritte.

In fase di addestramento, avviene la stima dei parametri del modello (sia acustico che del modello del linguaggio) sulla base dei dati raccolti precedentemente ed opportunamente ripuliti e suddivisi in insieme di addestramento e validazione. Dopo la fase di addestramento si misurano le prestazioni dei modelli così ottenuti su un insieme disgiunto di dati, detto 'insieme di valutazione'. Tale insieme ha tipicamente una dimensione adeguata al *task* che consiste in alcune ore di registrazione, tipicamente corrispondenti al 10-15% dei dati raccolti, al fine di produrre stime statisticamente significative (Ghai e Singh 2012). In alcuni casi, al segnale audio possono essere applicate tecniche di riduzione del rumore e/o di 'speech enhancement' al fine di migliorarne la qualità, oltre che una procedura di 'end-point detection' con l'obiettivo di eliminare i segmenti che non contengono segnale utile.

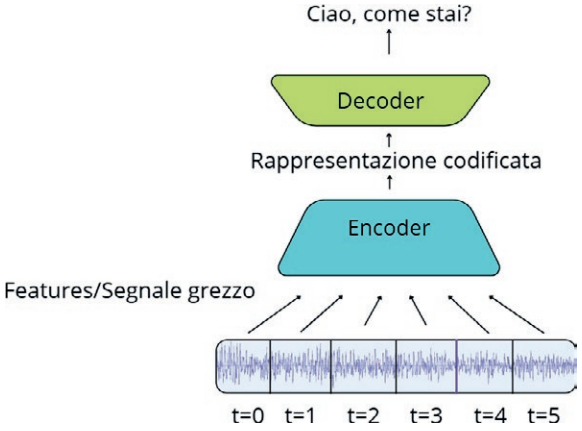

Figura 1. Rappresentazione schematica di un'architettura 'encoder-decoder'.

Per poter trasformare in stringhe di testo le forme d'onda date in input bisogna in primo luogo convertirle in una sequenza di vettori di caratteristiche acustiche. Questo processo coinvolge il *front-end* acustico che si occupa dell'elaborazione del segnale e dell'estrazione delle *feature*.

Il fine di questo processo è quello di elaborare una sequenza di vettori che fornisce una rappresentazione del segnale in input.

Prima di essere inviato al modello acustico (si veda il paragrafo successivo) il segnale di parlato subisce una fase di preelaborazione al fine di estrarre delle feature utili al modello acustico. Tipicamente questa preelaborazione può avvenire in vari modi. Storicamente le caratteristiche acustiche vengono estratte attraverso un processo di analisi spettrale a finestre (frames), ad esempio per generare i Mel Frequency Cepstral Coefficients, MFCCs (Bridle e Brown 1974; Këpuska e Elharati 2015). Più recentemente, alcuni modelli basati su reti neurali addestrano degli estrattori di feature convoluzionali con un processo di addestramento self-supervised (Baevski et al. 2020).

2.1. Gli ASR nell'approccio classico

Nella loro versione classica, o storica se vogliamo, i sistemi di riconoscimento automatico prevedevano due componenti addestrabili separatamente, cioè un modello acustico (AM) e un modello del linguaggio (LM). Il loro addestramento necessitava di due dataset separati con tre differenti forme di codifica. Il primo dataset per l'addestramento del modello acustico contiene sequenze di suoni mappate su di una trascrizione intermedia solitamente diversa da quella testuale (foni, fonemi o altro). Il dataset per l'addestramento del modello linguaggio consente la previsione del prossimo item lessicale.

Il modello acustico (AM) è una componente principale dell'architettura di un sistema di riconoscimento automatico del parlato. Esso, infatti, costituisce la parte predominante del carico computazionale (Ghai e Singh 2012). Nel modello classico l'unità di riferimento è un'unità di coarticolazione (detta anche fono contestuale o trifono) che considera sostanzialmente un suono e le sue transizioni dal suono precedente e verso quello successivo.

Un'altra componente fondamentale di un sistema di riconoscimento automatico del parlato è costituita dal modello del linguaggio (LM). Come abbiamo visto precedentemente questo modello ha il compito di prevedere quale potrebbe essere la parola successiva date quelle che il modello ha precedentemente riconosciuto. In pratica, l'LM è di supporto nella disambiguazione degli elementi generati dal modello acustico considerando il loro cotesto di occorrenza. Generalmente gli ASR classici utilizzano modelli del linguaggio basati sugli N-grammi, addestrati su milioni di parole, che individuano, su base statistica, la sequenza corretta di parole (Ghai e Singh 2012).

2.2. I moderni ASR End-to-End

I moderni sistemi di riconoscimento del parlato fondono le due componenti AM+LM in un unico modello che viene addestrato con un processo detto 'End-to-End' (E2E), che saranno meglio descritti nel paragrafo 3. In questa tipologia di processi viene solitamente utilizzata una sola tipologia di dataset che contiene la corrispondenza tra il segnale acustico e la relativa trascrizione senza alcuna forma di allineamento. Ciò da un lato permette di ridurre il costo relativo ai dati, mentre dall'altro riduce l'interpretabilità delle due singole componenti dato che la codifica intermedia è il risultato di un processo di addestramento complesso. Spesso questi modelli vengono combinati con i cosiddetti 'Large Language Models' che non sono però oggetto del presente lavoro.

3. L'EVOLUZIONE DEGLI ASR

L'idea di creare una macchina in grado di riconoscere e riprodurre il linguaggio naturale ha affascinato gli scienziati per lungo tempo. Considerato che la comunicazione orale costituisce il metodo primario di interazione tra gli esseri umani, sviluppare una macchina in grado di riconoscere e riprodurre parlato risulta essere il modo ottimale per facilitare l'interazione uomo-macchina (Malik et al. 2020). Il primo sistema di riconoscimento funzionante, Audrey, fu creato nel 1952 presso i laboratori Bell. Il sistema possedeva un vocabolario di dieci cifre che lo rendeva in grado di riconoscere vari numeri con un'accuratezza del 97-99%, ma solo se pronunciati da un unico parlante preselezionato (Li e Mills 2019). Altri sistemi simili furono creati negli anni '50, ad esempio presso gli RCA *Laboratories* venne costruito un sistema in grado di riconoscere 10 sillabe per un singolo parlante preselezionato. Al MIT, invece, fu sviluppato un riconoscitore con un vocabolario di 10 vocali indipendente dal parlante (Juang e Rabiner 2004). Questi primi sistemi di riconoscimento automatico del parlato si basavano sulla teoria della fonetica acustica (Juang e Rabiner 2004).

Dagli anni '50 ad oggi, i sistemi ASR sono stati oggetto di diverse innovazioni e la loro architettura interna, così come la natura dei dati impiegati per la loro formazione, sono notevolmente cambiati. Come abbiamo visto in precedenza, i sistemi ASR tradizionali si basano su due componenti separate: il modello acustico (AM) e il modello linguistico (LM). Per la creazione dell'AM erano impiegati gli *Hidden Markov Models* (HMM) e *Gaussian Mixture Models* (GMM) (Karpagavalli e Chandra 2016).

Con l'avvento delle Deep Neural Networks (DNN) (Hinton et al. 2012), in una prima fase, la componente AM ha continuato a essere usata per estrarre informazioni acustiche (ad

es. i MFCCs) tramite metodologie supervisionate, calcolando gli stati e le transizioni degli HMMs attraverso una DNN invece che con una densità di probabilità gaussiana. Un ASR molto usato fondato su questa funzionalità è stato realizzato tramite il framework KALDI (Povey et al. 2011), sviluppato dalla John Hopkins University a partire dal 2010 ed ha consentito notevoli miglioramenti di prestazioni rispetto all'uso dei tradizionali HMMs. Nonostante i costi e l'evoluzione delle tecnologie, le metodologie di etichettatura dei dati audio e l'impiego di modelli del linguaggio basati su n-grammi sono rimaste in voga per alcuni anni ancora grazie alla loro efficacia. Solo nel 2007 è stato introdotto il primo modello ASR (Graves 2012) basato su reti neurali ricorrenti che ottimizza una funzione di loss, *connectionist temporal classification* (CTC), che considera tutti i possibili allineamenti tra osservazioni di input e target.

L'ultima innovazione nell'ambito dei sistemi di riconoscimento automatico del parlato è data dai recenti modelli end-to-end (E2E-ASR) (Li 2022), in particolare dall'introduzione dell'architettura di rete Transformer (Vaswani et al. 2017). A differenza dei sistemi precedenti, gli ASR end-to-end non necessitano di alcuna rappresentazione intermedia nella conversione da segnale vocale a stringa di testo corrispondente, tutto ciò avviene in maniera diretta (come suggerisce il nome stesso di questi sistemi). Ciò vuol dire che tali sistemi presentano un solo obiettivo: eseguire la mappatura diretta del segnale vocale in una sequenza di caratteri (questi sistemi non usano informazione lessicale in prima istanza); per essere addestrati necessitano del solo set di dati acustici correlati dalla loro trascrizione e non fanno uso dell'LM.

L'architettura *Transformer* permette la creazione di sistemi composti da un *encoder* e un *decoder*, in grado di mappare direttamente una sequenza di suoni non allineati alla trascrizione corrispondente. Questi sistemi, addestrati con poche centinaia di ore di parlato trascritto, non allineato, mediante tecniche di apprendimento supervisionato, risultano avere prestazioni migliori rispetto alla generazione di modelli precedentemente sviluppata, restituendo una percentuale di errore pari al 5%[6] nel caso di sistemi *Transformer-Encoder*, o al 4% nel caso di sistemi *Conformer-Encoder*[7] (Gulati et al. 2020). Bisogna inoltre specificare che un ulteriore fattore che influisce notevolmente sulle prestazioni è costituito dalla scelta di implementazione di un modulo aggiuntivo che associa all'audio la sequenza di caratteri più probabile, ricorrendo ancora una volta a proprietà distributive della sequenza stessa. Il modulo è implementato come modello di *Connectionist Temporal Classification* (CTC) (Graves et al. 2006). La CTC è una tecnica di trascrizione del segnale vocale che raggruppa etichette di trascrizione consecutive, tutte uguali (carattere, pezzo di parola, ecc.) su un'unica etichetta: tale procedimento è illustrato nella figura 2. Ciò si rivela molto utile in situazioni comunicative in cui un parlante prolunga involontariamente la vocale finale di una parola, utilizzandola come pausa riflessiva o come riempitivo nel flusso del discorso. In questo caso la CTC elimina le vocali in eccesso trascrivendo la parola senza prolungamenti (Graves 2012).

L'approccio qui descritto, pur essendo basato su sistemi *Deep Neural Network end-to-end* quindi 'relativamente' moderno, genera una famiglia di sistemi che ancora richiedono ingenti quantità di dati corredati dalla loro trascrizione. Questi dati devono di norma essere prodot-

---

[6] Il *benchmark* è stato effettuato utilizzando LibriSpeech, un corpus che comprende 1000 ore di parlato letto in lingua inglese.

[7] Quest'ultimo sistema differisce dal *Transformer* in molti modi, ma principalmente per l'impiego di blocchi convoluzionali e per l'utilizzo di uno stile architetturale detto *Macaron-Style* (Kuchaiev et al. 2019).

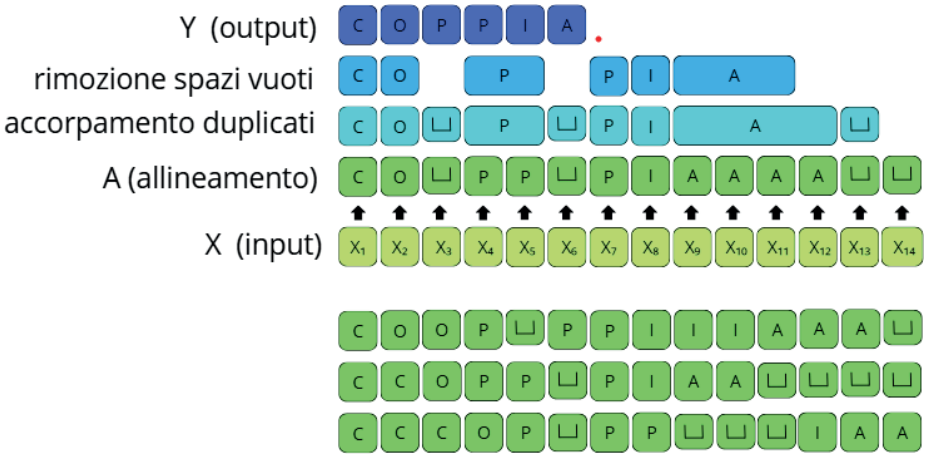

Figura 2. un'illustrazione per mostrare, in maniera semplificata, il meccanismo di funzionamento della CTC.

ti manualmente, o con altri sistemi automatici, e controllati *a posteriori* da operatori umani. In questo senso si può affermare che questi modelli appartengono alla famiglia dei sistemi basati sì su intelligenza artificiale, ma ancora legati alla necessità di una supervisione umana (*human in the loop*, HITL).

L'alternativa più moderna e per la quale i costi si spostano (ma non si riducono) dal ruolo del lavoro manuale verso investimenti in enormi quantità di dati e di risorse di calcolo, consiste nell'impiego di tecniche di apprendimento *self-supervised.*

Questi sistemi elaborano informazione acustica restando in 'ascolto' non mediato di migliaia di ore di parlato senza curarsi della trascrizione. Da questo pseudo-ascolto, utilizzando processi complessi ma totalmente autonomi, emerge un repertorio di categorie di suoni e delle loro proprietà distributive, in analogia a quanto farebbe un neonato esposto senza altri stimoli, al solo ascolto della lingua materna (Giordano Orsini, Vitale, e Cutugno 2023). Di fatto questi sistemi sono, finalmente, i primi veri ASR *end-to-end*, che 'apprendono' le strutture acustiche di una lingua senza ricorrere ad alcun apporto umano per tutta la fase di costruzione delle proprietà fonologiche[8] e fonotattiche della lingua parlata. Principali esponenti di questa famiglia di sistemi sono Wav2Vec2 (Baevski et al. 2020), HuBERT (Hsu et al. 2021) e Whisper (Radford et al. 2023). Scendendo un po' più in profondità nella descrizione del loro processo di funzionamento, possiamo osservare che questo comporta due fasi principali. La prima è la fase di *pre-training*, durante la quale, come detto in precedenza, vengono impiegate grandi quantità di dati vocali (nell'ordine delle decine di migliaia di ore, in alcuni casi di più lingue, in altri di una sola) senza fornire alcuna informazione aggiuntiva, cioè senza tra-

[8] Ovviamente, in questo caso il concetto di proprietà fonologiche deve essere considerato con una certa cautela. Questi sistemi estraggono proprietà fondanti della lingua parlata che possono essere impiegate negli stadi successivi del sistema per imparare a trascrivere, ma non ci è dato di comprendere come il repertorio "fonologico" è stato costruito e da quanti e quali elementi sia composto. Tuttavia, non c'è dubbio che i processi di astrazione operati dalla rete neurale si muovano esattamente nella direzione della creazione di un repertorio di tratti e di unità costruito dal basso ma che costituisce, in ogni caso, una base formale per la descrizione della lingua in esame. Questo repertorio rappresenta dunque un sistema complesso ma decidibile, quindi utile, come minimo, ma non solo, allo scopo della trascrizione.

scrizione. In questa fase si riconoscono e discretizzano le rappresentazioni di unità acustiche nascoste (che vanno a sostituirsi alle MFCC) utilizzando diversi processi, come la quantizzazione, direttamente dal campione audio grezzo o il *clustering*. In una seconda fase, il prodotto della prima fase di analisi del parlato non trascritto viene 'congelato' e produce l'input per un modulo altrettanto complesso e raffinato al quale bastano poche (dell'ordine di un centinaio) ore di parlato trascritto per addestrare il sistema a produrre la migliore sequenza di caratteri corrispondente ad un dato segnale acustico. Anche in questo caso, la catena di produzione è completata da un modulo CTC per perfezionare la scelta dell'output migliore. Esistono delle differenze sostanziali tra Wav2Vec e HuBERT: il primo, non opera alcuna preelaborazione del segnale vocale, analizza direttamente i campioni del segnale digitale e opera i suoi processi di quantizzazione e *clustering* in maniera totalmente autonoma. Il processo che ne deriva è complesso e richiede consistenti risorse di calcolo. HuBERT, al contrario, non lavora sui campioni del segnale vocale ma opera un'estrazione automatica dei parametri MFCC e usa quelli per il processo di quantizzazione e *clustering*. Questa soluzione semplifica la complessità della fase di *pre-training*.

Infine, Whisper si definisce come un sistema *multi-task* basato su *weak supervision* e *transfer learning*. Da un lato, per l'addestramento, invece di utilizzare materiale non trascritto, utilizza parlato che, pur essendo in principio non trascritto, viene automaticamente trasformato in testo ricorrendo a sistemi basati su tecnologie precedenti. Anche se queste trascrizioni presentano a volte significative imprecisioni, gli autori di Whisper dimostrano che le prestazioni dei sistemi addestrati con questi tipi di dati possono eguagliare lo stato dell'arte se non addirittura superarlo. Dall'altro lato, trattandosi di un modello *multi-task*, si sfruttano dati utilizzati per altre attività basate sul parlato, combinando la capacità di modellazione incrociata nativa del modello e sfruttandola per il *task* di trascrizione. La versione base di Whisper, rilasciata liberamente da OpenAI, è addestrata con circa 680.000 di ore di parlato così trattato, di cui 120.000 di lingue diverse dall'inglese a cui si aggiungono altre 120.000 ore di parlato inglese di cui è fornita la traduzione in altre lingue come trascrizione aggiuntiva.

Per tutti e tre i sistemi sopra indicati, i costi per produrre *in house* moduli di ASR sono difficilmente sostenibili da enti di ricerca, per non parlare della difficoltà di reperire tante ore di parlato per l'addestramento, anche se non trascritto. Tuttavia, ora che le politiche *open-source* vengono applicate anche a sistemi originariamente proprietari, l'azienda che ha prodotto Wav2Vec distribuisce in maniera gratuita alcune versioni del prodotto 'pronte all'uso' con vari processi di *pre-training* basati su differenti *data-set*. È poi, inoltre, possibile, con processi relativamente meno onerosi, applicare ulteriori fasi di *fine-tuning* per raffinare le prestazioni del sistema su lingue e/o su domini lessicali specifici. Va comunque specificato che per tutti questi sistemi il codice per sviluppare privatamente applicazioni, a patto di avere le risorse richieste, è in media liberamente disponibile in licenze aperte.

3.1. Metriche di valutazione

Una delle principali metriche di valutazione dell'accuratezza di un sistema di riconoscimento automatico del parlato è il *Word Error Rate* (WER). Il WER si basa sull'analisi della discrepanza tra la trascrizione manuale usata come riferimento e la stringa di testo generata dal riconoscitore. Una volta avvenuto il confronto, vengono rilevate le parole inserite per

errore (*Insertions*), le parole riconosciute in maniera errata (*Substitutions*) e quelle mancanti (*Deletions*). La formula per calcolare il *Word Error Rate* è la seguente:

$$\text{WER} = 100 \times (\textit{Insertions} + \textit{Substitutions} + \textit{Deletions}) \,/\text{Total number of words}$$

Un'altra metrica di valutazione delle prestazioni di un ASR è costituita dal *Character error rate* (CER). Il CER calcola la percentuale di caratteri errati (nel calcolo vengono considerati anche gli spazi) mediante la formula seguente:

$$\text{CER} = 100 \times (\text{Number of incorrect characters / Total number of characters})$$

Le metriche utilizzate per valutare questi strumenti si basano su calcoli che presentano limitazioni significative. Infatti, le metriche tradizionali, quali WER e CER, sebbene siano descritte come misure oggettive, costituiscono stime generiche prive di qualsiasi tipo di valutazione qualitativa che, al contrario, favorirebbe il miglioramento dei sistemi stessi e della loro applicabilità rispetto a specifici domini di utilizzo. Ad esempio, conoscere la natura degli errori agevolerebbe interventi risolutivi; inoltre nella trascrizione di un documento l'importanza del riconoscimento di ciascuna parola può variare a seconda del relativo carico di informazioni nel dominio specifico di riferimento.

Riconosciuti i limiti delle metriche standard tradizionali, alcuni studi hanno proposto la considerazione della quantità di informazione veicolata per ottenere misure più significative. Ad esempio, Morris e colleghi (2004) propongono metriche per valutare il riconoscimento del parlato connesso basate sulla percentuale di corrispondenze errate tra le parole in input e in output e un'approssimazione alla percentuale di informazione persa per parola. McCowan e colleghi (2005) poi sottolineano che la valutazione di sistemi di riconoscimento del parlato dovrebbe permettere un'analisi delle prestazioni approfondita e interpretabile rispetto al fine applicativo del sistema. Quindi dovrebbe essere non solo diretta e oggettiva, ovvero calcolabile in maniera automatica per favorire l'applicabilità su vasta scala e la comparabilità scientifica, ma anche interpretabile e modulare: il valore della misura dovrebbe fornire un'indicazione chiara dell'affidabilità del sistema e permettere di attribuire agli errori un peso diverso a seconda della specifica applicazione considerata. Gli autori propongono che tali requisiti potrebbero essere soddisfatti considerando misure legate al recupero di informazioni (*Information Retrieval*) che tengano conto quindi della quantità di informazioni inserite ma non realmente presenti (falsi positivi) e delle informazioni presenti ma non individuate (falsi negativi).

In ambito italiano, Palmerini e Savy (2014) evidenziano e affrontano la questione dei limiti delle metriche standard conducendo uno studio in cui si osserva che l'impiego di metriche meno ingenue dal punto di vista linguistico, più attente al tipo di errore che alla mera presenza di discrepanze tra quanto riconosciuto dal sistema e il dato di controllo (ovvero la trascrizione manuale allineata), permette non solo di identificare la frequenza degli errori, ma anche la loro natura.

Lo studio propone un nuovo paradigma di metadatazione. Pur mantenendo il criterio di individuazione di cancellazioni, inserzioni e sostituzioni, introduce ulteriori annotazioni di matrice strettamente linguistica che considerano parametri sui diversi piani di analisi lessicale, morfologico e fonetico-fonologico. Gli elementi facenti parte di questi domini possono di

volta in volta assumere valori in base ad alcune caratteristiche come ad esempio la complessità morfologica, la lunghezza fonologica e sillabica, i casi di coppie minime e così via. L'applicazione di questo sistema alla valutazione del sistema di riconoscimento considerato nello studio ha dimostrato una correlazione tra percentuale di errore e complessità morfologica, sillabica e fonologica: maggiore è la frequenza di parole lunghe e complesse, minore è la percentuale di errore. Anche per quanto riguarda la distribuzione degli errori nelle diverse categorie grammaticali ci sono risultati interessanti: in particolare lo studio mostra come le categorie maggiormente suscettibili di errore siano le parole funzionali e l'occorrenza di fenomeni di disfluenza. Pertanto, lo studio dimostra l'importanza di definire procedure e metriche di valutazione più specifiche che entrino nel merito dell'errore al fine di avere una comprensione più dettagliata e approfondita delle prestazioni dei sistemi ASR. Successivamente, sempre in ambito italiano, un altro studio affronta il problema dei limiti delle metriche tradizionali presentando un'analisi linguistica condotta sulle trascrizioni generate da ASR allo stato dell'arte, concentrandosi sulla considerazione delle parti del discorso maggiormente suscettibili di errore (Vitale, Tanda, e Cutugno 2024). In particolare, sono state prese come riferimento le trascrizioni manuali, sottoposte a *Part-of-speech tagging* (Pos), per confrontarle con le trascrizioni generate dagli ASR. L'analisi preliminare (in corso di sviluppo) evidenzia che in questo caso i problemi di riconoscimento riguardano perlopiù unità avverbiali e intende individuare strategie per il miglioramento dei sistemi, aumentandone l'efficacia e l'affidabilità.

4. PHONÉ

Il consorzio *Phoné* (dal greco *φωνή* 'suono linguistico', 'voce'), formato dall'Università di Napoli Federico II, dal CNR-ISTI di Pisa, e dalla Libera Università di Bolzano, nasce come un'iniziativa volontaria che si pone l'obiettivo generale di raccogliere dati di parlato italiano destinati al pubblico utilizzo da parte di chiunque voglia addestrare (sia in *fine-tuning* che *ex-novo*), testare o estendere l'utilizzo di sistemi di *Automatic Speech Recognition* di ultima generazione.

Come abbiamo già osservato, allo stato attuale può risultare difficile, se non impossibile, per enti pubblici reperire le risorse linguistiche e computazionali necessarie per l'addestramento da zero di un ASR. Più facilmente si ricorre alla valutazione e/o all'adattamento di sistemi pre-addestrati messi a disposizione fra quelli a cui abbiamo fatto riferimento in precedenza. Prima di mettere in esercizio questi sistemi, tuttavia, in primo luogo occorre valutarne le prestazioni in relazione alle caratteristiche generali dei sistemi testandoli su un dataset di riferimento. Occorre quindi un insieme di dati controllato e certificato rispetto alle principali caratteristiche (stile di parlato, bilanciamento rispetto al sesso dei parlanti, condizioni di rapporto segnale-rumore), che sia utilizzabile come riferimento standardizzato per potere obiettivamente confrontare la qualità dei sistemi in oggetto. Successivamente gli utenti con necessità applicative specifiche dovranno individualmente dotarsi di dati relativi al dominio specifico di interesse per lessico e stili di parlato. Fra gli obiettivi di *Phoné* c'è quello di mettere a disposizione della comunità scientifica una serie di strumenti molto articolata:

1) un corpus da usare per la valutazione dei sistemi rispetto alle loro caratteristiche generali;
2) corpora in lingua italiana per l'addestramento di nuovi sistemi sia da zero che per un eventuale *fine-tuning*, articolati in dati non trascritti (per i sistemi *self-supervised*) e in dati

trascritti da usare sia per l'ultimo stadio dei sistemi non supervisionati che per addestrare i sistemi supervisionati; a tale scopo è necessario raccogliere ingenti quantità di parlato non trascritto (circa 10.000 ore, equivalenti a poco più di un anno, potrebbero essere sufficienti), da utilizzare seguendo un approccio di autoapprendimento tipico di questi sistemi; al parlato non trascritto si affianca poi una quantità decisamente minore di parlato trascritto accuratamente che viene solitamente utilizzato per completare l'addestramento nella sua quota supervisionata o per raffinare le prestazioni su domini speciali.

3) due sistemi ASR, uno basato su un sistema non supervisionato e uno basato su ASR E2E supervisionato, entrambi addestrati da zero con dati in italiano, raccolti e controllati dal consorzio;
4) una infrastruttura per la sintesi vocale basata su tecniche E2E, proposta sia come applicazione pronta all'uso e basata sulla voce di due *speaker* italiani, uno maschile e uno femminile, sia come codice pronto all'uso e reso utilizzabile anche da persone non esperte, per costruire in proprio nuove voci;
5) metriche e procedure per la valutazione dei modelli/architetture/processi allo stato dell'arte in contesti specifici.

In questo lavoro, specificamente dedicato alle risorse orali parleremo del punto 2) e faremo cenno alle scelte progettuali per giungere alla descrizione di quanto indicato nel punto 3). Il dataset descritto nel punto 1), ovviamente è un sottoinsieme di quello che si produce nel lavoro relativo alla creazione del 2).

4.1. Fonti per la creazione del dataset di riferimento

Il dataset *Phoné* è stato creato raccogliendo materiale audio, in lingua italiana, pubblico ed eterogeneo, sia dal punto di vista diatopico che diamesico, che include diverse situazioni comunicative, tra cui dialoghi, monologhi, letto e altro. Il materiale è suddiviso in trascritto (la trascrizione è di tipo ortografica) e non trascritto.

Con riferimento ai primi due punti sopra indicati, al momento le fonti principali sono tre: la biblioteca digitale *LibriVox*, il corpus CLIPS e i video presenti su canali *Youtube* con licenza *Creative Commons*. Un'ulteriore fonte sarebbe costituita dai *TED Talks*, vale a dire discorsi tenuti nell'ambito di conferenze, gestite dal ramo italiano dall'organizzazione privata non-profit statunitense *Sapling Foundation*, note anche come appunto *TED* (*Technology Entertainment Design*) *talks*, inizialmente incentrate su tematiche legate al mondo della tecnologia e del design per poi includere argomenti di domini anche molto diversi. Tali registrazioni sono state usate in letteratura per supportare lo sviluppo di sistemi di traduzione automatica (Cettolo, Girardi, e Federico 2012). Tuttavia, sebbene anche questi video siano rilasciati con una licenza *Creative Commons*, il loro utilizzo è soggetto a specifiche restrizioni per le quali è necessario fare richiesta esplicita di autorizzazione (procedura in corso al momento della stesura dell'articolo)[9].

Infine, considerato l'impegno richiesto dalla raccolta della quantità di dati necessaria allo sviluppo e valutazione di sistemi di riconoscimento automatico del parlato, l'iniziativa si apre al coinvolgimento delle comunità scientifiche attive nello studio della lingua parlata.

---

[9] Per ulteriori informazioni si rimanda al sito https://www.ted.com/talks.

4.1.1. LibriVox

LibriVox è una biblioteca digitale online che offre audiolibri gratuiti di opere nel pubblico dominio. Fondata nel 2005 da Hugh McGuire con l'obiettivo di rendere accessibili a tutti i libri di dominio pubblico in formato audio, *LibriVox* conta su lettori volontari provenienti da diverse parti del mondo che registrano le loro letture ad alta voce per creare gli audiolibri.

Sebbene l'87% della collezione sia in lingua inglese, *LibriVox* produce anche audiolibri in altre 36 lingue, compreso l'italiano.

Librivox fornisce unicamente audiolibri di testi appartenenti al pubblico dominio, in particolare, la maggior parte dei libri di riferimento proviene da *Project Gutenberg*, una vasta biblioteca digitale libera che mette a disposizioni testi non coperti da diritto d'autore o copyright e che quindi possono essere utilizzati e distribuiti gratuitamente. *LibriVox* offre a chiunque voglia la possibilità di donare la propria voce per registrare questi testi, creando audiolibri che vengono poi nuovamente rilasciati nel pubblico dominio. Ciò permette a chiunque di utilizzare gli audiolibri per qualsiasi scopo, eliminando le barriere legate al diritto d'autore. Per ulteriori informazioni si rimanda alla pagina https://librivox.org/.

4.1.2. CLIPS

CLIPS (Corpora e Lessici di Italiano Parlato e Scritto, Savy e Cutugno 2009) è un corpus di italiano parlato interamente pubblico e pertanto disponibile liberamente per l'uso e la distribuzione. CLIPS fa parte di un progetto finanziato da MURST e MIUR, partito il 5 febbraio 1999 e concluso nel 2004. Il progetto era finalizzato alla messa a punto di strumenti per lo studio generale e per il trattamento automatico dell'italiano. Il corpus CLIPS è caratterizzato da una duplice stratificazione, diatopica e diafasica. Per quanto riguarda la varietà diafasica, CLIPS è suddiviso in cinque sottocorpora: Dialogico, Letto, Ortofonico, Telefonico (in questa prima fase nel dataset *Phoné* il sottocorpus telefonico non è stato preso in considerazione) e Radiotelevisivo. Per quanto riguarda la varietà diatopica, invece, il corpus comprende materiale raccolto in varie città italiane scelte in modo tale da essere rappresentative dal punto di vista della varietà di italiano, ovvero: Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia. Per ulteriori informazioni si rimanda al sito http://www.clips.unina.it/home.

4.2. Le architetture per il riconoscimento

4.2.1. Valutazione dei sistemi pre-addestrati attualmente disponibili per l'italiano

Prima di procedere alla determinazione dei sistemi che *Phoné* configurerà e addestrerà con dati solo italiani, si è deciso di condurre una prima valutazione, per ora estremamente preliminare, delle prestazioni di alcuni sistemi liberamente disponibili in rete e pre-addestrati a riconoscere l'italiano.

In questa fase iniziale, è stato utilizzato il corpus CLIPS, composto da 576, 8 minuti di parlato dialogico, letto, ortofonico e radiotelevisivo prodotto da 186 parlanti in diverse varietà regionali (si veda §4.1.2). Il corpus è stato segmentato in campioni di parlato consistenti in unità inter-

pausali e suddiviso in tre parti seguendo una proporzione del 60%, 20% e 20% le quali sono state utilizzate rispettivamente per le fasi di addestramento, sviluppo e valutazione dei modelli selezionati. Per il testing è stato utilizzato il 20% del subset, equivalente a 2511 campioni.

I sistemi che abbiamo valutato in questa fase preliminare sono:

a) *Wav2Vec2.0* è un ASR *end-to-end self-supervised* (Baevski et al. 2020) rilasciato da *Meta Platforms Inc.*, come abbiamo visto in precedenza, viene anche definito *Large Acoustic Model* a causa del suo processo di formazione che di solito comporta due fasi principali. Durante la prima, detta di *pre-training*, vengono impiegati grandi quantità di dati vocali non trascritti per raggruppare e discretizzare le rappresentazioni di unità acustiche nascoste utilizzando un processo di quantizzazione, direttamente da campioni di segnale, senza l'estrazione delle MFCC. Durante la seconda ed ultima fase avviene l'addestramento dell'ultimo strato finalizzato alla trascrizione, per il quale vengono utilizzati insiemi di dati trascritti di dimensione molto minore rispetto alla fase precedente (al massimo centinaia di ore o pochi minuti). L'obiettivo di questa seconda fase è di apprendere la corrispondenza tra gli elementi del repertorio appreso nella prima fase, e le corrispettive realizzazioni associate infine alle trascrizioni. Nello specifico, la versione utilizzata per la valutazione è *Wav2Vec2-Large-XLSR-53*.
b) *Conformer* (Gulati et al. 2020) è un'architettura basata su Transformer, che differisce da questo per la costruzione dei blocchi di codifica. NVIDIA ha pre-addestrato e rilasciato, tramite il suo *Framework* Nemo, tre varianti pre-addestrate del *Conformer*: *Conformer-CTC, Conformer Transducer e Fast-Conformer.* Il *Conformer-CTC* è un modello non auto-regressivo, che utilizza come funzione obiettivo (*loss)* e di decodifica la CTC (*Connectionist Temporal Classification).* Il Conformer-Transducer è un modello auto-regressivo che utilizza un decoder RNNT/*Transducer,* cioè un automa finito che tiene conto delle osservazioni precedenti (se presenti). Il *FastConformer* (Rekesh et al. 2023) consiste in una versione più leggera e veloce dei primi due, grazie ad alcuni accorgimenti architetturali. In particolare, le versioni dei modelli che abbiamo valutato sono: *Conformer CTC Large*, *Conformer Transducer Large* e *Fast Conformer-Hybrid Large*.

In relazione al punto 3) sopra indicato, occorreva individuare dei modelli disponibili liberamente rilasciati anche in forma di codice sorgente e che fosse possibile addestrare da zero con i nostri dati.

4.2.2. Le architetture da addestrare *from scratch*

Per quel che riguarda l'addestramento da zero, sono state selezionate due architetture, una per tipologia di addestramento (*supervised, self-supervised).*

Tabella 1. WER e CER dei modelli pre-addestrato valutati con il dataset selezionato.

Modello Pre-addestrato	WER (%)	CER (%)
Conformer CTC Large	30.06	13.40
Conformer Transducer Large	27.89	14.15
Fast Conformer- Hybrid Large	28.25	13.66
Wav2Vec2-Large-XLSR-53	34.5	13.94

a) *HuBERT* (Hsu et al. 2021) è un modello *self-supervised* basato sulla stessa architettura di *Wav2Vec2.0*, l'unica differenza risiede nel modo in cui si svolge la prima fase di addestramento, per la quale invece di un processo di quantizzazione ripetuto durante ogni epoca di addestramento, i campioni di parlato vengono raggruppati una volta sola tramite un algoritmo di *clustering* per il quale vengono utilizzate le feature MFCC. Anche in questo caso il repertorio ha una dimensione fissata inizialmente a cento e poi modificata a mano a mano che le fasi di pre-addestramento avanzano. Per quel che riguarda la seconda fase, quella di addestramento dello strato di trascrizione, i passi sono gli stessi.
b) Per la parte supervisionata abbiamo scelto l'architettura *FastConformer* che attualmente è la base per altre tipologie di modelli *multi-task,* cioè, addestrati a svolgere compiti diversi, e multilingua. Basato sull'architettura *Conformer*, introduce alcune ottimizzazioni che lo rendono più efficiente sia in fase di addestramento che in fase di inferenza. In particolare, l'*encoder* acustico risultante, combinato con un modello del linguaggio basato su *Transformer*, risulta essere nettamente migliore rispetto a tutti gli altri modelli supervisionati considerati.

4.3. Questioni legali ed etiche

Un punto cruciale nella gestione dei dati è rappresentato dagli aspetti relativi alla privacy e dai criteri per la redistribuzione degli stessi, al fine di garantirne e promuoverne un uso responsabile e sicuro. Il materiale per scopi interni non sarà distribuibile. Sarà invece possibile la distribuzione esterna di una parte del materiale raccolto, che auspicabilmente sarà massimizzata per un proficuo uso nelle mani di tutta la nostra comunità scientifica di riferimento, nei seguenti casi:

– dati per cui il consorzio *Phoné* ha ricevuto un'autorizzazione esplicita da parte dei proprietari dei dati donati alla iniziativa;
– dati con licenze pubbliche (ad esempio quelle *Creative Commons*) che il consorzio ha reperito da fonti che a monte avevano reso note le condizioni per le quali la proprietà intellettuale era riconosciuta e protetta;

I materiali di cui si dispone dell'autorizzazione alla distribuzione, ma che comunque potrebbero violare diritti di autore o di privacy, saranno sottoposti – prima della distribuzione – a processi quali randomizzazione (frammentando i file originali e ricomponendoli casualmente, in modo da non consentire la loro ricostruzione) e anonimizzazione al fine di assicurare la tutela di informazioni sensibili.

Al momento attuale abbiamo ricevuto come donazione materiali da diverse fonti e fornitori. Le condizioni legali relative alla proprietà di questi materiali variano a seconda della provenienza, e per ognuna di queste condizioni sono stati stipulati degli accordi specifici che liberano *Phoné* da rischi legati alla violazione della privacy o ai diritti di riproduzione dei parlanti. In alcuni casi, è il fornitore che garantisce pienamente la redistribuibilità dei dati, mentre in altri il materiale era stato già acquisito all'origine con licenze di pubblico dominio. Il fine del progetto è di creare un dataset pubblico destinato a chiunque voglia fare ricerca e a democratizzare l'accesso a tali risorse. Qualora un'azienda, o un privato in generale, intenda utilizzare parte del materiale contenuto nel dataset per fini che mal si accompagnano alle intenzioni che hanno portato alla creazione del progetto, quindi scopi commerciali ecc., questi dovrà fare

riferimento e dunque chiedere la licenza al proprietario e donatore dei dati, piuttosto che a *Phoné* stesso. A quel punto, sarà il fornitore a decidere liberamente in che modo gestire l'uso del materiale. Eventualmente, il consorzio potrà coordinare il rapporto tra le parti.

5. CONCLUSIONI

In questo articolo, ci siamo concentrati sulla descrizione del funzionamento e dell'architettura di sistemi di riconoscimento automatico del parlato (Malik et al. 2020), sulla loro evoluzione nel corso del tempo (Karpagavalli e Chandra 2016), fino ad arrivare ai più recenti modelli *End-to-End* (Li 2022) e all'introduzione dell'architettura di rete *Transformer* (Vaswani et al. 2017). Ciò ha evidenziato, da un lato, come la rapida evoluzione delle tecnologie legate all'elaborazione delle lingue nella forma scritta e parlata, basate quindi su *Large Language Models* (LLM) e i *Large Acoustic Models* (LAM), stia trasformando radicalmente l'elaborazione automatica delle lingue naturali e, di conseguenza, il nostro modo di interagire con il mondo digitale; d'altro canto, si è osservato che l'accesso limitato allo sviluppo e valutazione di questi modelli avanzati rappresenta una sfida significativa per la comunità accademica e per gli sviluppatori interessati a esplorare e utilizzare tali risorse. A tal proposito, è stato presentato il progetto *Phoné*, il quale si propone di colmare questa lacuna raccogliendo un dataset italiano per il riconoscimento automatico del parlato accessibile a tutti coloro che desiderano addestrare, testare o estendere l'uso di sistemi di *Automatic Speech Recognition* (ASR) di ultima generazione. Il progetto nasce dall'iniziativa collaborativa fra nuclei afferenti a tre enti di ricerca pubblica, l'Università di Napoli Federico II, il CNR-ISTI di Pisa e la Libera Università di Bolzano, che resta aperta anzi invita al coinvolgimento della comunità scientifica impegnata nella raccolta e studio di materiale parlato considerando che la condivisione di risorse possa essere determinante per affrontare l'impresa prefissata.

La valutazione condotta considerando modelli di riconoscimento *Conformer CTC Large*, *Conformer Transducer Large*, *Fast Conformer-Hybrid Large* e *Wav2Vec2-Large-XLSR-53* e utilizzando una parte del dataset *Phoné* ha evidenziato alcune delle sfide e delle opportunità nel campo del riconoscimento automatico del parlato italiano, tra cui i limiti delle metriche di valutazione tradizionali. Le attività in corso stanno riguardando l'ampliamento del dataset e la predisposizione di architetture per lo sviluppo di modelli di riconoscimento specifici per l'italiano. L'analisi degli errori riscontrati con le metriche di valutazione tradizionali sarà volta all'elaborazione di metriche più informative rispetto ai tipi di errore osservabili, che siano quindi indicative degli interventi necessari per il miglioramento del sistema in valutazione (Palmerini e Savy 2014). In un secondo momento si passerà al lavoro riguardante sistemi di sintesi da testo allo stato dell'arte.

Lo sviluppo di modelli specifici per l'italiano offrirebbe risorse applicabili nell'ambito della ricerca linguistica, ad esempio fornendo formalizzazioni computazionali utili per la valutazione sperimentale di teorie linguistiche o fornendo strumenti per la trascrizione automatica dotati di un buon grado di precisione.

Per concludere, il lavoro svolto su *Phoné* rappresenta un passo significativo verso la democratizzazione delle tecnologie linguistiche avanzate promuovendo l'accesso equo e l'innovazione nel campo del riconoscimento automatico del parlato e della generazione di produzioni sintetiche a partire da testi scritti.

BIBLIOGRAFIA

Baevski, Alexei, Zhou Yuhao, Mohamed Abdelrahman, e Michael Auli. 2020. "Wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations". *Advances in Neural Information Processing Systems,* 33: 12449-60.

Bridle, John Scott, e Michael D. Brown. 1974. "An Experimental Automatic Word-Recognition System". *Joint Speech Report,* 1003.5: 33.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, et al. 2024. "A Survey on Evaluation of Large Language Models". *ACM Transactions on Intelligent Systems and Technology,* 15(3): 1-45. https://doi.org/10.1145/3641289.

Ghai, Wiqas, e Navdeep Singh. 2012. "Literature Review on Automatic Speech Recognition". *International Journal of Computer Applications,* 41(8): 42-50. https://doi.org/10.5120/5565-7646.

Giordano Orsini, Luigi Maria, Vincenzo Norman Vitale, e Francesco Cutugno. 2023. "Large Scale Acoustic Models: A New Perspective". *Sistemi Intelligenti,* 35(2): 401-12. https://doi.org/10.1422/108137.

Goodfellow, Ian, Yoshua Bengio, e Aaron Courville. 2016. *Deep Learning*. MIT Press.

Graves, Alex, Santiago Fernández, Faustino Gomez, e Jürgen Schmidhuber. 2006. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". *Proceedings of the 23rd International Conference on Machine Learning:* 369-76.

Graves, Alex. 2012. "Sequence Transduction with Recurrent Neural Networks". *arXiv preprint*. https://doi.org/10.48550/arXiv.1211.3711.

Graves, Alex 2014. "Generating Sequences with Recurrent Neural Networks". *arXiv preprint*. https://doi.org/10.48550/arXiv.1308.0850.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, e Ruoming Pang. 2020. "Conformer: Convolution-augmented Transformer for Speech Recognition". *Proceedings of INTERSPEECH 2020, October 25-29, 2020, Shanghai, China*: 5036-5040. https://dx.doi.org/10.21437/Interspeech.2020-3015.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, e Brian Kingsbury. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". *IEEE Signal Processing Magazine*, 29(6): 82-97. https://doi.org/10.1109/MSP.2012.2205597.

Honnibal, Matthew, e Ines Montani. 2017. "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing". https://spacy.io/.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, e Abdelrahman Mohamed. 2021. "Hubert: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units". *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 29: 3451-60. https://doi.org/10.1109/TASLP.2021.3122291.

Ježek, Elisabetta, e Rachele Sprugnoli. 2023. *Linguistica computazionale. Introduzione all'analisi automatica dei testi*. Bologna: il Mulino.

Juang, Biing-Hwang, e Lawrence R. Rabiner. "Automatic speech recognition–a brief history of the technology development". *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1.67(2005): 1.

Jurafsky, Daniel, e James H. Martin. 2009. *Speech and Language Processing*. Pearson Education, Inc.

Karpagavalli, Shunmugam, e Erick Chandra. 2016. "A Review on Automatic Speech Recognition Architecture and Approaches". *International Journal of Signal Processing, Image Processing and Pattern Recognition,* 9: 393-404. https://doi.org/10.14257/ijsip.2016.9.4.34.

Këpuska, Veton Z., e Hussien A. Elharati. 2015. "Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RATA-PLP and Hidden Markov Mod-

el Classifier in Noisy Conditions". *Journal of Computer and Communications,* 3: 1-9. https://doi.org/10.4236/jcc.2015.36001.

Kuchaiev, Oleksii, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman et al. 2019. "Nemo: A Toolkit for Building AI Applications Using Neural Modules". *arXiv preprint.* https://arxiv.org/abs/1909.09577.

Li, Jinyu. 2022. "Recent Advances in End-to-end Automatic Speech Recognition". *APSIPA Transactions on Signal and Information Processing* 11: 1-64. https://doi.org/10.1561/116.00000050.

Malik, Mishaim, Muhammad Kamran Malik, Khawar Mehmood, e Imran Makhdoom. 2020. "Automatic Speech Recognition: a survey". *Multimedia Tools and Applications*, 80: 9411-57. https://doi.org/10.1007/s11042-020-10073-7

McCowan, Iain, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, e Hervé Bourlard. 2005. "On the Use of Information Retrieval Measures for Speech Recognition Evaluation". *IDIAP Research Report,* 04-73. IDIAP, Martigny, Switzerland.

Morris, Andrew Cameron, Viktoria Maier, e Phil Green. 2004. "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition". *Proceedings of Interspeech 2004. Jeju Island, Korea, 4-8 October 2004*: 2765-68. https://doi.org/10.21437/Interspeech.2004-668.

Nissim, Malvina, e Ludovica Pannitto. 2022. *Che cos'è la linguistica computazionale*. Carocci editore.

Palmerini, Maria, e Renata Savy. 2014. "Gli errori di un sistema di riconoscimento automatico del parlato: analisi linguistica e primi risultati di una ricerca interdisciplinare". *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*. Pisa, 9-11 December 2014: 281-5.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, e Karel Veselý. 2011. "The Kaldi Speech Recognition Toolkit". *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Hawaii, US.* IEEE Signal Processing Society.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, e Ilya Sutskever. 2023. "Robust Speech Recognition via Large-scale Weak Supervision". *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*: 28492-518.

Rekesh, Dima, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, e Boris Ginsburg. 2023. "Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition". *Proceedings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU):* 1-8. https://doi.org/10.1109/ASRU57964.2023.10389701.

Savy, Renata e Francesco Cutugno. 2009. "CLIPS. Diatopic, Diamesic and Diaphasic Variations in Spoken Italian". *Proceedings of the 5th Corpus Linguistics Conference (CL2009)*: 20-23.

Turing, Alan Mathison. 1950. "I.—COMPUTING MACHINERY AND INTELLIGENCE." *Mind,* LIX(236): 433-60. https://doi.org/10.1093/mind/LIX.236.433.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, e Illia Polosukhin. 2017. "Attention is All You Need". *Proceedings of the 31st International Conference on Neural Information Processing Systems*: 6000-10.

Vitale, Norman, Emilia Tanda e Francesco Cutugno. 2024. "Towards a Responsible Usage of AI-based Large Acoustic Models for Automatic Speech Recognition: On the Importance of Data in the Self-supervised Era". *Atti quarto Convegno Nazionale CINI sull'Intelligenza Artificiale – Ital-IA 2024*. https://hdl.handle.net/11588/974957.