Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Tesi di Laurea

Pipeline metodologica per l'analisi semantica e la valutazione degli errori di pronuncia nella lingua Thai

Relatore

Prof. **Loredana Caruccio**

Candidato

Damiana Buono

Matricola: 0522501592

Correlatore

Dott. **Bernardo Breve**

Anno Accademico 2024/2025

"Non si deve mai avere paura di dire "Non lo so" o "Non capisco" o di fare domande "stupide", poiché nessuna domanda è stupida. Bisogna continuare anche quando le cose sembrano impossibili, anche quando i cosiddetti esperti dicono che è impossibile; bisogna essere autonomi ed essere diversi; non aver paura di sbagliare o di ammettere i proprio errori, perché solo chi fallisce tanto, può avere tanto successo."

Margaret Heafield Hamilton

Tratto da un intervista del 2016

Abstract

L'apprendimento della lingua Thai rappresenta una sfida significativa per i parlanti non nativi, soprattutto a causa della sua natura tonale e della complessità fonetica. In questo contesto, la disponibilità di strumenti automatici di supporto può facilitare il miglioramento della pronuncia e fornire feedback mirati durante il processo di apprendimento.

La tesi propone una pipeline metodologica per l'analisi semantica e la valutazione automatica degli errori di pronuncia nella lingua Thai, con l'obiettivo di stimare non solo la correttezza, ma anche l'impatto comunicativo degli errori.

La pipeline è stata costruita a partire dal corpus *LOTUS*, manipolato per generare un insieme di audio contenenti errori tipici degli apprendenti. Questo corpus è stato utilizzato per il *fine-tuning* del modello di riconoscimento vocale *wav2vec2-large-xlsr-53-th*, adattato a produrre trascrizioni fedeli alle pronunce errate senza correzione automatica. Le trascrizioni ottenute sono state poi analizzate mediante il modello semantico *paraphrase-multilingual-MiniLM-L12-v2*, impiegato per misurare la similarità semantica tra la frase di riferimento e quella pronunciata.

I risultati mostrano che l'aumento della frequenza e della gravità degli errori di pronuncia comporta una riduzione proporzionale della similarità semantica, evidenziando la validità dell'approccio integrato. Il sistema fornisce così una valutazione congiunta fonetico-semantica utile per generare feedback mirati e supportare l'apprendimento della pronuncia in contesti di lingua tonale.

Indice

CAPITOLO 1

Introduzione

La crescente diffusione delle tecnologie vocali, che vanno dai sistemi di assistenza virtuale alle piattaforme per l'apprendimento delle lingue, ha suscitato un notevole interesse scientifico nell'ambito dell'elaborazione automatica del linguaggio naturale (NLP). Negli ultimi anni, questo campo ha fatto passi da gigante, grazie all'uso di modelli neurali profondi e all'accesso a vasti corpora testuali e orali in diverse lingue (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Liu et al., 2019). In particolare, l'Automatic Speech Recognition (ASR) ha tratto vantaggio dall'introduzione delle architetture Transformer e da tecniche di apprendimento auto-supervisionato, come nel caso di Wav2Vec2.0 (Baevski et al., 2020), che hanno permesso ai sistemi di raggiungere livelli di accuratezza molto elevati anche in ambienti acusticamente complessi.

Nonostante tali progressi, l'applicazione dei modelli ASR alle lingue tonali, tra cui il Thai, continua a presentare sfide specifiche (Wutiwiwatchai and Furui, 2006; Potisuk, 2010), riconducibili alla struttura fonologica della lingua e alla marcata dipendenza semantica dai toni e dalle durate vocaliche.

La lingua Thai si distingue per un sistema fonologico articolato, costituito da 44 consonanti, 16 simboli vocalici che danno origine ad almeno 32 combinazioni e cinque segni tonali[1]. Nonostante tale ricchezza grafemica, soltanto 21 fonemi consonantici sono effettivamente utilizzati nella lingua moderna, suddivisi in tre classi tonali (alta, media e bassa) che influenzano

[1] `https://en.wikipedia.org/wiki/Thai_script`

la realizzazione fonetica e il significato lessicale[2]. Il sistema vocalico comprende 27 vocali brevi e lunghe, la cui durata è fonemicamente distintiva e può determinare variazioni semantiche. Inoltre, il Thai presenta cinque toni fondamentali – medio, basso, cadente, alto e crescente – aventi valore distintivo e funzione semantica[3].

Un'ulteriore complessità è rappresentata dall'assenza di spazi tra le parole nel testo scritto, aspetto che rende la segmentazione automatica un compito particolarmente impegnativo. Tali caratteristiche richiedono dunque l'elaborazione di soluzioni specifiche, in grado di integrare conoscenze fonetiche e modelli di comprensione semantica, poiché i modelli ASR multilingue di tipo generalista tendono a trascurare informazioni prosodiche e fonetiche determinanti per la comprensione corretta dell'enunciato.

In questo contesto, la presente tesi introduce una **pipeline metodologica per l'analisi semantica e la valutazione degli errori di pronuncia nella lingua Thai**, finalizzata a integrare l'analisi fonetica con la valutazione semantica, al fine di misurare non solo la correttezza articolatoria ma anche l'impatto comunicativo degli errori. L'approccio proposto intende superare i limiti dei metodi tradizionali, che si concentrano sulla rilevazione di deviazioni fonetiche senza considerare le conseguenze semantiche prodotte dall'errore.

1.1 Motivazioni e Obiettivi

Dopo aver delineato le principali sfide linguistiche e tecnologiche connesse al Thai, è possibile definire le motivazioni alla base della presente ricerca.

L'apprendimento del Thai da parte di parlanti non nativi, in particolare di area europea, risulta complesso a causa della stretta dipendenza del significato lessicale da parametri tonali e temporali. Errori apparentemente minimi nella realizzazione di un tono o nella durata di una vocale possono determinare variazioni semantiche significative o compromettere la comprensione complessiva dell'enunciato. Tali considerazioni evidenziano la necessità di strumenti didattici in grado di valutare simultaneamente la correttezza fonetica e l'effetto semantico degli errori di pronuncia.

A tal fine, è stato sviluppato un **assistente intelligente per l'apprendimento della pronuncia Thai**, basato su un flusso di elaborazione che integra riconoscimento vocale, tokenizzazione e analisi semantica. La componente ASR impiega modelli pre-addestrati e successivamente ottimizzati per il Thai, quali *wav2vec2-large-xlsr-53-th* (Phatthiyaphaibun

[2] `https://en.wikipedia.org/wiki/Thai_phonology`

[3] `https://en.wikipedia.org/wiki/Thai_language`

et al., 2022), *Whisper* (Radford et al., 2022; Aung et al., 2024) e *Porjai* (Thatphithakkul et al., 2024), capaci di catturare le caratteristiche tonali e prosodiche del parlato. Le trascrizioni generate vengono successivamente segmentate mediante il modello *newmm* di PyThaiNLP (Phatthiyaphaibun et al., 2023), mentre la valutazione semantica si fonda su modelli di embedding contestuali quali *WangchanBERTa* (Lowphansirikul and et al., 2021), *PhayaThaiBERT* (Sriwirote and et al., 2023) e *LaBSE* (Feng and et al., 2020), che consentono di quantificare la similarità semantica tra la frase di riferimento e quella pronunciata.

Per l'addestramento e la validazione della pipeline è stato costruito un dataset di riferimento comprendente pronunce errate generate artificialmente mediante iniezione controllata di errori fonetici. Gli audio sintetici sono stati prodotti attraverso i modelli *KhanomTan TTS* (Wannaphong, 2022) e *gTTS* (Durette, 2025), al fine di garantire un elevato grado di realismo acustico e di controllo sui parametri di variazione tonale e vocalica. Tale corpus ha permesso di analizzare in modo sistematico la relazione tra deviazioni fonetiche e perdita di significato, costituendo una risorsa fondamentale per la progettazione di un sistema didattico avanzato.

L'obiettivo complessivo è la realizzazione di uno strumento in grado di fornire un feedback mirato, segnalando non solo la presenza di errori di pronuncia, ma anche il grado di alterazione semantica indotta, al fine di promuovere un apprendimento più consapevole ed efficace.

1.2 Risultati

La pipeline proposta integra moduli di riconoscimento vocale, tokenization, rappresentazione semantica e analisi della similarità, consentendo una valutazione congiunta dell'accuratezza fonetica e della coerenza semantica. Gli esperimenti condotti hanno evidenziato una correlazione negativa tra la complessità e la frequenza degli errori di pronuncia e il grado di similarità semantica tra la frase di riferimento e quella pronunciata. Tale relazione conferma la validità dell'approccio integrato per la valutazione automatica della pronuncia nella lingua Thai, dimostrando che le deviazioni fonetiche più marcate comportano una perdita proporzionale di coerenza semantica.

Il modello ASR *wav2vec2-large-xlsr-53-th* è stato sottoposto a un processo di *fine-tuning* sul corpus *LOTUS* (Sertsi et al., 2016), opportunamente manipolato per includere campioni audio contenenti errori di pronuncia tipici degli apprendenti di Thai. Tale procedura ha consentito di adattare il modello alla variabilità fonetica del parlato non nativo, migliorando in modo significativo la precisione di trascrizione e incrementando la robustezza del sistema rispetto alle variazioni tonali e prosodiche. L'analisi quantitativa ha mostrato una riduzione

sensibile del tasso di errore di carattere (Character Error Rate, CER) e un miglioramento nella distinzione automatica dei toni, dimostrando l'efficacia del fine-tuning mirato su dati realistici e controllati.

Parallelamente, l'impiego di modelli di embedding multilingue ha consentito di rappresentare in maniera più accurata le relazioni semantiche tra enunciati, favorendo una valutazione del significato indipendente dalle sole corrispondenze lessicali. L'integrazione del modello come *paraphrase- multilingual- MiniLM-L12-v2* (Mahmoud et al., 2025) ha reso possibile la misura della similarità semantica tra frasi mediante rappresentazioni vettoriali contestuali, permettendo la generazione di feedback linguistici più coerenti, informativi e sensibili al contesto comunicativo.

Infine, i moduli TTS (Durette, 2025) sviluppati si sono rivelati risorse di particolare utilità sia per la validazione acustica dei dati, consentendo un confronto oggettivo tra audio sintetici e reali, sia per l'integrazione in applicazioni didattiche interattive. In tale ambito, essi rappresentano uno strumento complementare per l'addestramento percettivo degli apprendenti, offrendo esempi di pronuncia corretta e simulazioni controllate di errori tonali e vocalici.

Nel complesso, i risultati ottenuti confermano la validità dell'approccio proposto, evidenziando come la combinazione di tecniche di riconoscimento vocale, rappresentazioni semantiche e sintesi vocale possa costituire una base solida per la realizzazione di sistemi intelligenti di supporto all'apprendimento e alla valutazione automatica della pronuncia nelle lingue tonali.

1.3 Struttura della tesi

La tesi è articolata in sette capitoli. Il **Capitolo 2** presenta lo stato dell'arte, analizzando i principali modelli di riconoscimento vocale, di rappresentazione semantica e di sintesi vocale per il Thai. Il **Capitolo 3** illustra le sfide specifiche delle lingue tonali e introduce l'approccio metodologico adottato. Il **Capitolo 4** descrive la costruzione del dataset e le procedure di *fine-tuning* dei modelli ASR, oltre alla pipeline di iniezione di errori fonetici. Il **Capitolo 5** approfondisce l'analisi semantica e la generazione automatica del feedback linguistico. Il **Capitolo 6** riporta la valutazione sperimentale e la discussione dei risultati ottenuti. Infine, il **Capitolo 7** presenta le conclusioni e delinea le prospettive future, tra cui l'estensione del sistema a lingue tonali affini e l'integrazione con modelli di apprendimento adattivo.

Nel complesso, la ricerca intende contribuire allo sviluppo di sistemi intelligenti per la valutazione automatica della pronuncia e per il supporto all'apprendimento delle lingue to-

nali, ponendo le basi per futuri studi sull'interazione tra fonetica, semantica e intelligenza artificiale.

CAPITOLO 2

Stato dell'arte

In letteratura si possono distinguere due principali linee di sviluppo per il riconoscimento vocale nelle lingue tonali. La prima riguarda i modelli *end-to-end* Speech-to-Phoneme (S2P), che mappano direttamente il segnale acustico in sequenze fonemiche. Questi modelli presentano il vantaggio teorico di limitare gli errori che si accumulano durante l'elaborazione sequenziale dei dati, consentendo di apprendere direttamente le relazioni tra segnali acustici e rappresentazioni fonetiche.Tuttavia, la loro applicazione al Thai incontra ostacoli significativi: la scarsità di dati fonetici annotati, la complessità tonale e prosodica, e la limitata flessibilità analitica dei modelli diretti ne riducono l'efficacia. Studi come quello di Wutiwiwatchai and Furui (2006) e Potisuk (2010) evidenziano come la mancanza di corpora fonetici e la complessità tonale rappresentino barriere sostanziali per lo sviluppo di modelli end-to-end affidabili.

La seconda linea di ricerca, che costituisce la base della presente tesi, si concentra sull'analisi semantica del parlato. Invece di valutare la correttezza fonetica dei singoli suoni, l'obiettivo è determinare se la frase pronunciata dallo studente trasmetta lo stesso contenuto della frase corretta. A questo scopo, si utilizza un sistema di trascrizione automatica (STT) basato su modelli di apprendimento auto-supervisionato [Baevski et al. (2020) e Phatthiyaphaibun et al. (2022)], seguito dalla rappresentazione delle trascrizioni tramite embedding contestuali [Liu et al. (2019) e Feng and et al. (2020)]. Tali rappresentazioni dense catturano informazioni lessicali e prosodiche, consentendo di confrontare le frasi mediante metriche di similarità come la cosine similarity. Questo approccio permette di quantificare la similarità

semantica tra la frase dello studente e quella corretta, evidenziando differenze di significato, concetti mancanti o divergenze semantiche, e fornendo un feedback mirato sul contenuto informativo anziché sui singoli errori fonetici.

Questa strategia guida la ricerca affrontata: pur riconoscendo i limiti dei modelli S2P per lingue tonali e a risorse limitate [Bisani and Ney (2008) e Řezáčková et al. (2021)], l'analisi semantica consente di focalizzarsi sugli aspetti più rilevanti per l'apprendimento del Thai, ossia la corretta trasmissione del significato delle frasi, integrando informazioni prosodiche e semantiche in un quadro interpretativo applicabile a contesti didattici reali.

Negli ultimi anni, diversi studi hanno esplorato lo sviluppo di sistemi *end-to-end* capaci di convertire direttamente il parlato in sequenze di fonemi, senza passaggi intermedi basati sulla trascrizione testuale. Questi modelli, noti come Speech-to-Phoneme (S2P), offrono il vantaggio teorico di ridurre gli errori accumulati nelle pipeline modulari e di apprendere direttamente la mappatura acustico-fonetica [Bisani and Ney (2008) e Nguyen et al. (2023)]. Tuttavia, l'applicazione di tali approcci al thailandese presenta sfide significative, legate principalmente alla scarsità di dati fonetici annotati, alla complessità tonale e prosodica della lingua e alla limitata flessibilità analitica dei modelli diretti (Wutiwiwatchai and Furui, 2006).

La disponibilità di corpora annotati a livello fonetico per il thailandese è estremamente limitata. Il sistema fonologico thailandese comprende 44 consonanti, circa 32 vocali (inclusi dittonghi e combinazioni vocaliche) e 5 toni lessicali, ciascuno dei quali può alterare il significato di una parola anche in presenza di minime variazioni nell'altezza o nella durata. La creazione di dataset sufficientemente ricchi e bilanciati risulta quindi complessa e costosa [Sertsi et al. (2016)e Ardila et al. (2019)]. La scarsità di risorse annotate riduce la capacità dei modelli end-to-end di apprendere le sottili variazioni prosodiche e tonali che caratterizzano il parlato thailandese [(Phatthiyaphaibun et al., 2022)]. Come sottolineato da Wutiwiwatchai and Furui (2006), la mancanza di corpora fonetici rappresenta una barriera significativa allo sviluppo di modelli diretti affidabili.

Un ulteriore ostacolo è rappresentato dalla complessità tonale e prosodica. I modelli S2P che hanno mostrato buone prestazioni su lingue non tonali tendono a faticare nel catturare le variazioni tonali e le differenze di durata sillabica tipiche del thailandese. Anche strumenti avanzati di segmentazione acustica rischiano di non rappresentare adeguatamente la dinamica tonale, compromettendo la qualità della conversione fonetica. Lo studio di Potisuk (2010) evidenzia come le caratteristiche acustiche specifiche del thailandese richiedano un'analisi dettagliata difficilmente integrabile in un modello end-to-end standard.

Parallelamente, la letteratura mostra come l'integrazione o il confronto con modelli Grapheme-

to-Phoneme (G2P) possa rappresentare un'alternativa o un complemento utile. Modelli G2P come T5G2P (Řezáčková et al., 2021) o CharsiuG2P (Zhu et al., 2022) trattano la conversione grapheme-to-phoneme come un problema di traduzione testuale e si sono dimostrati flessibili anche in contesti multilingue. L'uso di rappresentazioni fonemiche pre-addestrate, come in XPhoneBERT (Nguyen et al., 2023), può migliorare la generalizzazione anche su lingue tonali, catturando intonazione e coarticolazione. Tuttavia, anche questi modelli richiedono dati testuali e fonemici di qualità e spesso non considerano direttamente le variazioni acustiche specifiche del parlato reale.

Infine, i modelli diretti mostrano una limitata flessibilità analitica rispetto a pipeline modulari che combinano STT e G2P. Essi tendono a non sfruttare strumenti ottimizzati per compiti specifici e non consentono facilmente l'integrazione di informazioni aggiuntive, come annotazioni prosodiche o regole fonotattiche della lingua. In questo senso, gli approcci modulari che integrano componenti G2P dedicate possono superare alcune delle limitazioni pratiche dei modelli end-to-end quando applicati a lingue a risorse limitate e con alta complessità fonetica come il thailandese.

2.1 Approccio modulare: Speech-to-Text e Embedding semantico

Gli approcci end-to-end, pur avendo mostrato grandi potenzialità, presentano ancora limiti su lingue con fenomeni fonetici e prosodici complessi come il Thai. Una delle principali criticità riguarda la scarsa trasparenza e controllabilità dei processi interni: gli errori sono difficili da interpretare e correggere. Un'architettura modulare consente invece di intervenire su moduli specifici, migliorando robustezza e adattabilità.

- **Fase Speech-to-Text (STT):** Il segnale acustico viene trascritto in forma testuale. Modelli multilingue come Whisper (Radford et al., 2022), addestrati su grandi quantità di dati, garantiscono trascrizioni robuste anche in presenza di rumore o dialetti. Tecniche self-supervised come wav2vec 2.0 (Baevski et al., 2020) apprendono rappresentazioni acustiche trasferibili a diversi scenari linguistici. Adattamenti specifici al Thai (Phatthiyaphaibun et al., 2022; Aung et al., 2024) dimostrano l'efficacia anche su lingue a risorse limitate.

- **Modelli di embedding:** Il testo trascritto viene convertito in vettori che catturano informazioni semantiche. Modelli classici come Word2Vec (Mikolov et al., 2013) e GloVe

(Pennington et al., 2014) forniscono embedding statici, mentre fastText (Bojanowski et al., 2017) gestisce meglio parole rare o complesse. Modelli basati su Transformer come RoBERTa (Liu et al., 2019) e versioni monolingua per il Thai come WangchanBERTa (Mahmoud et al., 2025) e PhayaThaiBERT (Sriwirote et al., 2023) producono rappresentazioni dipendenti dal contesto. Per la valutazione della similarità tra frasi, modelli di *sentence embeddings* come LaBSE (Feng and et al., 2020) permettono confronti efficaci, mentre versioni più compatte e veloci dei modelli, come MiniLM e paraphrase-multilingual-MiniLM (Mahmoud et al., 2025) riducono i tempi di calcolo mantenendo buone prestazioni.

2.2 Corpus e Modelli per la Fase Speech-to-Text (STT)

Il riconoscimento automatico del parlato (ASR) per il Thai si fonda in larga misura sulla disponibilità di corpora annotati, che rappresentano la base per l'addestramento e la valutazione dei modelli. La varietà, la dimensione e la qualità dei dati determinano in modo cruciale le prestazioni dei sistemi: dai dataset telefonici storici e di dimensioni ridotte, fino ai corpora multilingue su scala web, la letteratura mostra una progressiva evoluzione delle risorse disponibili.

2.2.1 Corpora per il Riconoscimento Automatico del Parlato Thai

Lo sviluppo di sistemi di riconoscimento automatico del parlato (ASR) per la lingua Thai si fonda in maniera cruciale sulla disponibilità di corpora annotati, che costituiscono la base per l'addestramento, la valutazione e la comparazione dei modelli. La qualità, la varietà e la quantità dei dati disponibili determinano in modo significativo le prestazioni dei sistemi, influenzando la capacità dei modelli di apprendere le regolarità fonetiche e prosodiche della lingua. Nel corso degli anni, la letteratura mostra una progressiva evoluzione delle risorse, dai dataset relativamente piccoli e foneticamente bilanciati fino a collezioni più ampie e diversificate che includono parlato spontaneo, dialetti regionali e condizioni acustiche variabili. Questi corpora non solo supportano lo sviluppo di modelli ASR accurati, ma costituiscono anche la base per la costruzione di embedding di token e frasi, rappresentazioni fondamentali per attività di NLP quali l'analisi della similarità semantica, la comprensione del parlato e l'elaborazione del linguaggio naturale in contesti Thai.

Tra le prime risorse significative si colloca la serie di corpora LOTUS, sviluppata dal National Electronics and Computer Technology Center (NECTEC). Questa serie comprende

diverse versioni: *LOTUS* (parlato letto, circa 55 ore), *LOTUS-Cell 2.0* (parlato telefonico, 86–90 ore) e *LOTUS-SOC* (parlato spontaneo). Nonostante la dimensione limitata rispetto agli standard attuali, i dataset LOTUS hanno fornito materiale foneticamente bilanciato e trascrizioni manuali di elevata qualità, diventando fondamentali per le prime sperimentazioni sull'ASR Thai (Sertsi et al., 2016)[1]. In particolare, LOTUS-Cell 2.0 è stato progettato per applicazioni di traduzione vocale su reti telefoniche, fornendo un modello metodologico per studi pionieristici nel riconoscimento vocale della lingua.

Con l'avvento di raccolte crowdsourced, il progetto Common Voice di Mozilla ha introdotto una nuova generazione di corpora multilingue, comprendendo anche la lingua Thai con oltre 170 ore di parlato validate nelle versioni più recenti (v9–v12) (Ardila et al., 2019; Mozilla Foundation / Common Voice team)[2]. Questo corpus presenta una maggiore varietà di parlanti, accenti e condizioni di registrazione rispetto ai dataset storici, consentendo lo sviluppo di modelli più robusti e generalizzabili. La natura collaborativa della raccolta introduce variabilità nelle trascrizioni, rendendo Common Voice particolarmente adatto al *fine-tuning* di modelli pre-addestrati, come wav2vec 2.0 o Whisper, piuttosto che all'addestramento da zero.

Il Gowajee Corpus (Chuangsuwanich et al., 2020), sviluppato presso la Chulalongkorn University, rappresenta una risorsa open-source di dimensioni contenute, con circa 17 ore di parlato distribuite tra 188 speaker[3]. Sebbene non sia sufficiente per addestrare modelli di grandi dimensioni, il corpus si rivela estremamente utile per scopi didattici, prototipazione rapida di sistemi ASR e sperimentazioni preliminari, offrendo una base pratica per test e studi accademici focalizzati sul parlato Thai.

Un elemento distintivo della lingua Thai è l'elevata variabilità dialettale, che costituisce una sfida significativa per modelli ASR addestrati esclusivamente sul Thai standard. Per affrontare questa complessità è stato sviluppato il Thai-Dialect Corpus (Suwanbandit et al., 2023)[4], che raccoglie parlato da dieci varietà regionali. Questo corpus ha consentito di sperimentare approcci di *transfer learning* e *curriculum learning*, favorendo la riduzione del Word Error Rate (WER) [Baevski et al. (2020), Phatthiyaphaibun et al. (2022) e Aung et al. (2024)]

[1]Licenza: CC BY-SA-NC 3.0. Permette la condivisione e modifica per scopi non commerciali con attribuzione e mantenendo la stessa licenza.

[2]Licenza: CC0-1.0. Dominio pubblico, uso libero senza restrizioni.

[3]Licenza: CC BY-SA 4.0. Permette la condivisione e modifica anche a fini commerciali, con attribuzione e mantenendo la stessa licenza.

[4]Licenza: CC BY-SA 4.0. Permette la condivisione e modifica anche a fini commerciali, con attribuzione e mantenendo la stessa licenza.

e incrementando la robustezza dei modelli rispetto a sistemi addestrati unicamente sullo standard. Il Thai-Dialect Corpus rappresenta quindi una risorsa fondamentale per rendere l'ASR Thai realmente utilizzabile in scenari quotidiani, in cui la gestione delle differenze fonetiche e prosodiche tra dialetti influisce direttamente sulla qualità della trascrizione.

Tabella 2.1: Corpora per il Riconoscimento Automatico del Parlato Thai

Corpus	Anno	Ore di parlato	Caratteristiche principali
LOTUS (Sertsi et al., 2016)	2016	55 ore	Primo corpus foneticamente bilanciato per il Thai. Trascrizioni manuali di alta qualità, utile per esperimenti pionieristici in ASR. Licenza CC BY-SA-NC 3.0.
LOTUS-Cell 2.0 (Sertsi et al., 2016)	2016	86–90 ore	Progettato per traduzione vocale su reti telefoniche. Dataset metodologicamente accurato ma di diffusione limitata.
LOTUS-SOC (Sertsi et al., 2016)	2016	172 ore	Include parlato naturale e spontaneo, utile per studiare scenari reali di trascrizione vocale.
Common Voice Thai (Ardila et al., 2019; Mozilla Foundation / Common Voice team)	2019+	170 ore	Corpus crowdsourced, variabilità di parlanti, accenti e condizioni di registrazione. Licenza aperta CC0-1.0, ideale per fine-tuning di modelli pre-addestrati.
Gowajee Corpus (Chuangsuwanich et al., 2020)	2020	17 ore	Open-source, dimensioni contenute ma utile per prototipazione, didattica e test preliminari. Licenza CC BY-SA 4.0.
Thai Dialect Corpus (Suwanbandit et al., 2023)	2023	840 ore	Copre 10 varietà regionali. Supporta transfer learning e curriculum learning. Riduce WER su dialetti rispetto a sistemi addestrati solo sul Thai standard.

2.2.2 Modelli Pre-addestrati per la Fase STT/ASR

Negli ultimi anni, lo sviluppo di modelli pre-addestrati per il riconoscimento automatico del parlato (ASR) ha subito una rapida evoluzione, soprattutto grazie ai progressi nell'ap-

prendimento auto-supervisionato e alle architetture basate su Transformers. Questi modelli permettono di catturare rappresentazioni acustiche ricche a partire da grandi quantità di audio non etichettato, riducendo la dipendenza da dati annotati manualmente, tradizionalmente costosi e difficili da ottenere per lingue meno rappresentate come il Thai.

Il modello **Wav2Vec2.0 Thai** rappresenta un adattamento del modello multilingue **XLSR-Wav2Vec2** alla lingua Thai (Baevski et al., 2020). La versione originale di Wav2Vec 2.0 combina un encoder convoluzionale per l'estrazione di caratteristiche dai segnali audio grezzi, un modulo di quantizzazione per creare rappresentazioni discreti dei segmenti acustici e un Transformer che produce rappresentazioni contestualizzate tramite un loss contrastivo auto-supervisionato. Questa struttura consente al modello di apprendere efficacemente le regolarità fonetiche e prosodiche della lingua senza supervisionare direttamente le trascrizioni.

Il fine-tuning su **Wav2Vec2.0 Thai** è stato effettuato utilizzando il corpus **Common-Voice V8**, con uno split studiato per evitare sovrapposizioni tra parlanti nei set di addestramento e test (Phatthiyaphaibun et al., 2022). Tale accorgimento metodologico è cruciale, poiché versioni precedenti del dataset mostravano la presenza ripetuta degli stessi speaker in più split, portando a stime di performance eccessivamente ottimistiche. Grazie alla combinazione di pretraining auto-supervisionato e dati annotati, **Wav2Vec2.0 Thai** cattura in maniera efficace le caratteristiche fonetiche del Thai standard e fornisce rappresentazioni acustiche trasferibili ad altri corpora, come il **Gowajee Corpus** (Chuangsuwanich et al., 2020) e **LOTUS** (Sertsi et al., 2016). L'integrazione di tali risorse aumenta la robustezza del modello, anche se la gestione di varianti regionali e di parlato spontaneo richiede ulteriori adattamenti.

Un secondo modello di rilievo è **Whisper**, introdotto da OpenAI e addestrato su circa 680.000 ore di audio raccolto dal web (Radford et al., 2022). Whisper si distingue per la sua capacità di generalizzazione in scenari multilingue e rumorosi, rendendolo particolarmente adatto a lingue con caratteristiche fonetiche complesse come il Thai. Studi sperimentali dimostrano che Whisper fornisce trascrizioni di alta qualità anche in condizioni acustiche non ottimali, caratteristica fondamentale per applicazioni reali come sistemi di dettatura, assistenti vocali o trascrizione automatica di contenuti multimediali. Per migliorare ulteriormente le prestazioni sul Thai, è stato sviluppato **Thonburian Whisper**, un'estensione fine-tuned su dataset eterogenei che comprendono CommonVoice, Gowajee, Thai Dialect Corpus, Thai Elderly Speech e altre risorse curate (Aung et al., 2024). L'utilizzo di questi dati diversificati consente al modello di gestire meglio parlato spontaneo, varianti dialettali e parlanti anziani, riducendo significativamente il Word Error Rate (WER) (Baevski et al., 2020) rispetto alla

versione originale e sottolineando l'importanza di un adattamento locale.

Infine, il modello **Porjai** (Thatphithakkul et al., 2024), sviluppato da Chulalongkorn University e NECTEC, è stato concepito per affrontare la notevole variabilità dialettale presente in Thailandia. Il training è stato condotto principalmente sul **Thai Dialect Corpus** (Suwanbandit et al., 2023), che raccoglie parlato da dieci varietà regionali. Questa impostazione consente a **Porjai** di ridurre gli errori di trascrizione rispetto a modelli addestrati unicamente sul Thai standard, evidenziando l'importanza di includere rappresentazioni bilanciate dei dati dialettali. La letteratura sottolinea come l'inclusione di più varietà regionali migliori la robustezza dei sistemi ASR e permetta di affrontare efficacemente le sfide derivanti dalle differenze fonetiche, prosodiche e lessicali tra dialetti.

In generale, l'adozione di modelli pre-addestrati come Wav2Vec2.0 Thai, Whisper e Porjai mostra chiaramente i vantaggi dell'approccio auto-supervisionato e della fine-tuning su dataset specifici per la lingua target. Questi modelli non solo riducono la necessità di grandi quantità di dati annotati, ma consentono anche di sviluppare sistemi più resilienti e adattabili alle varianti regionali e ai contesti reali di parlato spontaneo, rappresentando un passo fondamentale verso sistemi ASR di alta qualità per il Thai.

Tabella 2.2: Confronto tra modelli STT/ASR per la lingua Thai

Modello	**Anno**	**Dataset Principale**	**Caratteristiche principali**
Wav2Vec2.0 Thai (Phatthiyaphaibun et al., 2022)	2022	CommonVoice V8	Adattamento del modello auto-supervisionato Wav2Vec2 per il Thai. Fine-tuning con split senza overlap tra parlanti. Buone performance su parlato standard e robustezza trasferibile a Gowajee e LOTUS. WER stimato: 8-12% su test set standard.
Whisper (Radford et al., 2022)	2022	680.000 ore multilingue	Modello multilingue OpenAI, robusto a rumore e varianti fonetiche. Buone prestazioni su parlato spontaneo e condizioni acustiche difficili. WER stimato: 10-15% sul Thai standard.

Modello	Anno	Dataset Principale	Caratteristiche principali
Thonburian Whisper (Aung et al., 2024)	2024	CommonVoice, Gowajee, Thai Dialect Corpus, Thai Elderly Speech	Fine-tuning locale su dataset diversificati per gestire varianti dialettali e parlanti anziani. Riduce significativamente il WER rispetto a Whisper standard, migliorando l'accuratezza su parlato spontaneo e code-switching. WER stimato: 6-10%.
Porjai (Suwanbandit et al., 2023)	2023	Thai Dialect Corpus	Addestrato su 10 varietà regionali, ottimizzato per la gestione delle differenze dialettali e fonetiche. Migliora la robustezza rispetto a modelli monolingua standard. WER stimato: 7-11% su dialetti.

2.2.3 Modelli di Embedding

La costruzione di embedding per la lingua Thai richiede tre componenti fondamentali: corpora ampi e diversificati, algoritmi capaci di trasformare il testo in rappresentazioni numeriche ricche di informazione e metriche adeguate per valutarne la qualità. Solo l'integrazione di queste tre dimensioni consente di ottenere modelli che riflettano non soltanto la semantica, ma anche le peculiarità morfologiche e sintattiche della lingua. Nel caso del Thai, tali esigenze si amplificano per via della mancanza di spazi tra le parole, della forte presenza di prestiti linguistici non completamente adattati al sistema ortografico e fonologico e della variazione significativa tra registri comunicativi. Questi elementi rendono particolarmente sfidante il compito di costruire embedding robusti, in grado di generalizzare in scenari applicativi eterogenei.

I primi modelli sviluppati, come **Word2Vec** (Mikolov et al., 2013) e **GloVe** (Pennington et al., 2014), hanno rappresentato una pietra miliare nello studio delle rappresentazioni distribuite. **Word2Vec**, attraverso architetture come Continuous Bag of Words e Skip-Gram, apprende vettori basandosi sul contesto locale delle parole. **GloVe**, invece, sfrutta le statistiche globali di co-occorrenza per posizionare le parole in uno spazio semantico coerente. Entrambi i modelli hanno permesso progressi significativi nei task di similarità semantica e classificazione, ma presentano limitazioni intrinseche: le rappresentazioni generate sono sta-

tiche, ovvero una parola polisemica viene associata a un solo vettore, indipendentemente dal contesto. Ciò risulta particolarmente problematico in Thai, lingua caratterizzata da un'ampia polisemia e da fenomeni di ambiguità lessicale. A complicare ulteriormente il quadro, l'assenza di segmentazione esplicita porta questi modelli a dipendere fortemente da algoritmi di tokenizzazione esterni, che non sempre riescono a gestire correttamente la morfologia o la composizione delle parole.

Una prima risposta a queste criticità è stata proposta da **fastText** (Bojanowski et al., 2017), che arricchisce la rappresentazione delle parole con informazioni sui sotto-componenti morfologici tramite n-grammi di caratteri. Questa caratteristica si è dimostrata particolarmente adatta al Thai, lingua che non segna esplicitamente i confini tra le parole e che presenta numerosi composti e varianti ortografiche. Grazie alla modellazione sublessicale, fastText è in grado di generare embedding anche per parole rare o non viste in fase di addestramento, superando in parte i problemi di out-of-vocabulary. Tuttavia, il modello rimane comunque vincolato alla logica statica: il significato di una parola rimane invariato anche se il contesto circostante ne altera l'interpretazione. In questo senso, fastText si colloca a metà strada tra le soluzioni puramente lessicali e i modelli contestuali basati su Transformer.

L'introduzione di architetture Transformer ha segnato un punto di svolta radicale, permettendo la creazione di rappresentazioni contestuali. In questo scenario, **WangchanBERTa** (Lowphansirikul and et al., 2021), basato sull'architettura **RoBERTa** (Liu et al., 2019), rappresenta il primo grande modello monolingua specifico per il Thai. È stato addestrato su circa 78 GB di testi provenienti da fonti eterogenee, incluse piattaforme social, articoli giornalistici e risorse pubbliche. Una delle scelte più innovative riguarda la preservazione degli spazi come token speciali: questa strategia si è rivelata essenziale per compensare l'assenza di delimitatori chiari tra le parole e per migliorare la segmentazione interna. Inoltre, gli autori hanno sperimentato diverse strategie di tokenizzazione, confrontando granularità a livello di parola, sillaba e subword, con l'obiettivo di massimizzare le prestazioni nei compiti downstream. WangchanBERTa ha mostrato prestazioni superiori a modelli multilingua come **mBERT** e **XLM-R**, dimostrando che un modello progettato su misura per il Thai riesce a catturare meglio le peculiarità della lingua. Tuttavia, il modello soffre ancora di alcune limitazioni: la dimensione del corpus, pur ampia, rimane inferiore a quella disponibile per lingue a maggiore disponibilità di risorse; inoltre, la gestione dei prestiti linguistici non assimilati risulta ancora parziale.

Per affrontare questa ultima sfida è stato sviluppato **PhayaThaiBERT** (Sriwirote and et al., 2023), che rappresenta un'estensione critica del lavoro precedente. Partendo dai pesi

di WangchanBERTa, PhayaThaiBERT integra un vocabolario più ampio, arricchito da token importati da **XLM-R** (Conneau et al., 2019). L'obiettivo principale era gestire in maniera più efficace i prestiti linguistici, in particolare quelli dall'inglese, che spesso compaiono in testi Thai senza adattamenti grafici. L'addestramento è stato condotto su un corpus ancora più ampio, pari a circa 156 GB, consentendo una maggiore copertura lessicale e stilistica. I risultati sperimentali hanno confermato che l'espansione del vocabolario e l'incremento dei dati di addestramento migliorano le prestazioni in scenari di code-switching, dove i testi alternano Thai e inglese. Nonostante ciò, la maggiore complessità del vocabolario comporta un aumento della frammentazione delle rappresentazioni e una crescita dei costi computazionali, rendendo il modello meno accessibile in contesti a risorse limitate.

Un ruolo particolare nel panorama degli embedding Thai è svolto da **XLM-R**, un modello multilingua addestrato su circa 100 lingue. Sebbene non raggiunga le stesse performance dei modelli monolingua nei task specifici, XLM-R si dimostra estremamente utile in applicazioni cross-lingua, come il transfer learning o il retrieval multilingue. La sua forza risiede nella possibilità di condividere rappresentazioni semantiche tra lingue differenti, facilitando lo sviluppo di sistemi in scenari dove i dati Thai sono limitati. La limitazione principale, tuttavia, è che il modello non è ottimizzato per catturare le specificità linguistiche del Thai, in quanto il vocabolario e le strategie di tokenizzazione sono progettati per essere generici e universali.

Infine, negli ultimi anni si sono diffusi modelli progettati non per singole parole, ma per intere frasi, con l'obiettivo di catturare significati a un livello semantico superiore. Tra questi, **LaBSE** (Feng and et al., 2020) rappresenta un modello di riferimento, capace di generare embedding multilingua per oltre 100 lingue. Il suo principale punto di forza risiede nella capacità di produrre rappresentazioni comparabili tra lingue diverse, rendendolo particolarmente adatto a compiti di retrieval cross-lingua e allineamento di corpora paralleli. **paraphrase-multilingual-MiniLM-L12-v2** (Mahmoud et al., 2025), invece, si colloca come una soluzione più leggera, derivata da un processo di distillazione su MiniLM. Il modello offre un compromesso tra accuratezza e velocità di inferenza, risultando utile in applicazioni real-time o in contesti con vincoli computazionali. Entrambi, pur non essendo specificamente progettati per il Thai, hanno dimostrato buone prestazioni in compiti di similarità semantica e rappresentano strumenti flessibili per sistemi multilingua.

Tabella 2.3: Confronto tra modelli di embedding per il Thai e multilingua

Modello	**Anno**	**Dataset Principale**	**Caratteristiche principali**
Word2Vec (Mikolov et al., 2013)	2013	Google News, Wikipedia	Basato su CBOW e Skip-Gram, produce vettori statici per parola. Semplice ed efficiente, ma non gestisce la polisemia né l'assenza di spazi tipica del Thai.
GloVe (Pennington et al., 2014)	2014	Common Crawl, Wikipedia	Usa statistiche globali di co-occorrenza. Robusto su similarità semantica, ma resta statico e dipende dalla qualità della tokenizzazione.
fastText (Bojanowski et al., 2017)	2017	Wikipedia, Common Crawl	Introduce n-grammi di caratteri, utile per lingue non segmentate e con composti come il Thai. Riesce a generare embedding per parole rare, ma rimane insensibile al contesto.
WangchanBERTa (Lowphansirikul and et al., 2021)	2021	OSCAR Thai (78 GB)	Basato su RoBERTa e addestrato solo su Thai. Gestisce spazi come token speciali e sperimenta diverse granularità di tokenizzazione. Ottime prestazioni sul Thai, ma meno generalizzabile ad altre lingue.
PhayaThaiBERT (Sriwirote and et al., 2023)	2023	Thai OSCAR + XLM-R (156 GB)	Estende WangchanBERTa con vocabolario arricchito da XLM-R e corpus di 156 GB. Migliora il code-switching Thai-Inglese, ma aumenta frammentazione e costi computazionali.
XLM-R (Conneau et al., 2019)	2020	Common Crawl (2.5 TB, 100+ lingue)	Basato su RoBERTa, addestrato su oltre 100 lingue. Utile per transfer learning e retrieval multilingue, ma meno accurato sul Thai rispetto ai modelli monolingua.

Modello	Anno	Dataset Principale	Caratteristiche principali
LaBSE (Feng and et al., 2020)	2020	Multilingual Wikipedia + ParaCrawl	Ottimizzato per embedding di frasi in 100+ lingue. Produce rappresentazioni comparabili cross-lingua, ideale per retrieval e allineamento di corpora paralleli.
paraphrase-multilingual-MiniLM-L12-v2 (Mahmoud et al., 2025)	2020	Multilingual paraphrase corpus (N/A)	Versione distillata di MiniLM. Più leggera e veloce, adatta a scenari real-time e dispositivi a risorse limitate. Buon compromesso tra accuratezza e costo computazionale.

2.3 Tecniche Complementari: Tokenizzazione e Sintesi Vocale

L'elaborazione automatica della lingua Thai richiede l'integrazione di tecniche complementari che permettano di garantire la coerenza tra rappresentazione testuale e fonetica. Tra queste, la **tokenizzazione** e la **sintesi vocale** rappresentano due componenti centrali nei sistemi di elaborazione del linguaggio naturale (NLP) e di tecnologia vocale.

La tokenizzazione, come descritto nella documentazione della libreria `PyThaiNLP` (Phatthiyaphaibun et al., 2023), costituisce il primo passo del processo di analisi linguistica e ha il compito di suddividere il testo continuo in unità lessicali significative. Nel caso del Thai, questa operazione è particolarmente complessa a causa dell'assenza di spazi tra le parole e della natura grafemica della lingua, che include segni tonali e diacritici. Studi come quelli di Rakpong Kittinaradorn et al. (2019) e AttaCut (Chormai et al., 2019) hanno mostrato come approcci basati su reti neurali profonde e rappresentazioni sillabiche possano migliorare sensibilmente la precisione della segmentazione, superando le limitazioni dei metodi puramente basati su dizionario.

Parallelamente, la sintesi vocale mira a trasformare il testo scritto in parlato naturale e intelligibile, preservando le caratteristiche tonali e prosodiche proprie del Thai. Il modello **KhanomTan TTS**, proposto da Wannaphong (2022), rappresenta uno degli esempi più avanzati di sintesi neurale specificamente progettata per la lingua Thai, grazie alla capacità di gestire tono, prosodia e variazioni lessicali. In contesti più generali, strumenti come **gTTS**

(Durette, 2025) offrono invece un approccio pratico e immediato, pensato per applicazioni che privilegiano la rapidità di generazione rispetto alla personalizzazione del parlato.

2.3.1 Tokenizzazione delle Parole in Thai

La tokenizzazione del testo in lingua Thai rappresenta una delle sfide più complesse nell'elaborazione del linguaggio naturale. A differenza di molte lingue occidentali, infatti, il Thai non utilizza spazi per separare le parole e presenta una struttura grafemica e fonetica peculiare, con vocali che possono trovarsi prima, dentro o dopo la consonante, segni diacritici e simboli tonali che ne aumentano ulteriormente la complessità. Questa caratteristica rende difficile identificare i confini tra le unità lessicali e ha portato allo sviluppo di diverse soluzioni, sia basate su dizionario sia su modelli di apprendimento automatico, che si distinguono per accuratezza, velocità e robustezza rispetto a domini linguistici diversi.

2.3.2 Tokenizzazione del Thai

La tokenizzazione del Thai rappresenta una fase cruciale nell'elaborazione del linguaggio naturale, poiché la lingua non utilizza spazi tra le parole e presenta una morfologia complessa e ricca di regolarità fonologiche e ortografiche. Una segmentazione accurata è essenziale non solo per compiti di analisi lessicale e sintattica, ma anche per la costruzione di rappresentazioni semantiche robuste, l'estrazione di informazioni e l'elaborazione automatica di testi su larga scala. Inoltre, la qualità della tokenizzazione influisce direttamente sulle prestazioni di modelli di embedding e sistemi di sintesi vocale, in quanto un input mal segmentato può degradare significativamente l'accuratezza dei modelli downstream.

Tra i primi approcci di tokenizzazione si colloca il modello **newmm** (Phatthiyaphaibun et al., 2023), integrato nella libreria `PyThaiNLP`. Si tratta di un sistema basato su dizionario, che sfrutta una tecnica di maximum matching combinata con regole derivate dai Thai Character Clusters (TCC), unità ortografiche minime in grado di rispettare i vincoli della scrittura. Questo approccio garantisce una segmentazione coerente e veloce, rendendo **newmm** adatto a sistemi con risorse computazionali limitate o a scenari che richiedono elaborazione in batch di grandi quantità di testo. La variante **newmm-safe** migliora ulteriormente le prestazioni su testi lunghi, introducendo meccanismi per ridurre l'ambiguità derivante da possibili confini lessicali multipli. Nonostante la sua efficienza e semplicità, **newmm** presenta limitazioni tipiche dei metodi basati esclusivamente su dizionario: è meno efficace nel gestire parole non registrate (out-of-vocabulary) e può produrre segmentazioni errate quando il contesto non è sufficiente a risolvere ambiguità complesse.

Per superare tali limiti, sono stati sviluppati approcci più sofisticati basati su reti neurali, come **DeepCut** (Rakpong Kittinaradorn et al., 2019). Questo modello affronta la tokenizzazione come un problema di classificazione binaria a livello di carattere, prevedendo se ciascun simbolo rappresenti l'inizio di una parola. L'utilizzo di Reti Neurali Convoluzionali (CNN) permette di catturare informazioni contestuali sui caratteri circostanti, migliorando la gestione di parole sconosciute e di ambiguità derivanti dal contesto linguistico. Addestrato su corpus ampi come il BEST2010, **DeepCut** (Rakpong Kittinaradorn et al., 2019) ha dimostrato performance elevate, con valori di F1 superiori al 97%, evidenziando la sua capacità di generalizzare a nuovi domini. Pur garantendo maggiore accuratezza e robustezza rispetto a **newmm**, **DeepCut** richiede risorse computazionali più elevate e tempi di elaborazione maggiori, fattori da considerare in applicazioni real-time o in sistemi a bassa capacità hardware.

Una proposta più recente, **AttaCut** (Chormai et al., 2019), combina i punti di forza di **newmm** e **DeepCut**, bilanciando efficienza e precisione. Questo modello utilizza convoluzioni dilatate e integra informazioni fonetiche e ortografiche sotto forma di syllable embeddings, catturando pattern locali tipici della lingua Thai. Sono disponibili due varianti principali: la versione "c", basata esclusivamente su caratteristiche dei caratteri, e la versione "sc", che include anche informazioni sillabiche. Tale architettura consente a **AttaCut** di ridurre drasticamente i tempi di inferenza, risultando fino a sei volte più veloce di **DeepCut**, con perdita minima di accuratezza. Grazie a queste caratteristiche, **AttaCut** si rivela ideale per applicazioni in tempo reale o su larga scala, dove è necessario un compromesso ottimale tra prestazioni computazionali e qualità della segmentazione.

Negli ultimi anni, gli approcci basati sui meccanismi di attenzione hanno mostrato risultati ancora più avanzati, come evidenziato dal lavoro di Chay-intr and Hidetaka Kamigaito (2021). Il modello proposto da Chay-intr and Hidetaka Kamigaito (2021) utilizza una rappresentazione basata sui caratteri combinata con meccanismi di **multiple attention**, che consentono di catturare contesti più ampi e dipendenze a lungo raggio tra i caratteri. Questa strategia migliora la precisione della segmentazione, soprattutto in presenza di ambiguità lessicali e parole sconosciute, superando alcune limitazioni dei modelli CNN tradizionali. I risultati sperimentali sul corpus di riferimento dimostrano l'efficacia degli approcci basati su attention nella tokenizzazione del Thai, aprendo la strada a integrazioni future con sistemi di embedding contestuali e sintesi vocale ad alta fedeltà.

Tabella 2.4: Modelli di tokenizzazione per la lingua Thai

Modello	Anno	Dataset Principale	Caratteristiche principali
`newmm` (Phatthiyaphaibun et al., 2023)	2023	Basato su dizionario `thai_words`	Algoritmo di segmentazione lessicale basato su *maximum matching* e Thai Character Cluster. Utilizza un dizionario ampio e ottimizzato per la lingua thai; non richiede addestramento su corpus. La variante `newmm-safe` introduce controlli aggiuntivi per ridurre l'ambiguità e gestire testi lunghi o rumorosi.
`DeepCut` (Rakpong Kittinaradorn et al., 2019)	2019	BEST2010	Classificazione binaria a livello di carattere. Gestisce parole sconosciute e ambiguità contestuali. Richiede più risorse rispetto ai metodi basati su dizionario.
`AttaCut` (Chormai et al., 2019)	2019	BEST2010, Orchid	Combina efficienza e precisione. Versioni "c" (solo caratteri) e "sc" (include informazioni sillabiche). Inferenza fino a 6 volte più veloce di `DeepCut`.

2.3.3 Sintesi Vocale del Thai

La sintesi vocale per la lingua Thai ha conosciuto progressi significativi grazie all'adozione di modelli neurali avanzati e all'accesso a corpora di parlato sempre più ampi e diversificati. La lingua Thai presenta sfide peculiari: la struttura tonale richiede che ogni sillaba venga pronunciata con il tono corretto, mentre la morfologia e la prosodia influenzano la naturalezza del parlato. L'obiettivo della ricerca è dunque generare un parlato sintetico non solo intelligibile, ma che preservi le caratteristiche prosodiche e tonali del Thai, garantendo un ascolto naturale e vicino alla voce umana. Questo è particolarmente rilevante in applicazioni come assistenti vocali, strumenti didattici, sistemi di dialogo e piattaforme multimediali interattive.

Tra i modelli più utilizzati, **KhanomTan TTS** (Wannaphong, 2022) rappresenta una

soluzione avanzata open-source progettata specificamente per il Thai, con capacità multilingue che includono anche l'inglese. Basato sull'architettura YourTTS, variante avanzata di Coqui-TTS, integra meccanismi di adattamento vocale e gestione multilingua, permettendo di generare parlato naturale e realistico. L'addestramento su corpora come TSync 1 e TSync 2 (Wutiwiwatchai and Hansakunbuntheung, 2004) garantisce varietà prosodica e lessicale sufficiente per una resa fedele del parlato, preservando le caratteristiche tonali fondamentali per la comprensione. Grazie a queste caratteristiche, **KhanomTan TTS** è particolarmente indicato per applicazioni che richiedono alta naturalezza e precisione del parlato, come sistemi educativi, assistenti vocali avanzati e piattaforme multimediali personalizzate.

In parallelo, **gTTS** (Durette, 2025) offre un approccio più generale, basato sull'interfaccia con l'API di Google Translate per la conversione del testo in parlato. Supportando oltre trenta lingue e diversi dialetti, **gTTS** produce file audio in formato MP3 in tempi rapidi, risultando particolarmente utile in applicazioni in cui semplicità, rapidità e facilità di integrazione sono prioritarie, come prototipi di chatbot, strumenti didattici o applicazioni mobili. Tuttavia, la dipendenza da un servizio esterno e la limitata flessibilità nella gestione dei parametri prosodici rendono **gTTS** meno adatto a scenari che richiedono personalizzazione avanzata della voce o fedeltà tonale elevata.

Tabella 2.5: Modelli di sintesi vocale per la lingua Thai

Modello	**Anno**	**Dataset Principale**	**Caratteristiche principali**
`KhanomTan TTS` (Wannaphong, 2022)	2022	TSync 1, TSync 2	Progettato specificamente per la lingua Thai. Gestisce correttamente prosodia e toni, con supporto multilingua (Thai e Inglese). Addestrato su TSync 1 e 2, offre elevata naturalezza nella sintesi vocale.
`gTTS` (Durette, 2025)	2025	N/A	Supporta oltre 30 lingue. Rapido e semplice da integrare. Meno flessibile nella gestione di toni e prosodia e dipendente da un servizio esterno.

2.4 Integrazione dei Modelli nella Pipeline di Elaborazione del Parlato Thai

La realizzazione della pipeline per l'analisi del parlato Thai si è basata sull'integrazione di modelli pre-addestrati e strumenti specifici per il riconoscimento vocale, la tokenizzazione, la rappresentazione semantica e la sintesi vocale, con l'obiettivo di ottenere un sistema capace di gestire la complessità fonetica, tonale e prosodica della lingua Thai. Per la fase di trascrizione automatica (STT), è stato adottato il modello `wav2vec2-large-xlsr-53-th`, un adattamento per la lingua Thai del modello auto-supervisionato Wav2Vec2 (Phatthiyaphaibun et al., 2022). Questo modello combina un approccio self-supervised per l'apprendimento di rappresentazioni audio generali con un fine-tuning supervisionato specifico per la lingua Thai, consentendo di catturare efficacemente le caratteristiche tonali e fonetiche peculiari della lingua.

L'utilizzo di `wav2vec2-large-xlsr-53-th` ha permesso di ottenere trascrizioni accurate anche in presenza di variabilità di accento, ritmo e intonazione tipiche dei parlanti non madrelingua, riducendo significativamente gli errori di riconoscimento fonetico. Le trascrizioni ottenute sono state quindi sottoposte a un processo di iniezione controllata di errori fonetici, generando dati sintetici rappresentativi degli errori più comuni commessi dagli studenti. Tale approccio ha costituito la base per il fine-tuning di un modello ASR capace di trascrivere fedelmente le pronunce degli studenti, senza correggerle automaticamente, preservando quindi gli errori fonetici originali come oggetto di analisi. Parallelamente, il modello è stato utilizzato anche per trascrivere frasi corrette generate tramite il motore di sintesi vocale *gTTS* (Durette, 2025), fornendo un riferimento standardizzato utile per la valutazione della qualità della pronuncia e per il confronto semantico.

Per la tokenizzazione delle trascrizioni Thai, è stato impiegato il modello *newmm* (Phatthiyaphaibun et al., 2023) della libreria `PyThaiNLP`. Questo strumento implementa algoritmi di segmentazione basati su regole linguistiche e statistiche, che tengono conto della morfologia e della fonologia della lingua Thai. La segmentazione coerente delle frasi ha permesso di ridurre la possibilità di confini artificiali tra parole, garantendo una maggiore accuratezza nella costruzione di embedding semantici e nell'analisi del contenuto informativo delle trascrizioni.

L'estrazione della rappresentazione semantica delle frasi è stata affidata al modello *paraphra*

se-multilingual-MiniLM-L12-v2 (Mahmoud et al., 2025), il quale converte le trascrizioni in vettori semantici compatti adatti al calcolo di metriche di similarità tra frasi. L'utilizzo di un

modello multilingue basato su architettura transformer ha permesso di mantenere elevate prestazioni anche in presenza di frasi brevi, incomplete o contenenti errori fonetici, supportando l'individuazione di divergenze concettuali, concetti mancanti o approssimazioni semantiche, e consentendo di generare feedback dettagliato e mirato per gli studenti.

Per la sintesi vocale, sono stati impiegati due modelli complementari. Il motore *gTTS* (Durette, 2025) è stato utilizzato per generare audio delle frasi corrette, fornendo una base di confronto standardizzata. Il modello *KhanomTan TTS* (Wannaphong, 2022), invece, è stato impiegato per creare un corpus vocale con errori di pronuncia controllati. Questo modello sfrutta reti neurali sequenza-a-sequenza in grado di riprodurre con precisione la prosodia, i toni e le inflessioni tipiche della lingua Thai, consentendo la generazione di dataset sintetici realistici senza introdurre correzioni indesiderate. Inoltre, è stato utilizzato il corpus *Lotus* (Sertsi et al., 2016), che non era stato impiegato per l'addestramento di alcuna versione del modello STT, per applicare l'injection degli errori di pronuncia thailandese, garantendo dati indipendenti e realistici.

La scelta dei modelli utilizzati nella pipeline è stata guidata dalla necessità di garantire un equilibrio tra accuratezza fonetica e robustezza semantica. Il modello `wav2vec2-large-xlsr-53-th` offre un riconoscimento altamente accurato, ma richiede dati ben segmentati e pre-processati, mentre la sintesi vocale controllata tramite *KhanomTan TTS* permette di generare errori fonetici realistici, ma necessita di una gestione attenta della prosodia e del tono. L'integrazione con la tokenizzazione avanzata e l'embedding semantico consente di mitigare le limitazioni individuali dei singoli modelli, garantendo una pipeline complessiva capace di affrontare le complessità fonetiche, tonali e semantiche della lingua Thai.

CAPITOLO 3

Automatic Speech Recognition per la lingua thailandese

L'**Automatic Speech Recognition (ASR)**, o riconoscimento automatico del parlato, è una tecnologia che consente di trascrivere in forma testuale segnali vocali prodotti da un parlante umano. L'ASR costituisce un elemento chiave nell'interazione uomo-macchina basata sulla voce, ed è alla base di sistemi quali assistenti virtuali, strumenti di dettatura automatica, sottotitolazione di contenuti audiovisivi.

Lo sviluppo dei sistemi ASR ha seguito una traiettoria evolutiva significativa. I primi approcci, basati su modelli statistici come gli Hidden Markov Models (HMM) (Gales and Young, 2008) e i modelli acustici Gaussian Mixture Models (GMM) (El-Emary and Khafaga, 2015), hanno rappresentato per lungo tempo lo standard, permettendo di modellare sequenze temporali con una discreta efficacia . Successivamente, l'introduzione delle reti neurali profonde ha inaugurato una nuova fase, con l'adozione di architetture ricorrenti (RNN, LSTM) e, più recentemente, di modelli Transformer. Questi ultimi, integrati in sistemi end-to-end, hanno rivoluzionato il settore grazie alla capacità di apprendere direttamente il mapping tra segnali acustici e testo, riducendo la dipendenza da componenti progettati manualmente e migliorando sensibilmente le prestazioni.

Negli ultimi anni, l'integrazione di modelli di deep learning ha permesso di ottenere livelli di accuratezza elevati, soprattutto per lingue ad alta risorsa come l'inglese e il cinese mandarino. L'adozione di approcci end-to-end, combinata con la disponibilità di grandi corpora vocali annotati, ha consentito ai sistemi di apprendere in modo efficace le caratteristiche fonetiche, prosodiche e sintattiche delle lingue, riducendo la necessità di componenti manual-

mente progettati. Tuttavia, persistono sfide significative, legate sia a problemi generali del riconoscimento del parlato sia a caratteristiche specifiche di alcune lingue, come il thailandese.

3.1 Automatic Speech Recognition: Sfide Aperte

Lo sviluppo di sistemi di **Automatic Speech Recognition (ASR)** rappresenta in generale una sfida complessa, dovuta a molteplici fattori di natura sia tecnica sia linguistica. Un aspetto fondamentale riguarda la **variabilità acustica**, ossia le differenze nel segnale vocale generate da fattori intrinseci ed estrinseci. Le differenze tra parlanti, dovute a età, genere, timbro della voce, accenti regionali o inflessioni individuali, si combinano con variazioni legate all'ambiente, come rumori di fondo, riverberi o interferenze, e alla qualità o al tipo di microfono impiegato. Questa variabilità introduce una complessità significativa nella fase di modellazione, poiché i sistemi devono essere in grado di generalizzare efficacemente anche in presenza di condizioni non osservate durante l'addestramento, senza subire perdite rilevanti in termini di accuratezza.

Un'altra difficoltà riguarda il **parlato spontaneo**, caratterizzato da fenomeni quali esitazioni, pause irregolari, autocorrezioni, disfluenze e variazioni di velocità nell'articolazione. A differenza del parlato letto o registrato in ambienti controllati, il parlato spontaneo introduce rumori linguistici e variazioni prosodiche complesse che rendono più difficile la segmentazione corretta dei fonemi e la trascrizione accurata. Gestire tali fenomeni richiede approcci avanzati di modellazione acustica e linguistica, spesso combinati con algoritmi in grado di adattarsi dinamicamente al contesto e alla cadenza del parlante.

Nei contesti **multilingue**, il fenomeno del **code-switching** rappresenta un'ulteriore fonte di complessità. Quando un parlante alterna più lingue all'interno della stessa frase o conversazione, i sistemi ASR devono non solo rilevare quale lingua sia attiva in un determinato istante, ma anche applicare correttamente le regole fonetiche, lessicali e sintattiche specifiche di ciascuna lingua. Questa capacità di adattamento in tempo reale è particolarmente rilevante nelle comunità bilingue o multilingue, dove l'alternanza linguistica è frequente e può coinvolgere anche prestiti lessicali o termini tecnici non comuni.

La gestione delle parole **out-of-vocabulary (OOV)** costituisce un'altra sfida critica. Nomi propri, neologismi, termini tecnici e prestiti linguistici spesso non sono presenti nei vocabolari utilizzati dai modelli ASR, che si basano su statistiche apprese durante l'addestramento. La presenza di OOV può comportare trascrizioni errate o omissioni, compromettendo la comprensibilità del testo generato. Approcci come la modellazione subword, l'uso di rappre-

sentazioni fonetiche e le tecniche G2P (*grapheme-to-phoneme*) sono stati proposti per ridurre tali problematiche, offrendo una copertura lessicale più ampia e flessibile.

Le lingue low-resource rappresentano un ulteriore contesto di criticità nello sviluppo di sistemi di riconoscimento vocale e NLP. La scarsità di corpora ampi e accuratamente annotati limita le capacità di apprendimento dei modelli, rendendo difficile ottenere prestazioni comparabili a quelle delle lingue ad alta risorsa. Per affrontare questa problematica, sono stati sviluppati approcci come il transfer learning (Gupta et al., 2023), in cui modelli pre-addestrati su lingue abbondantemente rappresentate trasferiscono conoscenze fonetiche, sintattiche e semantiche a lingue meno documentate, riducendo la quantità di dati richiesta per l'addestramento e migliorando la generalizzazione del modello. Parallelamente, tecniche di data augmentation (Gupta et al., 2023)sono state adottate per generare varianti artificiali dei dati esistenti, ad esempio mediante modifiche al segnale audio, variazioni di velocità, pitch-shifting o sintesi vocale controllata, al fine di aumentare la diversità e la quantità di esempi disponibili e favorire l'adattamento del modello a scenari reali più complessi. Modelli multilingue end-to-end rappresentano un'ulteriore strategia, poiché consentono di condividere rappresentazioni acustiche e linguistiche tra lingue diverse e di mitigare, almeno in parte, le limitazioni derivanti dalla mancanza di dati, offrendo una soluzione scalabile e robusta per l'elaborazione di lingue a bassa risorsa.

Le **lingue tonali e con prosodia complessa**, come il cinese, il vietnamita o il thailandese, introducono ulteriori sfide. In queste lingue, il significato di una parola dipende non solo dai suoni segmentali (consonanti e vocali), ma anche dal tono con cui vengono pronunciati. Ciò implica che i sistemi ASR debbano integrare informazioni prosodiche e discriminare tra diverse tonalità, al fine di evitare ambiguità fonetiche e semantiche. La corretta modellazione dei toni richiede reti neurali in grado di catturare dipendenze temporali sottili e informazioni di contesto, così da distinguere parole foneticamente simili ma semanticamente differenti.

Infine, un aspetto trasversale riguarda i **bias nei dati di training**. Con questo termine si intende la presenza di una distribuzione non uniforme delle caratteristiche linguistiche all'interno dei dataset utilizzati per l'addestramento. Alcune varietà linguistiche, accenti, fasce d'età, generi o situazioni comunicative possono essere sovra-rappresentate, mentre altre risultano sottorappresentate o del tutto assenti. Di conseguenza, i modelli tendono a funzionare meglio con le tipologie più frequenti, riducendo la robustezza e l'equità del sistema quando si trovano a trascrivere parlanti o situazioni meno comuni. La mitigazione di tali bias è essenziale per garantire prestazioni uniformi, ridurre disparità sociali e assicurare affidabilità nelle applicazioni reali. Un riepilogo critico di queste problematiche e delle strategie adottate

per affrontarle, come l'impiego di modelli multilingue, la segmentazione subword o il bilanciamento dei dati, consente di evidenziare come la ricerca stia progredendo verso sistemi sempre più adattabili, sebbene molte sfide restino ancora aperte.

Le sfide generali presentate, quali la variabilità acustica, la gestione del parlato spontaneo, il code-switching e la scarsità di dati, assumono caratteristiche particolari quando ci si concentra sul thailandese. La natura tonale della lingua, la complessità ortografica e fonetica, la mancanza di spazi tra le parole e la variabilità dialettale rendono più marcata la difficoltà di sviluppare sistemi ASR robusti ed efficaci, evidenziando come le problematiche generali diventino sfide specifiche che richiedono soluzioni dedicate.

3.2 ASR per la lingua thailandese (Sfide Aperte)

Il thailandese presenta caratteristiche linguistiche uniche che rendono il riconoscimento automatico del parlato particolarmente complesso. In primo luogo, si tratta di una **lingua tonale** con cinque toni distintivi: un errore nella trascrizione del tono può alterare radicalmente il significato della parola, rendendo la discriminazione tonale un aspetto cruciale per i sistemi ASR (Wutiwiwatchai and Furui, 2006). Ad esempio, la sillaba *mai* può assumere significati completamente diversi a seconda del tono: *máai* ("legno"), *mài* ("nuovo"), *măi* (particella interrogativa) o *mâi* ("non"). Un riconoscimento scorretto del tono comporta quindi un'interpretazione semantica errata del messaggio. In questo contesto, oltre alla tradizionale metrica del Word Error Rate (WER) [Radford et al. (2022) e Baevski et al. (2020)], la letteratura ha proposto indicatori complementari come il Character Error Rate (CER)[Radford et al. (2022) e Baevski et al. (2020)], più adatto a lingue con scritture complesse, che permettono di misurare in modo più accurato la capacità di un sistema di catturare errori prosodici e differenze di tono.

Un'altra difficoltà deriva dall'**assenza di spazi tra le parole** nella scrittura thai. Poiché le parole non sono separate graficamente, l'ASR deve affidarsi a modelli linguistici avanzati per segmentare correttamente il flusso continuo di fonemi in unità lessicali, senza introdurre errori di nei processi di tokenization (Rakpong Kittinaradorn et al., 2019; Chay-intr and Hidetaka Kamigaito, 2021; Phatthiyaphaibun et al., 2023). Approcci che applicano subword tokenization sono stati adottati per affrontare questo problema, mostrando una maggiore capacità di generalizzazione.

L'**ortografia complessa** del thailandese contribuisce ulteriormente alla sfida: 44 consonanti, numerosi simboli vocalici e segni di tono disposti in diverse posizioni attorno alla

consonante principale rendono difficile l'allineamento grafema-fonema (Bisani and Ney, 2008; Řezáčková et al., 2021). Sistemi G2P specifici per il thai sono stati sviluppati per migliorare l'accuratezza di tale conversione, con l'obiettivo di supportare in maniera più efficace la modellazione acustica e linguistica.

A ciò si aggiunge una **fonetica ricca**, con distinzioni tra vocali lunghe e corte, consonanti aspirate e non aspirate, nonché suoni simili per articolazione, che aumentano la probabilità di errori di trascrizione, soprattutto in condizioni di rumore o con parlanti meno chiari (Wutiwiwatchai and Furui, 2006). Per migliorare la robustezza dei modelli, sono state proposte tecniche di *data augmentation* (Gupta et al., 2023) che includono variazioni artificiali del pitch, della durata e della velocità dell'eloquio.

La **variabilità dialettale** rappresenta un'altra sfida significativa. Oltre al thai standard, esistono numerose varianti regionali che differiscono per pronuncia, intonazione e lessico. I sistemi ASR devono essere capaci di generalizzare tra queste varianti, evitando cali di performance quando il parlante non segue la norma standard (Suwanbandit et al., 2023). A questo si aggiunge il fenomeno del **code-switching con l'inglese**, frequente in ambiti tecnologici, educativi e quotidiani, che richiede ai modelli la capacità di distinguere dinamicamente le lingue e applicare regole fonetiche e lessicali differenti senza compromettere la trascrizione complessiva (Aung et al., 2024).

Un ostacolo cruciale per l'ASR in thailandese è la **scarsità di risorse**. Corpus come *Lotus* (Sertsi et al., 2016), *CommonVoice Thai* (Ardila et al., 2019) e *TALPCo* (Phatthiyaphaibun et al., 2022) hanno contribuito alla ricerca, ma restano limitati in termini di dimensione, varietà di parlanti e diversità dei contesti linguistici coperti. Inoltre, tali dataset sono costituiti quasi esclusivamente da parlato di madrelingua, mentre nella letteratura non risultano disponibili corpora su larga scala dedicati a parlanti non nativi di thai.

La mancanza di **corpora L2**[1] rappresenta una criticità rilevante. Nel corso di questo lavoro di tesi è stato osservato come non vi siano dataset che raccolgano sistematicamente produzioni di studenti di thailandese L2, comprensive di errori fonetici e prosodici caratteristici del parlato non nativo. Questa mancanza condiziona direttamente la progettazione dei modelli ASR: i sistemi esistenti, addestrati prevalentemente su parlato standard, tendono a *normalizzare* o *correggere* le trascrizioni, restituendo output che non riflettono fedelmente le produzioni degli apprendenti. Un fenomeno analogo è stato documentato in studi su ASR per altre lingue, dove si evidenzia come i sistemi ottimizzati sul parlato nativo non riescano a rappresentare adeguatamente le peculiarità fonetiche e prosodiche degli utenti L2. Nel

[1] L2 indica la seconda lingua, ovvero la lingua appresa da parlanti non nativi.

caso del thai, tale problematica appare particolarmente significativa proprio per l'assenza di risorse dedicate.

Negli ultimi anni, i progressi nei modelli di *self-supervised learning* hanno ampliato le possibilità per le lingue a bassa risorsa come il thai. Modelli come **wav2vec 2.0** (Baevski et al., 2020) hanno dimostrato la capacità di apprendere rappresentazioni acustiche robuste senza richiedere grandi quantità di dati annotati, favorendo l'adattabilità a lingue con corpora limitati. L'estensione specifica **Thai Wav2Vec2.0** (Phatthiyaphaibun et al., 2022), addestrata su CommonVoice, ha mostrato miglioramenti significativi nelle metriche WER e CER rispetto a sistemi precedenti basati su HMM o RNN, sebbene la sua efficacia sia ancora vincolata alla disponibilità e qualità dei dati di training. Un altro sviluppo rilevante è costituito dai modelli multilingue di grandi dimensioni, come **Whisper** (Radford et al., 2022), che hanno mostrato robustezza rispetto a rumore, code-switching e variabilità di accenti. Versioni adattate al thai, come **Thonburian Whisper** (Aung et al., 2024), dimostrano che la combinazione di pre-training multilingue e fine-tuning specifico per lingua può ridurre significativamente gli errori tonali e lessicali, aprendo la strada a sistemi ASR più versatili e inclusivi.

In generale, i modelli ASR per il thailandese si articolano tra approcci basati su architetture end-to-end, come wav2vec 2.0 adattato alla lingua, e modelli multilingue di grandi dimensioni come Whisper. Gli approcci end-to-end offrono il vantaggio di ridurre la dipendenza da componenti manuali, come dizionari fonema-grafema, ma restano vincolati alla quantità e alla qualità dei dati disponibili. I modelli multilingue, invece, consentono di trasferire conoscenze da lingue ad alta risorsa, migliorando la robustezza rispetto a rumore, accenti e code-switching, ma possono incontrare difficoltà nell'apprendere caratteristiche tonali e ortografiche peculiari del thailandese senza fine-tuning specifico.

Oltre agli aspetti tecnici, è necessario considerare dimensioni sociolinguistiche ed etiche. La maggior parte dei sistemi è addestrata sul thai standard di Bangkok, con il rischio di marginalizzare le varietà regionali e di generare discriminazioni nei confronti dei parlanti che non aderiscono alla norma. In ambito educativo, inoltre, vi è il pericolo che un ASR eccessivamente normalizzante finisca per occultare gli errori degli studenti, riducendo la sua utilità come strumento di supporto all'apprendimento. Questo rappresenta un caso concreto di **bias nei dati di addestramento**, fenomeno già discusso a livello generale (Radford et al., 2022), che nel thai si manifesta in maniera particolarmente marcata.

3.3 Motivazioni e Approccio Proposto

Le problematiche discusse evidenziano la necessità di sviluppare soluzioni mirate per l'**Automatic Speech Recognition** della lingua thailandese, in grado di affrontare simultaneamente le sfide derivanti dalla natura tonale, dalla complessità ortografica e dalla scarsità di dati annotati, in particolare per il parlato non nativo. L'assenza di corpora L2 rappresenta un limite rilevante: i sistemi ASR esistenti, addestrati quasi esclusivamente su parlato di madrelingua, tendono a normalizzare o correggere automaticamente le produzioni deviate, impedendo una rappresentazione accurata degli errori fonetici e prosodici tipici degli apprendenti.

In tale contesto, il presente lavoro propone un approccio basato sulla **creazione di un corpus vocale dedicato al parlato non nativo** e sul successivo **fine-tuning di un modello STT** (Speech-to-Text)(Phatthiyaphaibun et al., 2022). Il corpus viene generato a partire da trascrizioni thailandesi modificate mediante un sistema di *iniezione controllata di errori sul parlato*, che simula in modo realistico le produzioni L2, includendo variazioni su consonanti, vocali e toni. Tali dati sintetici consentono di addestrare un modello STT affinato per riconoscere le pronunce deviate senza applicare correzioni automatiche, preservando così la fedeltà rispetto al parlato effettivo degli apprendenti.

L'approccio prevede inoltre la definizione di una **pipeline completa di elaborazione**, che comprende la generazione automatica degli audio, la gestione delle trascrizioni foneticheche alterate, il processo di addestramento supervisionato e la valutazione delle prestazioni mediante metriche specifiche come il CER (Baevski et al., 2020) . Tale metodologia mira a colmare la mancanza di risorse per il thai L2 e a costituire una base per lo sviluppo di sistemi ASR più inclusivi, utili non solo per la trascrizione del parlato non nativo, ma anche come strumenti di supporto all'apprendimento della pronuncia.

CAPITOLO 4

Fine-Tuning di un Modello ASR

Il presente capitolo descrive il processo di addestramento di un modello di **Automatic Speech Recognition (ASR)** per la lingua Thai, progettato per trascrivere il parlato di studenti L2 contenente errori fonetici tipici. A differenza dei modelli ASR convenzionali, che tendono a correggere automaticamente le deviazioni dalla pronuncia standard, il nostro modello è stato concepito per **preservare fedelmente gli errori di pronuncia**, fornendo trascrizioni realistiche del parlato imperfetto.

Questo approccio presenta vantaggi significativi sia in ambito didattico sia linguistico: consente di analizzare le difficoltà fonetiche degli studenti, valutare la loro competenza e identificare schemi di errore ricorrenti. In particolare, nella lingua Thai, preservare gli errori è fondamentale, poiché la variazione di un solo tono può alterare completamente il significato di una frase. Disporre di trascrizioni che mantengono tali deviazioni rende possibile studiarne l'impatto semantico e progettare sistemi di feedback automatico più efficaci.

Il modello scelto per il fine-tuning è `airesearch/wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al., 2022), una variante di Wav2Vec2 pre-addestrata su dati Thai, capace di gestire sequenze audio di lunghezza variabile e di adattarsi alle produzioni linguistiche non native, permettendo al modello di trascrivere fedelmente le pronunce imperfette.

4.1 Creazione di dataset di trascrizione fedele

La costruzione di un dataset mirato alla trascrizione fedele del parlato L2 rappresenta un passaggio cruciale per l'addestramento di modelli ASR destinati ad ambienti educativi e di analisi linguistica. I sistemi ASR tradizionali, infatti, sono generalmente addestrati su parlato nativo e tendono a interpretare il segnale audio in base a forme canoniche della lingua. Questo approccio, sebbene massimizzi l'accuratezza rispetto a metriche convenzionali, introduce un bias nelle trascrizioni, eliminando deviazioni fonetiche e prosodiche tipiche degli studenti. Tale fenomeno può compromettere la validità delle analisi di errore o la capacità del modello di riconoscere pronunce non standard, risultando in una rappresentazione incompleta o distorta del parlato L2.

Per superare questa limitazione, è stata sviluppata una procedura automatica di generazione di trascrizioni fedeli al parlato degli studenti. Poiché non era disponibile la collaborazione di parlanti madrelingua, le trascrizioni non sono state annotate manualmente; invece, sono stati inseriti automaticamente errori fonetici simulati, basati su studi linguistici e osservazioni degli errori più comuni commessi dagli studenti non madrelingua thailandese. In questo modo, il dataset rappresenta il parlato L2, includendo fenomeni quali omissioni di suoni, sostituzioni consonantiche, riduzioni vocaliche, modifiche della durata delle vocali e variazioni tonali.

Come base di partenza è stato utilizzato il dataset **LOTUS** (Sertsi et al., 2016), costituito da registrazioni di parlanti madrelingua in contesti diversificati (Clean e Office), selezionato per la ricchezza dei fenomeni tonali e fonetici presenti. Il dataset finale, derivante dall'integrazione di trascrizioni originali e iniezioni automatiche di errori fonetici, rappresenta dunque le caratteristiche del parlato L2. Grazie a questo approccio, il modello ASR può essere addestrato su trascrizioni realistiche, migliorando la capacità di riconoscere pronunce non native e supportando analisi pedagogiche, valutazioni della competenza linguistica e studi di linguistica applicata.

4.2 Injection di errori

Per facilitare la comprensione del flusso operativo della pipeline, la Figura 4.1 fornisce una rappresentazione grafica dell'intero processo, mostrando le fasi principali dalla trascrizione automatica alla generazione del dataset sintetico e all'addestramento del modello fine-tuned. Questo schema permette di avere una visione immediata della struttura complessiva, agevolando la lettura dei dettagli dei singoli step presentati in seguito.

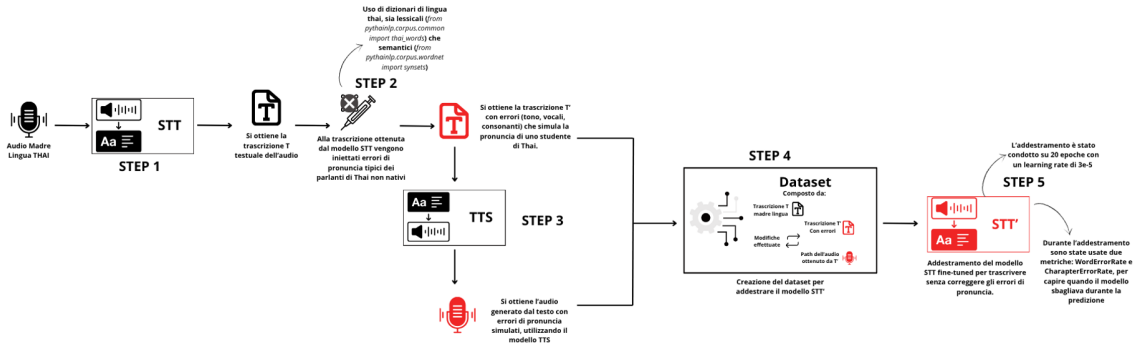

Figura 4.1: Schema generale della pipeline per la generazione del dataset di pronuncia sintetica con errori di pronuncia.

La costruzione del dataset sintetico con errori di pronuncia si fonda su una pipeline operativa articolata in più fasi sequenziali. Tale pipeline consente di trasformare un corpus di parlato nativo in un insieme di dati alterati foneticamente, mantenendo la tracciabilità di ogni modifica e la coerenza tra rappresentazione testuale e acustica. Il processo parte dai file audio originali e impiega un modello ASR per la trascrizione, un modulo di iniezione di errori per la simulazione del parlato L2, un modello TTS per la sintesi vocale e infine una fase di salvataggio e addestramento del modello fine-tuned.

- **Step 1: Trascrizione automatica (ASR).** La pipeline inizia con il caricamento e la normalizzazione dei file audio del corpus di riferimento. Ogni campione viene processato dal modello `airesearch/wav2vec2-large-xlsr-53-th`, una variante di Wav2Vec2 ottimizzata per la lingua thailandese. Questo modello è stato selezionato per garantire un equilibrio tra accuratezza e fedeltà, producendo trascrizioni pulite senza eccessiva normalizzazione ortografica, preservando eventuali fenomeni linguistici naturali e la variabilità prosodica del parlato. Ad esempio, partendo da un file audio del parlato originale, otteniamo la seguente trascrizione corretta:

  มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ

  Questa rappresenta la versione priva di errori, cioè la pronuncia corretta prodotta da un parlante madrelingua.

  Prima di procedere con l'injection di errori, la trascrizione viene segmentata mediante il tokenizer `newmm` (Phatthiyaphaibun et al., 2023), che suddivide correttamente la frase in unità linguistiche coerenti con la fonologia thai. Questa tokenizzazione è indispensabile

per consentire alla fase successiva di identificare correttamente i punti in cui applicare le modifiche fonetiche.

- **Step 2: Injection controllata di errori fonetici.** Una volta ottenuta la sequenza tokenizzata, il sistema applica la funzione `maybe_inject_pronunciation_multi_scaled`, responsabile dell'introduzione controllata di errori di pronuncia. L'algoritmo opera sulla trascrizione testuale, simulando le principali difficoltà degli studenti L2. Le alterazioni possono riguardare le consonanti iniziali, le vocali, le consonanti finali o i toni, e vengono decise in modo casuale ma regolato da probabilità scalate, distanza minima tra token modificati e parametri di severità (light o heavy). La funzione genera varianti fonetiche plausibili per ogni parola, filtrando quelle lessicalmente valide o semanticamente coerenti tramite un dizionario di parole thailandesi e il WordNet Thai. In questo modo, ogni parola può essere modificata in maniera realistica, producendo sostituzioni consonantiche, vocaliche o tonali coerenti con gli errori osservati in parlanti non nativi. Al termine dell'elaborazione, la funzione restituisce la frase modificata, un log dettagliato delle trasformazioni effettuate e un flag che indica la presenza o meno di alterazioni.

Il funzionamento generale dell'algoritmo è descritto nel seguente pseudocodice:

Algorithm 1: maybe_inject_pronunciation_multi_scaled (parte 1)

Input: frase, frase_riferimento, probabilità_base, severità, distanza_minima

Output: frase_modificata, log_cambiamenti, flag_modifiche

1 Dividi la frase e la frase_riferimento in token;

2 Inizializza lista_parole_modificate, log_cambiamenti, flag_modifiche = False;

3 Imposta indice_ultima_parola_modificata = $-distanza\_minima - 1$;

4 Calcola probabilità scalata di modifica (massimo 50%);

Algorithm 1: (continua) maybe_inject_pronunciation_multi_scaled (parte 2)

```
for i ← 0 to len(parole_originali) − 1 do
    if i − indice_ultima_parola_modificata ≤ distanza_minima then
        mantieni token e aggiorna log;
    else
        determina livello di severità (light o heavy);
        con probabilità = probabilità scalata decidi se modificare il token;
        if il token viene modificato then
            genera varianti fonetiche plausibili e filtra quelle valide;
            if esiste almeno una variante valida then
                sostituisci token con variante scelta;
                aggiorna log e imposta flag_modifiche = True;
                indice_ultima_parola_modificata = i;
        aggiungi token (modificato o originale) alla lista_parole_modificate;
Riallinea la frase_modificata con la frase_riferimento;
return frase_modificata, log_cambiamenti, flag_modifiche;
```

La funzione genera varianti fonetiche plausibili. Ad esempio, partendo dalla frase corretta:

> มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ

l'algoritmo può produrre la versione alterata:

> มี ดอกเตอร์ โดม ว โลตม สิบ กอ คอดิต ต่อ เล้า ดอกเตอร์ นิติ ยา กาอน วัน เป็น ประธาน โครงการ

Le modifiche principali introdotte sono:

- **SUONO_VOCALE** – อุดม: 'อ' → 'โ';
- **Rimozione vocale** – Eliminato il simbolo '◌ุ' da อุดม;
- **SUONO_VOCALE** – และ: 'แ' → 'เ';
- **SUONO_VOCALE** – และ: 'ะ' → 'า'.

- **Step 3: Rigenerazione acustica (TTS).** Completata l'iniezione degli errori, la pipeline procede alla rigenerazione dell'audio corrispondente alla trascrizione alterata. Questa operazione viene effettuata mediante il modello di sintesi vocale `khanomtan`, specificamente addestrato per la lingua thai. La scelta di questo modello si basa sulla sua capacità di rispettare fedelmente l'input testuale, riducendo al minimo la normalizzazione automatica. L'audio sintetico prodotto riproduce accuratamente le alterazioni fonetiche introdotte, consentendo la creazione di un corpus di pronuncia L2 coerente dal punto di vista acustico e fonetico. Il risultato è un segnale vocale realistico che conserva le caratteristiche prosodiche e fonetiche alterate della frase di input.

  Ad esempio, consideriamo la seguente sequenza:

  1. **Trascrizione originale (dal corpus Lotus):** มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ
  2. **Trascrizione alterata dopo l'iniezione degli errori:** มี ดอกเตอร์ โดม ว โลตม สิบ กอ คอดิต ต่อ เล้า ดอกเตอร์ นิติ ยา กาอน วัน เป็น ประธาน โครงการ

  La seconda frase viene quindi fornita al modello `khanomtan` per la generazione dell'audio corrotto. L'output risultante è un segnale vocale sintetico che contiene pronunce modificate come la sostituzione del suono iniziale อ con โ in อุดม, la rimozione della vocale ุ, e l'allungamento vocalico in และ, trasformato in เล้า. Queste alterazioni riproducono fedelmente errori comuni tra parlanti non madrelingua, garantendo un dataset acusticamente coerente con le deviazioni fonetiche simulate.

- **Step 4: Archiviazione dei dati sintetici.** Una volta generato l'audio corrotto, la pipeline analizza il log delle modifiche prodotto durante la fase di iniezione. Se il flag di alterazione risulta negativo, viene registrata l'assenza di modifiche. In caso contrario, il sistema elabora una descrizione testuale dettagliata delle trasformazioni fonetiche applicate, specificando la natura degli errori introdotti e le corrispondenze tra i token originali e quelli modificati. Questa fase consente di mantenere la completa tracciabilità del processo e di associare a ogni campione un insieme di metadati descrittivi, utili per l'analisi linguistica e per l'addestramento supervisionato.

  Concluse le fasi di generazione e analisi, la pipeline registra i risultati in un formato strutturato. Ogni entry del dataset include il nome del file, la trascrizione originale, la versione modificata, il flag di alterazione, i dettagli delle modifiche e il percorso dell'audio sintetico generato.

Un esempio concreto di registrazione dei dati è il seguente:

- **Trascrizione originale:** มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ
- **Trascrizione errata:** มี ดอกเตอร์ โดม ว โลตม สิบ กอ คอดิต ต่อ เล้า ดอกเตอร์ นิติ ยา กาอน วัน เป็น ประธาน โครงการ
- **Dettagli delle modifiche:** `SUONO_VOCALE` – อุดม: 'อ' → 'โ'; Rimozione di '◌ุ' da อุดม; `SUONO_VOCALE` – และ: 'แ' → 'เ'; `SUONO_VOCALE` – และ: 'ะ' → 'า'.
- **Percorso dell'audio corrotto**

Il dataset finale è composto da 3.255 frasi. Gli errori sono stati introdotti secondo due criteri principali: la lunghezza della frase, che aumenta la probabilità di contenere più modifiche, e un parametro di probabilità che regola la possibilità di inserire un errore in una determinata posizione. Le tipologie di errori simulate appartengono a tre categorie principali: tono, consonanti e vocali. Gli errori tonali consistono nella sostituzione di uno dei cinque toni della lingua Thai, con conseguenze spesso significative sul significato della frase.

La quantità di errori per frase è variabile: alcune frasi rimangono integre, altre presentano uno o due errori, mentre nei casi più complessi si possono riscontrare fino a tredici modifiche. Le tipologie di errori possono inoltre combinarsi tra loro, producendo frasi con più alterazioni contemporanee, come una sostituzione di tono combinata a un errore consonantico o un errore consonantico unito a una modifica vocalica. Questa varietà rende il dataset una risorsa ricca e realistica, in grado di riprodurre le principali difficoltà fonetiche incontrate dagli studenti di Thai come L2. Pur essendo controllato nei parametri di generazione, il corpus sintetico consente di allenare e valutare modelli ASR capaci di gestire le deviazioni dalla pronuncia standard.

- **Step 5: Addestramento del modello fine-tuned.** L'ultima fase della pipeline riguarda l'utilizzo del dataset sintetico per l'addestramento di un modello di riconoscimento vocale fine-tuned. Prima di procedere all'addestramento, è stata eseguita una fase di preprocessing accurata dei dati, fondamentale per garantire che audio e trascrizioni fossero correttamente formattati e compatibili con il modello Wav2Vec2. Il dataset originale era organizzato in un file CSV contenente i percorsi dei file audio, generati tramite sintesi vocale, e le trascrizioni corrispondenti, comprendenti sia la versione corretta della frase sia una versione modificata con errori fonetici simulati. Durante l'addestramento, sono

state utilizzate esclusivamente le trascrizioni contenenti errori come target, con l'obiettivo di addestrare il modello a riprodurre fedelmente il parlato con le caratteristiche di pronuncia tipiche di parlanti non nativi, senza tentare di correggerlo.

Il dataset preprocessato è stato suddiviso in tre sottoinsiemi distinti: l'80% per l'addestramento, il 10% per la validazione e il 10% per il test. Tale ripartizione ha permesso di monitorare in maniera efficace le prestazioni del modello su dati non visti, riducendo il rischio di overfitting. Particolare attenzione è stata rivolta a mantenere la rappresentatività dei diversi tipi di errori fonetici, assicurando che ogni sottoinsieme includesse una distribuzione bilanciata di errori di consonanti iniziali, vocali, consonanti finali e toni. Questo approccio ha garantito che il modello fosse esposto a una varietà ampia di fenomeni fonetici, migliorando la sua capacità di generalizzazione su parlato non nativo.

Per quanto riguarda l'elaborazione dell'audio, tutte le registrazioni sono state convertite in mono a 16 kHz e normalizzate per uniformare l'ampiezza dei segnali. I silenzi iniziali e finali sono stati rimossi mediante algoritmi di *silence trimming* (Baevski et al., 2020), riducendo rumore inutile e migliorando la qualità delle sequenze audio. Successivamente, tramite il processor di Wav2Vec2, le registrazioni sono state trasformate in sequenze numeriche (`input_values`) utilizzabili direttamente dal modello. Questo passaggio ha consentito di rappresentare il parlato come sequenze di feature acustiche standardizzate, preservando le informazioni fonetiche rilevanti per il riconoscimento delle pronunce errate. Parallelamente, le trascrizioni testuali sono state tokenizzate (Phatthiyaphaibun et al., 2023) utilizzando un vocabolario adatto alla lingua thai e convertite in sequenze di etichette numeriche, con padding gestito in modo da garantire la compatibilità tra sequenze di lunghezza diversa durante le operazioni di *batching* (Baevski et al., 2020). Tale impostazione ha permesso un addestramento efficiente e stabile, con la possibilità di monitorare sia la *training loss* sia la *validation loss* per valutare l'apprendimento del modello su dati di addestramento e su dati non visti.

- Il dataset è stato, poi sottoposto a una fase di validazione finale, finalizzata a verificare la coerenza interna dei dati. Sono stati controllati la corrispondenza tra file audio e trascrizioni, la correttezza delle lunghezze delle sequenze, la compatibilità con le operazioni di batching e collation, nonché l'integrità delle etichette numeriche. Questa verifica ha assicurato che il modello ricevesse dati consistenti, rappresentativi del parlato non nativo e pronti per il fine-tuning, riducendo il rischio di errori durante l'addestramento e garantendo una base solida per la generazione di trascrizioni accurate anche in pre-

senza di pronunce foneticamente alterate. Particolare attenzione è stata inoltre posta alla qualità fonetica dei dati, verificando che le variazioni introdotte simulassero realisticamente errori di parlanti non nativi, coerenti con i pattern fonetici della lingua thailandese.

Il modello di partenza impiegato per il fine-tuning è `airesearch/wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al., 2022), pre-addestrato su dati Thai e particolarmente adatto a catturare la struttura fonetica e intonativa della lingua. L'obiettivo del fine-tuning era adattarlo a trascrivere fedelmente il parlato L2, mantenendo gli errori di pronuncia. Durante l'addestramento, i valori numerici di input sono stati generati dal processor e, per gestire batch contenenti sequenze di lunghezza variabile, è stato implementato un sistema di padding dinamico tramite una *data collator* personalizzata, che ignorava i token di padding durante il calcolo della loss. Le trascrizioni sono state tokenizzate (Phatthiyaphaibun et al., 2023) in maniera *batched* per migliorare l'efficienza, saltando il processo qualora fossero già presenti come ID numerici.

L'addestramento è stato eseguito tramite la classe `Trainer` di HuggingFace (Baevski et al., 2020) (Phatthiyaphaibun et al., 2022), con iperparametri scelti per bilanciare prestazioni e risorse computazionali: *learning rate* pari a 3×10^{-5}, batch size di quattro esempi per dispositivo, un massimo di venti epoche e un criterio di *early stopping* che interrompeva l'addestramento dopo cinque epoche senza miglioramenti nel CER di validazione. È stato abilitato il *gradient checkpointing* per ridurre il consumo di memoria e impostato il salvataggio automatico del modello con il CER più basso sul validation set.

La valutazione delle prestazioni è stata effettuata tramite il *Character Error Rate* (CER) (Baevski et al., 2020).

- Durante l'addestramento, le predizioni venivano decodificate dai logit e confrontate con le trascrizioni di riferimento, escludendo i token di padding. L'andamento delle curve di training e validation loss, riportato nella Figura 4.2, mostra una decrescita regolare della training loss con l'aumentare degli step e una rapida stabilizzazione della validation loss su valori bassi e relativamente costanti, indicando una buona capacità di generalizzazione e l'assenza di overfitting. Al termine del processo, i pesi del modello e il processor sono stati salvati, garantendo riproducibilità e riutilizzo nelle fasi successive. Prima dell'addestramento completo, è stato inoltre condotto uno *smoke test* su un mini-

batch per verificare le dimensioni dei tensori e il corretto funzionamento del forward pass, riducendo il rischio di errori durante l'addestramento vero e proprio.

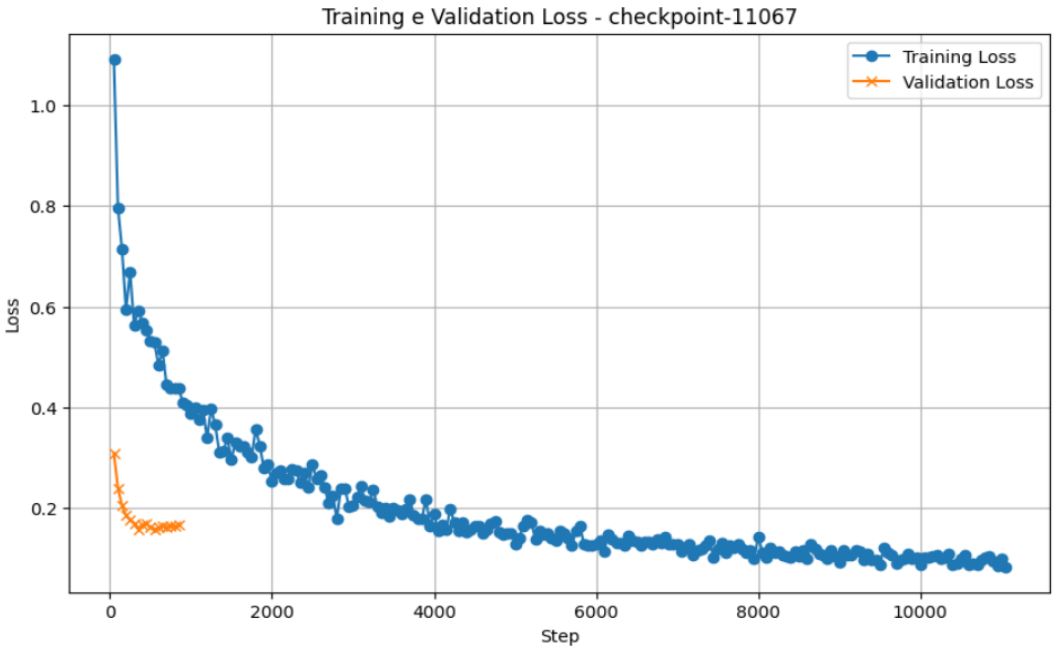

Figura 4.2: Andamento della training loss e della validation loss durante il fine-tuning del modello.

Nel suo insieme, la pipeline realizza un flusso completo che, a partire da registrazioni native, produce un corpus sintetico di parlato alterato e ne utilizza i risultati per l'addestramento di un modello adattivo. Ogni passaggio è progettato per garantire coerenza, tracciabilità e realismo, consentendo la generazione di un dataset riproducibile, utile per lo studio del parlato L2 e per il miglioramento dei modelli di riconoscimento vocale in contesti multilingue.

4.3 Verso lo Scenario Applicativo

L'intero processo descritto in questo capitolo definisce le basi per la realizzazione di un sistema ASR in grado di riconoscere e trascrivere fedelmente il parlato non nativo in lingua Thai, mantenendo le caratteristiche fonetiche che lo differenziano dal parlato standard. Tale risultato non rappresenta un obiettivo fine a sé stesso, ma costituisce il presupposto per un'applicazione concreta in contesti di apprendimento linguistico assistito.

Il modello fine-tuned, infatti, viene impiegato all'interno di un più ampio scenario sperimentale, in cui il parlato degli studenti viene analizzato sia in termini di pronuncia che contenuti semantici. Questo scenario, presentato nel capitolo successivo, esplora un caso di

studio realistico denominato *"Sara al Mercato Locale"*, concepito per valutare la capacità del sistema di interpretare produzioni linguistiche spontanee e fornire un feedback personalizzato.

In tale contesto, il modello ASR costituisce il primo modulo della pipeline di analisi del parlato, seguito da componenti di elaborazione semantica e generazione automatica di suggerimenti correttivi. L'obiettivo finale è la creazione di un sistema di supporto didattico intelligente, capace di analizzare la pronuncia, comprendere il contenuto del messaggio e generare risposte mirate per migliorare l'apprendimento della lingua Thai come L2.

CAPITOLO 5

Caso di Studio: Analisi Semantica delle Produzioni Linguistiche in Tailandese

In questo capitolo viene presentato un caso di studio focalizzato sull'analisi delle produzioni linguistiche degli studenti non madrelingua in lingua tailandese. L'obiettivo principale è sviluppare uno strumento di supporto all'apprendimento linguistico, capace di valutare la coerenza semantica delle frasi pronunciate e la correttezza della pronuncia, fornendo un feedback immediato, personalizzato e interpretabile.

Il sistema utilizza modelli di *sentence embedding* (Mahmoud et al., 2025) per rappresentare le frasi come vettori in uno spazio semantico continuo. Questo consente di misurare la similarità tra la produzione dello studente e una versione di riferimento corretta, identificando discrepanze concettuali, omissioni o errori lessicali, anche in presenza di variazioni sintattiche o fonetiche minori. L'integrazione con l'analisi fonetica e tonale permette di generare report completi, combinando informazioni quantitative (punteggi di similarità, gravità degli errori) e qualitative (suggerimenti fonetici e lessicali).

5.1 Scenario di Analisi

Il sistema è progettato per studenti e utenti non madrelingua con livelli di competenza compresi tra principiante e intermedio. A differenza degli strumenti tradizionali basati su esercizi predeterminati, l'approccio adottato è flessibile e adattivo: le frasi pronunciate vengono valutate rispetto a una versione di riferimento corretta sia semantica che fonetica-

mente, stimando quanto il messaggio mantenga coerenza e comprensibilità rispetto all'intento comunicativo.

Il sistema supporta sia la produzione autonoma di frasi sia il perfezionamento della pronuncia in contesti reali o simulati. Agendo come assistente linguistico interattivo, fornisce un riscontro immediato e mirato, promuovendo un apprendimento attivo, riflessivo e graduale, in cui l'utente può monitorare i progressi e adattare le proprie strategie di studio.

Le registrazioni analizzate possono provenire da ambienti eterogenei, caratterizzati da rumore, variazioni di velocità o differenze di accento. Per garantire robustezza, il sistema integra tecniche di normalizzazione acustica, filtraggio del rumore e trascrizione automatica basata su modelli ASR robusti agli errori fonetici tipici del parlato L2. La trascrizione dello studente viene poi confrontata con la versione di riferimento tramite *sentence embeddings* (Mahmoud et al., 2025) e metriche di similarità, evidenziando divergenze concettuali, omissioni, sostituzioni lessicali o alterazioni sintattiche rilevanti. La misura di similarità viene convertita in una classificazione della gravità degli errori su quattro livelli: *nessuno*, *lieve*, *medio* e *grave*, consentendo la generazione di un feedback chiaro e personalizzato.

5.2 Analisi del Parlato e Generazione del Feedback

La Figura 5.1 illustra la pipeline completa, articolata in fasi che vanno dalla trascrizione automatica del parlato alla produzione del feedback linguistico, combinando modelli di riconoscimento vocale, sintesi vocale, analisi semantica e fonetica per una valutazione coerente e riproducibile.

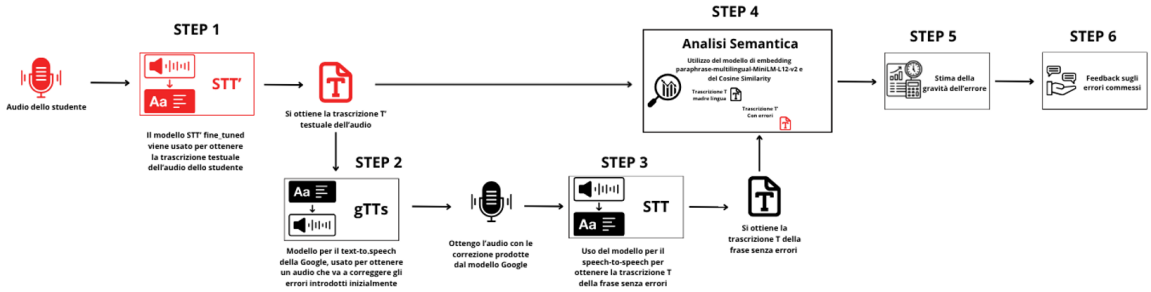

Figura 5.1: Schema generale della pipeline per l'analisi del parlato e la generazione del feedback.

Funzionalità principali della pipeline:

- **Trascrizione automatica del parlato (ASR):** acquisizione e normalizzazione dell'audio tramite modello `wav2vec2` fine-tuned per pronunce con errori fonetici; produce trascrizioni accurate senza correzioni automatiche.

- **Generazione dell'audio di riferimento (TTS):** creazione di un modello standard di pronuncia mediante sintesi vocale di alta qualità, utile per confronti fonetici e tonali.
- **Trascrizione dell'audio di riferimento (ASR di riferimento):** produzione della versione testuale standardizzata basata su parlato nativo, riducendo falsi positivi nelle valutazioni fonetiche e semantiche.
- **Analisi semantica:** calcolo della similarità tramite *sentence embeddings* multilingue a livello di frase e di parola, identificando discrepanze concettuali, omissioni o sostituzioni lessicali.
- **Classificazione della gravità degli errori:** categorizzazione in quattro livelli (*nessuno*, *lieve*, *medio*, *grave*) per rendere il feedback interpretabile e adattabile al livello dell'utente.
- **Analisi fonetica e tonale:** confronto parola per parola, identificazione di omissioni, aggiunte, errori di tono e differenze fonetiche mediante tokenizzazione (Phatthiyaphaibun et al., 2023) e strumenti di confronto fonetico; produzione di suggerimenti mirati.
- **Generazione del feedback linguistico:** integrazione dei risultati semantici e fonetici in un report completo, con punteggi, suggerimenti pratici e indicazioni operative per il miglioramento progressivo delle competenze.

Questa architettura modulare consente non solo di valutare accuratamente la produzione linguistica dello studente, ma anche di tracciare lo storico delle prestazioni e di adattare i parametri di analisi in base al livello e alle esigenze dell'utente, supportando così un apprendimento personalizzato e continuo.

- **Step 1: Trascrizione automatica del parlato (ASR).**

  Il primo step consiste nella trascrizione automatica del parlato prodotto dallo studente mediante un modello di riconoscimento vocale fine-tuned basato su `wav2vec2` (Baevski et al. (2020), Phatthiyaphaibun et al. (2022)). Il modello è stato adattato per gestire pronunce contenenti errori fonetici, consentendo di ottenere trascrizioni robuste anche in presenza di deviazioni tipiche del parlato L2. L'input è costituito da file audio in formato `.wav` o `.mp3`, che vengono convertiti in segnale mono a 16 kHz per uniformare il preprocessing. Questa fase viene realizzata tramite la funzione `prepara_audio()`, che utilizza la libreria `torchaudio` per il caricamento, la normalizzazione e l'eventuale trimming dei silenzi. La trascrizione prodotta rappresenta la base di confronto per

tutte le analisi successive, permettendo di preservare fedelmente gli errori di pronuncia presenti nel parlato dello studente, senza correzioni automatiche.

- **Step 2: Generazione dell'audio di riferimento (TTS).**

  Il secondo step prevede la creazione di un audio di riferimento corretto a partire dalla trascrizione dello studente, mediante un sistema di sintesi vocale basato su `gTTS` (Durette (2025)). Questo audio funge da modello standard di pronuncia e costituisce il riferimento acustico per la valutazione fonetica e tonale. L'adozione di un TTS di alta qualità garantisce una produzione vocale chiara, con intonazione e durata coerenti, indispensabile per confronti accurati tra la voce dello studente e quella corretta. Il file sintetizzato viene salvato e successivamente trascritto da un modello ASR di riferimento per ottenere una versione testuale standardizzata, utile nelle fasi di analisi semantica e fonetica.

- **Step 3: Trascrizione dell'audio generato (ASR di riferimento).**

  Nel terzo step, l'audio sintetizzato viene elaborato da un secondo modello ASR, `wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al. (2022), Aung et al. (2024)), addestrato su parlato nativo thailandese. La trascrizione prodotta rappresenta la versione di riferimento ("target") con cui confrontare la produzione dello studente. Questa fase garantisce che le analisi successive si basino su dati consistenti, privi di errori e rispecchianti la pronuncia madrelingua, riducendo il rischio di falsi positivi nella valutazione fonetica e semantica.

- **Step 4: Analisi semantica.**

  La quarta fase consiste nella valutazione della similarità semantica tra la trascrizione dello studente e quella corretta. Entrambe le frasi vengono convertite in vettori semantici mediante il modello multilingue `paraphrase-multilingual-MiniLM-L12-v2` (Mahmoud et al. (2025)), basato su `SentenceTransformer`. La similarità è calcolata tramite la *cosine similarity* (Mahmoud et al., 2025), con valori compresi tra 0 e 1. Per migliorare la granularità, il sistema effettua un confronto a livello di parola tramite `word_level_similarity()`, che segmenta le frasi con `PyThaiNLP` (Phatthiyaphaibun et al., 2023) e individua la similarità massima tra token corrispondenti. Il punteggio finale, calcolato come media delle similarità reciproche, rappresenta una misura equilibrata della coerenza semantica, identificando differenze concettuali, omissioni o sostituzioni lessicali.

- **Step 5: Classificazione della gravità degli errori.**

  In questa fase, il punteggio di similarità semantica viene tradotto in una categoria qualitativa di gravità dell'errore tramite la funzione `gravita_semantic_similarity()`. Le soglie predefinite distinguono tra assenza di errori, errori lievi, medi o gravi, permettendo di rappresentare in modo chiaro la qualità del parlato dello studente. Questa classificazione consente di rendere il feedback interpretabile, facilitando l'individuazione delle aree di miglioramento. Le etichette vengono registrate insieme ai punteggi di similarità e al feedback generato nel file `risultati.csv`, e le soglie possono essere adattate in base al livello degli studenti o al contesto didattico.

- **Step 6: Analisi fonetica e tonale.**

  Il sesto passaggio approfondisce le differenze a livello fonetico e tonale. La funzione `analizza_differenze_fonetiche()` confronta le frasi parola per parola, utilizzando la tokenizzazione di `PyThaiNLP` (Phatthiyaphaibun et al., 2023). Vengono identificate omissioni, aggiunte, errori di tono attraverso i caratteri tonali Unicode mappati in categorie (alto, basso, discendente, ascendente, neutro) e differenze fonetiche mediante la romanizzazione e il confronto con `SequenceMatcher`. Il sistema genera suggerimenti dettagliati, come "Hai sbagliato il tono nella parola X" o "Hai omesso la parola Y", fornendo indicazioni precise sulle aree fonetiche da correggere. Queste informazioni costituiscono il nucleo del feedback personalizzato, integrando dettagli semantici e fonetici.

- **Step 7: Generazione del feedback linguistico.**

  L'ultima fase unisce i risultati delle analisi semantiche e fonetiche per produrre un feedback finale completo. La funzione `genera_feedback()` combina punteggio di similarità, gravità degli errori e suggerimenti fonetici per creare un messaggio descrittivo, chiaro e mirato. Il feedback finale fornisce indicazioni operative per migliorare la pronuncia, correggere omissioni o errori tonali e consolidare la comprensione semantica. Tutte le informazioni vengono registrate nel file di output, permettendo allo studente di ricevere una valutazione completa e personalizzata.

  Ad esempio, consideriamo la frase del corpus Lotus utilizzata nello Step 1 per la creazione del dataset "มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ":

– **Caso con errori fonetici e semantici:** La frase มี ดอกเตอร์ โดม ว โร ตม สิ ก ค ดิ ต เล้า ดอกเตอร์ นิติ ยา กา ญจน วัน เป็น ประธาน โครงการ è molto simile a มี ดอกเตอร์ โนม ว โลตม สิบ กอ คอดิต ต่อ เล้า ดอก เตอร์ นิติ ยา กาอน วัน เป็น ประธาน โครงการ (0.91). Solo piccole variazioni, ma il significato resta chiaro.

Suggerimenti:

* Migliora la pronuncia di โดม, dovrebbe essere โน
* Migliora la pronuncia di ว, dovrebbe essere ม
* Migliora la pronuncia di โร, dovrebbe essere ว
* Migliora la pronuncia di ตม, dovrebbe essere โล
* Migliora la pronuncia di สิ, dovrebbe essere ตม
* Migliora la pronuncia di ก, dovrebbe essere สิบ
* Migliora la pronuncia di ค, dovrebbe essere กอ
* Migliora la pronuncia di ดิ, dovrebbe essere คอ
* Migliora la pronuncia di ต, dovrebbe essere ดิ
* Migliora la pronuncia di เล้า, dovrebbe essere ต
* Migliora la pronuncia di ดอกเตอร์, dovrebbe essere ต่อ
* Migliora la pronuncia di นิติ, dovrebbe essere เล้า
* Migliora la pronuncia di ยา, dovrebbe essere ดอก
* Migliora la pronuncia di กา, dovrebbe essere เตอร์
* Migliora la pronuncia di ญจน, dovrebbe essere นิติ
* Migliora la pronuncia di วัน, dovrebbe essere ยา
* Migliora la pronuncia di เป็น, dovrebbe essere กา
* Migliora la pronuncia di ประธาน, dovrebbe essere อน
* Migliora la pronuncia di โครงการ, dovrebbe essere วัน

– **Caso corretto:** La frase พลังงาน คือ ความสามารถ ทำงาน è molto simile a พลังงาน คือ ความสามารถ ทำงาน (1.00). Ottimo lavoro! Solo piccole variazioni fonetiche potrebbero essere migliorate.

5.3 Utilità Dell'Approccio Nel Caso di Studio Considerato

Il caso di studio presentato ha l'obiettivo principale di dimostrare l'efficacia di un sistema integrato per l'analisi delle produzioni linguistiche in lingua tailandese da parte di studenti

non madrelingua. La pipeline proposta è progettata per fornire una valutazione accurata della coerenza semantica e della correttezza delle frasi pronunciate, anche in presenza di errori tipici del parlato L2, e allo stesso tempo per generare un feedback immediato, interpretabile e personalizzato, capace di guidare l'utente nel miglioramento progressivo delle competenze linguistiche.

L'approccio combinato di trascrizione automatica, sintesi vocale, analisi semantica e confronto fonetico consente di integrare informazioni quantitative, come la similarità semantica e la gravità degli errori, con informazioni qualitative. Questa integrazione permette di ottenere una valutazione completa e modulare della produzione orale. La scelta di utilizzare modelli robusti sia al parlato con errori sia alle variazioni prosodiche deriva dalla necessità di garantire risultati affidabili anche in scenari reali, caratterizzati da rumore ambientale o pronunce divergenti.

In conclusione, questa sezione illustra le potenzialità della pipeline e introduce in maniera naturale il capitolo successivo, dedicato alla valutazione sperimentale del sistema. Nel capitolo seguente verranno presentati i test condotti per rispondere alle domande di ricerca, l'analisi dei risultati ottenuti e la discussione sull'efficacia dell'approccio proposto, fornendo un riscontro qualitativo sul funzionamento del sistema.

CAPITOLO 6

Valutazione sperimentale

In questo capitolo viene presentata la valutazione sperimentale del sistema sviluppato per la trascrizione del parlato thailandese con injection controllato di errori fonetici. L'obiettivo principale di questa fase sperimentale consiste nell'analizzare in maniera sistematica come differenti tipologie e quantità di errori fonetici possano influenzare la capacità del modello di trascrivere correttamente il parlato, e al contempo valutare come tali errori incidano sulla percezione del significato delle frasi da parte degli utenti.

La valutazione è organizzata attorno a tre domande di ricerca principali, ciascuna delle quali è supportata da esperimenti mirati e dall'analisi di specifici sottoinsiemi del dataset di test. Le prime due domande (RQ1 e RQ2) si concentrano sulla fedeltà della trascrizione e utilizzano come metrica il *Character Error Rate* (CER) (Phatthiyaphaibun et al., 2022), mentre la terza domanda (RQ3) analizza in maniera approfondita il significato percepito delle frasi tramite una metrica di *similarità semantica* (Mahmoud et al., 2025) e un'analisi statistica complementare.

6.1 Configurazioni sperimentali

Gli esperimenti sono stati condotti in un ambiente controllato, utilizzando procedure implementate in Python che integrano le fasi di pre-processing, inferenza automatica e analisi statistica dei risultati. Tutte le fasi di addestramento, validazione e test sono state eseguite

su hardware a risorse costanti, al fine di garantire la replicabilità dei test e la coerenza dei risultati.

Il dataset di riferimento utilizzato è il **LOTUS** (Sertsi et al., 2016), un corpus di parlato in lingua thailandese ampiamente impiegato nella ricerca sul riconoscimento automatico del parlato. A partire da questo corpus, è stata creata una versione estesa contenente frasi con *injection* controllato di errori fonetici, simulando in maniera sistematica le tipiche deviazioni introdotte dai parlanti non nativi. Le modifiche hanno interessato le consonanti iniziali, le vocali, le consonanti finali e i toni, consentendo di controllare sia la tipologia sia la quantità di errori fonetici presenti nelle frasi generate.

Ogni istanza del dataset è accompagnata da metadati che includono la trascrizione corretta, la versione con errori fonetici, la tipologia di errore introdotto e la relativa posizione nella frase. Questa struttura consente un'analisi dettagliata e mirata degli errori, facilitando l'individuazione dei pattern ricorrenti e la valutazione della sensibilità del modello alle diverse categorie di deviazione fonetica.

Inoltre, per ciascuna frase è stata generata una corrispondente registrazione audio sintetica tramite modelli *Text-to-Speech* (TTS), al fine di disporre di una versione multimodale del corpus (testo e parlato). Ciò rende il dataset adatto sia ad attività di addestramento e valutazione di sistemi di riconoscimento vocale automatico. La Figura **??** riporta un estratto del dataset impiegato, illustrando la sua organizzazione interna e la struttura dei dati raccolti.

Il dataset così ottenuto è stato suddiviso in tre sottoinsiemi destinati rispettivamente alle fasi di addestramento, validazione e test, secondo una proporzione dell'80%, 10% e 10%. La suddivisione è stata effettuata in modo stratificato, in modo da garantire una rappresentatività equilibrata delle diverse tipologie di errore in ciascun sottoinsieme. Il test set risultante comprende complessivamente 326 registrazioni audio, ciascuna associata alla corrispondente trascrizione di riferimento.

Per rispondere alle specifiche esigenze delle domande di ricerca, il test set è stato ulteriormente organizzato. Per la prima domanda di ricerca (RQ1), sono state selezionate dieci frasi contenenti esclusivamente una singola tipologia di errore, al fine di isolare l'effetto della tipologia di errore sulla capacità del modello di produrre trascrizioni fedeli. Per la seconda domanda di ricerca (RQ2), invece, è stato necessario garantire un numero minimo di esempi per ciascun livello di quantità di errori per frase, fino a dieci esempi per livelli da uno a dieci errori, generando file supplementari quando necessario. Questo ha permesso di analizzare l'impatto della quantità di errori mantenendo comparabilità tra i livelli.

Per la terza domanda di ricerca (RQ3), il test set è stato ulteriormente arricchito con una

colonna contenente il numero totale di errori presenti in ciascuna frase e un indice di similarità semantica tra la trascrizione alterata e quella corretta. Quest'ultimo è stato stimato tramite un modello di embeddings multilingue, *paraphrase-multilingual-MiniLM-L12-v2* (Mahmoud et al., 2025), in grado di generare rappresentazioni vettoriali e di fornire un punteggio normalizzato compreso tra 0 e 1, indicativo di quanto la frase trascritta preservi il significato dell'originale.

Il modello impiegato in tutti gli esperimenti, denominato **Eval Model**, corrisponde alla versione fine-tuned del sistema di riconoscimento vocale automatico `wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al., 2022), sviluppato per questo studio. L'addestramento è stato condotto per un totale di 20 epoche, utilizzando un *learning rate* di 1×10^{-4}, una *batch size* di sedici e l'ottimizzatore Adam (Kingma and Ba, 2014). Il modello è stato addestrato su dati contenenti errori fonetici generati artificialmente, con l'obiettivo di incrementarne la robustezza nei confronti di pronunce non standard o distorte. Non sono state applicate tecniche specifiche di *data augmentation* (Gupta et al., 2023), ad eccezione della regolarizzazione standard prevista dall'architettura, al fine di preservare la naturale variabilità del dataset e garantire stabilità durante l'addestramento.

6.2 Domande di ricerca

La presente sezione illustra le tre domande di ricerca che hanno guidato la progettazione e la valutazione degli esperimenti. Ciascuna domanda è stata formulata per indagare un diverso aspetto della robustezza del modello di riconoscimento vocale automatico `wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al., 2022) in presenza di errori fonetici generati artificialmente. L'obiettivo complessivo è comprendere in che misura il modello riesca a mantenere la fedeltà fonetica e semantica delle trascrizioni al variare della tipologia e della quantità di errori introdotti. Le analisi descritte di seguito si basano sulle configurazioni sperimentali introdotte nella sezione precedente, applicando metriche specifiche in funzione dell'obiettivo di ciascuna domanda di ricerca.

6.2.1 RQ1. Quanto impatta la tipologia di errore rispetto alla capacità di trascrizione fedele?

La prima domanda di ricerca analizza l'impatto della **tipologia di errore fonetico** (toni, consonanti, vocali) sulla capacità del modello di produrre una trascrizione fedele. Questa analisi consente di identificare quali categorie di errore compromettono maggiormente la

prestazione del sistema, fornendo indicazioni utili per l'ottimizzazione o l'adattamento del modello a specifici tipi di deviazioni fonetiche.

Per rispondere alla domanda, è stato utilizzato il test set derivato dal preprocessing del dataset originale, comprendente 326 registrazioni audio suddivise in training, validation e test set. Dal test set sono state selezionate dieci frasi per ciascuna categoria di errore: solo errori sui toni, solo errori sulle consonanti e solo errori sulle vocali. Questa selezione ha permesso di isolare l'effetto di ciascuna tipologia di errore e di osservare in che modo essa influenzi la fedeltà della trascrizione.

Per semplificare l'analisi, la colonna `dettaglio_modifica` del dataset, che riportava informazioni estremamente specifiche (come sostituzioni di consonanti o aggiunte/rimozioni di toni), è stata ridotta a una rappresentazione compatta contenente solo le categorie di errore. Ad esempio, una voce del dataset come `SUONO_CONSONANTE - กัด: ก→ข | Aggiunto tono: ้ in ขัด | Rimosso ้ da กัด` è stata trasformata in `TONO; CONSONANTE`. Questa procedura ha permesso di focalizzare l'analisi sul ruolo della tipologia di errore, riducendo la complessità descrittiva dei dati e mantenendo la comparabilità tra le classi di errore.

6.2.2 RQ2. Quanto impatta la quantità di errori rispetto alla capacità di trascrizione fedele?

La seconda domanda di ricerca si concentra sull'effetto della **quantità di errori fonetici iniettati** sulla capacità del modello di mantenere una trascrizione fedele. Lo scopo è valutare la **robustezza del modello** al progressivo aumento del rumore fonetico, osservando come la qualità della trascrizione degradi all'aumentare del numero di deviazioni rispetto alla pronuncia corretta.

Gli errori iniettati variavano da uno a tredici, ma per i livelli più alti (oltre sette) il dataset originario conteneva pochi esempi, rendendo necessaria un'integrazione controllata dei dati. Per garantire almeno dieci esempi per ciascun livello di errore (da uno a dieci), sono stati generati file supplementari con frasi manipolate fino a raggiungere il numero desiderato di deviazioni, in particolare per le classi con otto, nove e dieci errori. In totale sono stati prodotti settanta esempi aggiuntivi. Questa procedura ha consentito di analizzare sistematicamente l'impatto della quantità di errori sulla fedeltà della trascrizione, mantenendo omogeneità tra i livelli e assicurando la comparabilità dei risultati.

6.2.3 RQ3: Quanto il modello di valutazione degli errori altera il significato semantico della frase rispetto alla versione corretta?

La terza domanda di ricerca si concentra sull'analisi della capacità del modello di valutazione degli errori, denominato "**Eval Model**", di preservare il significato semantico delle frasi in presenza di errori fonetici previamente iniettati. In questa fase, l'interesse non riguarda la generazione di errori da parte del modello, ma la coerenza semantica della trascrizione prodotta rispetto alla versione corretta. Per ciascun esempio del test set, la trascrizione prodotta dall' **Eval Model** è stata confrontata con la trascrizione di riferimento mediante un modello di embedding `paraphrase-multilingual-MiniLM-L12-v2` (Mahmoud et al., 2025), che rappresenta ciascuna frase come un vettore numerico nello spazio semantico. Successivamente, è stata calcolata la *cosine similarity* (Mahmoud et al., 2025) tra i vettori delle due trascrizioni, ottenendo un valore compreso tra 0 e 1, interpretabile come misura della similarità semantica: valori prossimi a 1 indicano un alto grado di corrispondenza semantica, mentre valori più bassi riflettono una maggiore discrepanza nel significato. Questo procedimento consente di quantificare in modo rigoroso l'impatto degli errori iniettati sul contenuto trascritto, valutando la misura in cui le deviazioni fonetiche alterino la fedeltà semantica rispetto all'originale.

6.3 Metriche di valutazione

La fase di valutazione sperimentale ha avuto l'obiettivo di quantificare in modo sistematico le prestazioni del sistema proposto, distinguendo tra gli aspetti di accuratezza fonetica e quelli di coerenza semantica. Poiché ciascuna domanda di ricerca indaga una dimensione specifica del comportamento del modello, sono state impiegate metriche di valutazione differenti ma complementari, selezionate in funzione della natura dei fenomeni analizzati.

In particolare, per le prime due domande di ricerca (RQ1 e RQ2) l'attenzione è stata rivolta alla **fedeltà della trascrizione** rispetto al riferimento, analizzando la quantità e il tipo di errori commessi dal modello. Le metriche utilizzate in questa fase mirano quindi a valutare quanto il modello riproduca correttamente il contenuto dell'audio, senza concentrarsi sulla correttezza fonetica o lessicale, ma sulla capacità di trascrivere in modo fedele la sequenza di parole o suoni presenti nell'audio di input.

Per la terza domanda di ricerca (RQ3), invece, l'obiettivo si sposta dal livello puramente fonetico a quello **semantico**, esaminando in che misura gli errori di trascrizione incidano sul significato complessivo delle frasi. In questo caso, la valutazione si fonda su metriche che misurano la similarità semantica tra la trascrizione generata e la frase di riferimento,

permettendo di stimare non solo la quantità di errori, ma anche il loro impatto sul contenuto informativo e sulla comprensibilità del messaggio.

Questa distinzione metodologica consente di analizzare in modo più articolato il comportamento dell' **Eval Model**, separando l'accuratezza superficiale dalla fedeltà semantica. Le metriche così definite offrono quindi una prospettiva integrata: da un lato, quantificano gli errori fonetici e strutturali, dall'altro, valutano la preservazione del significato, elemento cruciale per determinare l'effettiva qualità delle trascrizioni generate.

6.3.1 RQ1 e RQ2: Accuratezza della trascrizione

Per RQ1 (tipologia di errore) e RQ2 (quantità di errori), la metrica principale utilizzata è il *Character Error Rate* (CER) (Phatthiyaphaibun et al., 2022), una misura standard nel riconoscimento automatico del parlato che quantifica la discrepanza tra la trascrizione prodotta dal modello e il testo di riferimento.

Il CER si calcola secondo la formula:

$$CER = \frac{S + D + I}{N}$$

dove S rappresenta il numero di sostituzioni, D le cancellazioni, I gli inserimenti e N il numero totale di caratteri nel testo di riferimento. Un valore di CER pari a 0 indica una trascrizione perfetta, mentre valori crescenti evidenziano un aumento della discrepanza rispetto al testo originale.

6.3.2 RQ3: Impatto semantico degli errori

Per RQ3, la metrica principale adottata è stata la *similarità semantica*, calcolata come *cosine similarity* tra vettori di embedding (Mahmoud et al., 2025), con valori compresi tra 0 e 1: punteggi più elevati indicano una maggiore vicinanza semantica tra la trascrizione prodotta dall'**Eval Model** e la frase di riferimento.

La *cosine similarity* misura il grado di vicinanza direzionale tra due vettori nello spazio semantico ed è definita come:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{6.3.1}$$

dove A e B rappresentano i vettori di embedding delle due frasi, $A \cdot B$ indica il prodotto scalare tra i vettori, e $\|A\|$ e $\|B\|$ sono le loro norme euclidee. Il valore risultante varia tra 0 e 1, dove punteggi più elevati indicano una maggiore somiglianza semantica.

La **gravità** degli errori è stata definita come il numero totale di errori iniettati per ciascuna trascrizione, classificata in quattro livelli qualitativi, secondo il seguente criterio:

- **Nessun errore**: 0 errori iniettati;
- **Lieve**: fino a 3 errori iniettati;
- **Media**: da 4 a 6 errori iniettati;
- **Grave**: più di 6 errori iniettati.

Per ogni livello sono state calcolate la media, la deviazione standard e la distribuzione dei valori di similarità semantica, al fine di valutare l'impatto crescente della gravità sulla preservazione del significato.

6.4 Analisi dei risultati

In questa sezione vengono presentati e descritti i risultati ottenuti dagli esperimenti delineati nel capitolo precedente, articolati in funzione delle tre domande di ricerca (RQ1, RQ2 e RQ3). L'obiettivo principale è fornire una visione complessiva e sistematica dell'andamento delle metriche di valutazione adottate, mantenendo un approccio descrittivo e oggettivo, privo di interpretazioni critiche che saranno invece approfondite nella sezione successiva.

I risultati riportati derivano dalle valutazioni condotte sul *test set* del corpus modificato, costruito a partire dal dataset *Lotus* (Sertsi et al., 2016) e arricchito con frasi contenenti errori fonetici iniettati artificialmente. Tale struttura ha consentito di misurare l'impatto degli errori di pronuncia, in termini sia di tipologia sia di quantità, sulla capacità del modello di riconoscimento vocale di restituire trascrizioni accurate e semanticamente coerenti.

Il modello oggetto di analisi, denominato **Eval Model**, corrisponde alla versione fine-tuned di `wav2vec2-large-xlsr-53-th` (Phatthiyaphaibun et al., 2022), addestrato specificamente su dati contenenti errori, rappresentativi delle tipiche deviazioni fonetiche e linguistiche commesse dagli studenti L2. Tale configurazione ha lo scopo di rafforzare la robustezza del sistema di fronte a pronunce errate tipiche dei parlanti non nativi, assicurando al contempo che il modello mantenga un comportamento coerente con la lingua di riferimento.

Dal punto di vista metodologico, le valutazioni sono state eseguite considerando due prospettive complementari: da un lato, l'**accuratezza fonetica** e ortografica della trascrizione, misurata attraverso il *Character Error Rate (CER) Phatthiyaphaibun et al. (2022)*; dall'altro,

la **coerenza semantica** del contenuto trascritto, valutata tramite una metrica di *similarità semantica* (Mahmoud et al., 2025) basata su rappresentazioni vettoriali multilingue.

In accordo a quanto specificato precedentemente, le analisi sono organizzate attorno alle tre domande di ricerca principali:

- RQ1: valutare l'influenza della *tipologia di errore iniettato* (es. tono, consonante, vocale) sulla fedeltà della trascrizione;
- RQ2: stimare l'impatto della *quantità di errori* presenti in ciascuna frase sulla precisione complessiva della trascrizione;
- RQ3: analizzare in che misura gli errori di trascrizione alterino il *significato semantico* della frase rispetto alla versione corretta.

I risultati vengono pertanto presentati in tre sottosezioni corrispondenti, ognuna dedicata a una specifica domanda di ricerca. Le tabelle e le rappresentazioni grafiche incluse mostrano l'andamento dei valori medi, la variabilità delle metriche e le tendenze emergenti nei diversi scenari di errore. In questa fase l'obiettivo è fornire una descrizione dettagliata e comparabile dei risultati, che costituirà la base per l'interpretazione critica e la discussione finale delle implicazioni linguistiche e computazionali del sistema.

6.4.1 Analisi dei risultati RQ1:

Per rispondere alla prima domanda di ricerca (**RQ1: Quanto impatta la tipologia di errore rispetto alla capacità di trascrizione fedele?**), è stata condotta un'analisi approfondita volta a valutare l'influenza della tipologia di errore fonetico sulla fedeltà della trascrizione. La comprensione di questo aspetto è fondamentale per determinare se il modello necessita di strategie di correzione mirate per specifici tipi di errore, oppure se può mantenere prestazioni consistenti indipendentemente dalla natura dell'errore presente nel testo.

In particolare, sono stati calcolati i valori di *Character Error Rate* (CER) (Phatthiyaphaibun et al., 2022) per ciascuna categoria di errore: toni, vocali e consonanti. Il calcolo del CER è stato effettuato sia considerando il valore medio su ciascun gruppo di frasi, sia valutando il CER a livello di singolo campione, mediante l'utilizzo della libreria `jiwer`. Il CER, essendo una misura basata sul confronto tra sequenze di caratteri, fornisce un'indicazione precisa del numero di inserimenti, cancellazioni o sostituzioni necessarie per ottenere la trascrizione corretta. I risultati ottenuti sono riportati in Tabella 6.1, con valori medi pari a 0.1738 per le frasi con errori sui toni, 0.1622 per quelle con errori sulle vocali e 0.1529 per le frasi con errori sulle consonanti.

Tabella 6.1: Sintesi dei valori medi di CER e dei risultati del test di normalità Shapiro–Wilk per ciascuna categoria di errore fonetico.

Categoria	**CER medio**	**W Shapiro–Wilk**	**p-value**
Toni	0.1738	0.9503	0.6721
Vocali	0.1622	0.9390	0.5415
Consonanti	0.1529	0.9273	0.4222

Prima di procedere al confronto tra le categorie, è stata verificata la normalità delle distribuzioni del CER mediante il test di Shapiro–Wilk. Questo test prevede la valutazione dell'ipotesi nulla secondo cui un campione proviene da una popolazione normalmente distribuita, restituendo un valore W e un corrispondente p-value. I risultati indicano che tutte le distribuzioni sono compatibili con l'ipotesi di normalità, come evidenziato dai valori di W e dai rispettivi p-value riportati in Tabella 6.1. Questa condizione ha reso possibile l'utilizzo di un test parametrico per confrontare le medie dei gruppi, garantendo l'adeguatezza statistica dell'analisi. Inoltre, il controllo della normalità permette di evitare interpretazioni errate dovute a distribuzioni fortemente asimmetriche o contenenti outlier estremi.

Per determinare se le differenze osservate tra le categorie fossero statisticamente significative, è stata quindi condotta un'ANOVA one-way. Tale test è appropriato quando i dati seguono una distribuzione normale e consente di confrontare le medie di più gruppi, fornendo una stima dell'influenza della variabile indipendente (tipologia di errore) sulla variabile dipendente (CER). L'ANOVA restituisce due valori principali: il valore F, che rappresenta il rapporto tra la varianza spiegata dal modello e la varianza residua non spiegata, e il p-value, che indica la probabilità di osservare un valore di F almeno così estremo assumendo che l'ipotesi nulla sia vera. I risultati dell'ANOVA sono riportati in Tabella 6.2. L'analisi ha restituito un valore di F pari a 1.0187 con un p-value di 0.3745, indicando che le differenze nei CER medi tra le tre categorie non sono statisticamente significative. Questo risultato suggerisce che, a livello quantitativo, il tipo di errore fonetico non incide in maniera rilevante sulla capacità del modello di produrre trascrizioni accurate.

Tabella 6.2: Risultati dell'ANOVA one-way sul CER per le tre categorie di errore fonetico.

Test	F	p-value
ANOVA one-way	1.0187	0.3745

Per facilitare la comprensione visiva delle differenze tra le categorie di errore, sono stati realizzati due grafici complementari. Il primo (Figura 6.1) è un *barplot*, comunemente utilizzato per confrontare i valori medi di una variabile tra differenti gruppi. Nel presente caso, le barre illustrano i valori medi di CER per ciascuna categoria di errore fonetico (*toni*, *vocali* e *consonanti*). Il grafico mostra che i toni presentano un valore medio leggermente superiore (17.38%), seguiti dalle vocali (16.22%) e dalle consonanti (15.29%). Sebbene tali differenze siano visibili, esse risultano contenute e non suggeriscono una variazione sostanziale nella capacità del modello di produrre trascrizioni fedeli.

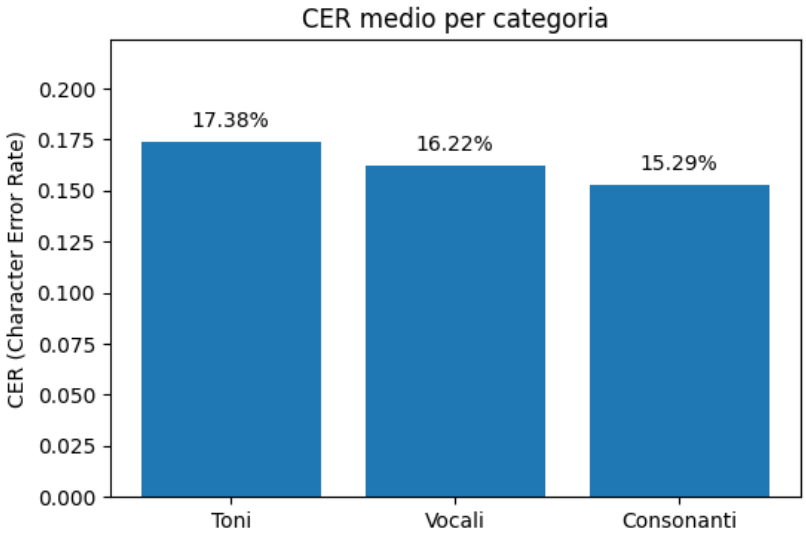

Figura 6.1: CER medio per ciascuna categoria di errore fonetico.

Il secondo grafico (Figura 6.2) è un *boxplot*, che mostra la distribuzione statistica dei valori di CER, evidenziando mediana, quartili, variabilità interna e valori anomali. Le distribuzioni risultano simili tra loro, con mediane comparabili e sovrapposizione degli intervalli interquartili, indicando che la variabilità interna delle categorie è analoga. Questa analisi visiva permette di osservare eventuali pattern o outlier che potrebbero non emergere dai valori medi, fornendo una rappresentazione più completa del comportamento del modello.

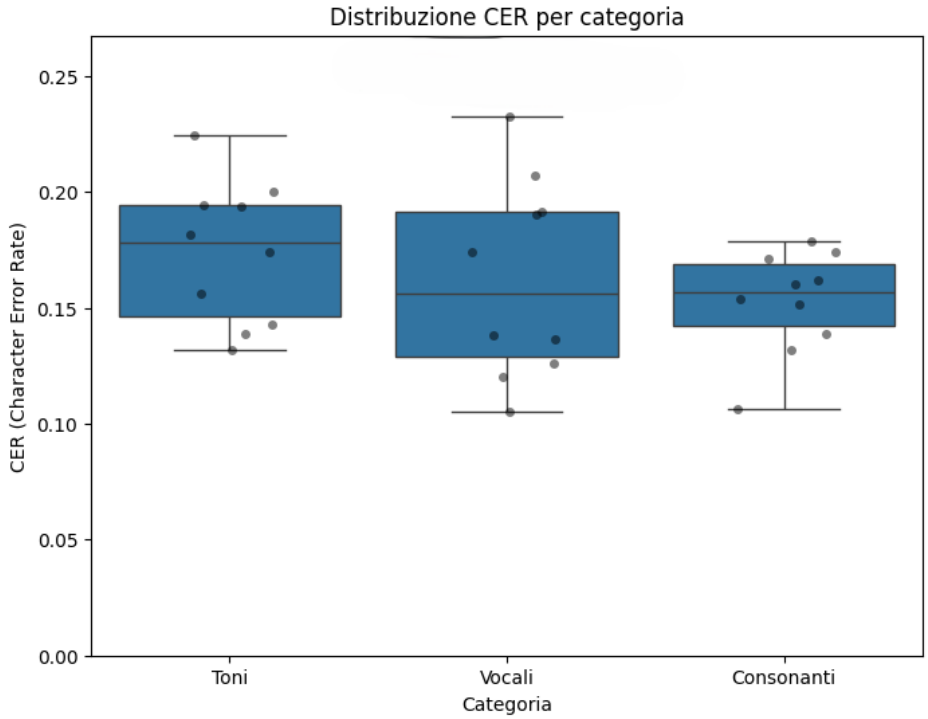

Figura 6.2: Distribuzione dei CER a livello di singolo campione per le tre categorie di errore fonetico.

Dal punto di vista pratico, l'assenza di differenze significative suggerisce che il modello mantiene una fedeltà costante nella trascrizione indipendentemente dal tipo di errore fonetico. Ciò implica che eventuali strategie di correzione o ottimizzazione del modello non devono necessariamente concentrarsi su una specifica categoria di errore (toni, vocali o consonanti), ma possono essere indirizzate in modo più generale alla riduzione complessiva del CER. Inoltre, queste evidenze supportano la robustezza del modello nel trattare trascrizioni contenenti errori fonetici di diversa natura, aumentando la fiducia nella sua applicabilità a scenari reali con input variabili.

Complessivamente, l'analisi fornisce una risposta chiara e quantitativamente fondata alla **RQ1**, sottolineando la robustezza del modello rispetto a diverse forme di errore fonetico e indicando le direttrici più appropriate per futuri interventi di ottimizzazione.

6.4.2 Analisi dei risultati RQ2:

Per rispondere alla seconda domanda di ricerca (**RQ2: Quanto impatta la quantità di errori rispetto alla capacità di trascrizione fedele?**), è stata condotta un'analisi volta a valutare l'impatto quantitativo della presenza di errori multipli sui risultati del modello. Comprendere questa relazione è cruciale per identificare soglie critiche oltre le quali le prestazioni del sistema degradano sensibilmente.

Gli esperimenti sono stati condotti su un dataset costituito da frasi di diversa lunghezza e complessità linguistica. Gli errori inseriti artificialmente nelle frasi appartengono a tre categorie principali: **modifiche di tono, vocale** e **alterazioni di consonanti**. Per la presente analisi, sono stati considerati tutti gli errori senza distinzione di tipologia, poiché l'obiettivo principale era valutare l'impatto *quantitativo* del numero totale di errori iniettati su ciascuna frase, indipendentemente dalla natura specifica della perturbazione.

I valori medi di CER, calcolati per ciascun numero di errori da 1 a 10, mostrano una tendenza chiara all'aumento del tasso di errore con l'aumentare del numero di errori iniettati. In particolare, per una frase con un solo errore, il CER medio risulta pari a 0.1675, mentre per due e tre errori i valori medi sono rispettivamente 0.1770 e 0.1774. Con quattro e cinque errori, il CER medio cresce leggermente, raggiungendo 0.1808 e 0.1946. Il gruppo con sei errori mostra un valore medio di 0.2111, mentre per sette errori si osserva una riduzione a 0.1796, evidenziando una variabilità nella risposta del modello a errori intermedi. Il CER medio aumenta poi drasticamente con otto, nove e dieci errori, raggiungendo valori pari a 0.3615, 0.3849 e 0.3565, confermando un punto critico oltre il quale le prestazioni del sistema peggiorano sensibilmente.

Per ciascun gruppo è stato effettuato un test di Shapiro-Wilk per verificare la normalità della distribuzione dei valori di CER. La maggior parte dei gruppi risulta normalmente distribuita, ad eccezione del gruppo con sei errori (Shapiro-Wilk $p = 0.0361$), mentre tutti gli altri presentano valori di p superiori a 0.13, indicando la non significatività della deviazione dalla normalità . Alla luce di questi risultati, si è scelto di utilizzare la **correlazione di Spearman** per valutare l'associazione tra numero di errori e CER medio, ottenendo un coefficiente pari a 0.891 con $p = 0.001$, a conferma di una forte correlazione positiva: all'aumentare degli errori, il CER cresce in modo consistente.

I risultati sono ulteriormente illustrati attraverso tre rappresentazioni grafiche complementari. Il boxplot della Figura 6.3 mostra la distribuzione dei valori di CER per ciascun numero di errori, evidenziando un aumento della mediana con l'incremento degli errori e una maggiore dispersione a partire da sette errori, con la presenza di outlier che indicano una diversa sensibilità dei testi alla perturbazione.

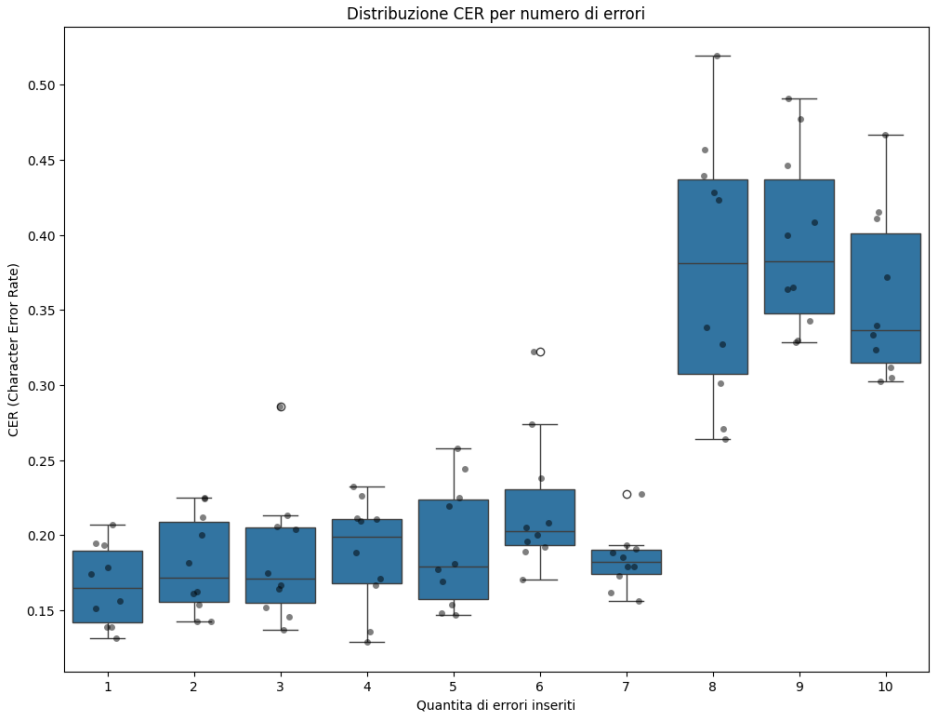

Figura 6.3: Distribuzione del CER per numero di errori inseriti.

Il grafico mostrato in Figura 6.4 mette in evidenza il trend crescente del CER medio, con un incremento marcato tra sette e otto errori, coerente con i valori numerici riportati sopra.

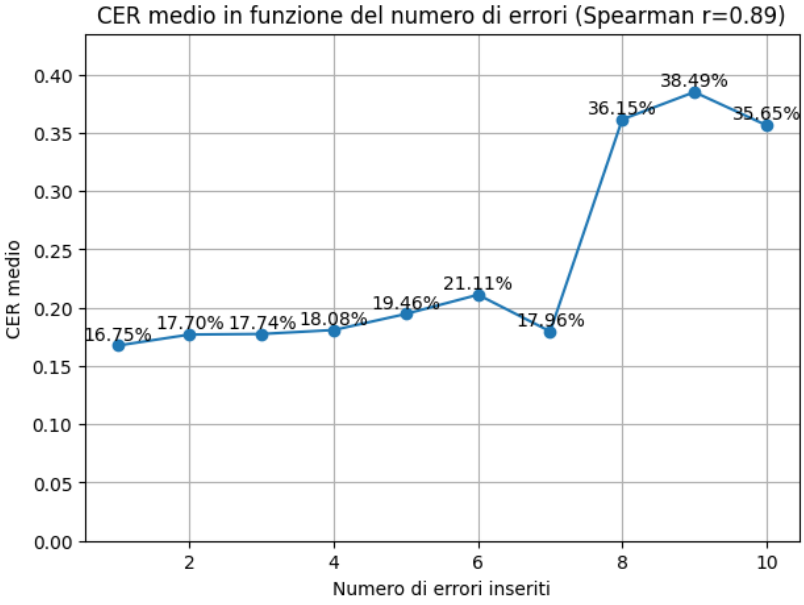

Figura 6.4: Trend del CER medio in funzione del numero di errori.

Infine, il grafico a barre della Figura 6.5 mostra come i valori medi di CER restino relativamente stabili tra uno e sette errori, oscillando tra circa 16.75% e 21.11%, per poi raddoppiare

a partire da otto errori, raggiungendo valori compresi tra 35.65% e 38.49%, evidenziando un punto critico nella capacità del sistema di gestire errori multipli.

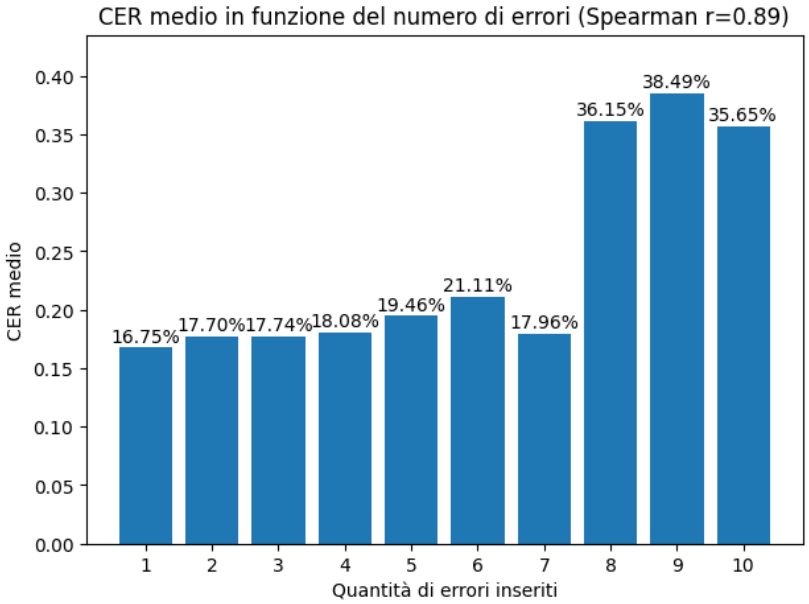

Figura 6.5: CER medio per ciascun numero di errori inseriti.

Un'analisi comparativa delle tre rappresentazioni grafiche evidenzia come ciascun plot metta in luce aspetti diversi del comportamento del modello. Il boxplot permette di osservare la variabilità dei singoli valori di CER e la presenza di outlier, fornendo informazioni sulla dispersione e sulla sensibilità di specifiche frasi agli errori iniettati. Il grafico a linea enfatizza il trend generale del CER medio, mostrando chiaramente l'incremento progressivo e l'identificazione della soglia critica oltre la quale le prestazioni peggiorano rapidamente. Infine, il grafico a barre sintetizza i valori medi in maniera immediata, evidenziando la stabilità relativa del modello fino a sette errori e il drastico aumento del tasso di errore oltre tale soglia. La combinazione di queste tre visualizzazioni fornisce dunque una comprensione completa del fenomeno, con un approccio sia quantitativo sia visivo, utile per supportare interpretazioni approfondite e decisioni di ottimizzazione del sistema.

Dal punto di vista pratico, i risultati indicano che il modello mantiene prestazioni accettabili fino a un numero moderato di errori, ma oltre la soglia critica di sette errori la capacità di trascrizione fedele diminuisce drasticamente. L'analisi condotta dimostra in modo empiricamente fondato che l'aumento progressivo del rumore linguistico compromette la coerenza fonetica e ortografica delle trascrizioni, evidenziando la necessità di introdurre meccanismi di robustezza specifici per la gestione di input fortemente alterati. Complessivamente, questa evidenza fornisce una risposta quantitativa e dettagliata alla **RQ2**, delineando le basi per

futuri interventi di ottimizzazione mirati al miglioramento della resilienza del sistema.

6.4.3 Analisi dei risultati RQ3:

Per rispondere alla terza domanda di ricerca (**RQ3:Quanto l' Eval Model altera il significato semantico della frase rispetto alla versione corretta?**), è stata condotta un'analisi statistica approfondita volta a valutare come l'aumento della gravità degli errori influenzi la capacità del modello di preservare il significato originario del testo trascritto. Comprendere questa relazione è fondamentale per identificare potenziali criticità nella gestione di errori più gravi e per indirizzare eventuali strategie di ottimizzazione del modello.

In primo luogo, è stata calcolata la correlazione di Spearman tra gravità e similarità semantica, scelta per la sua capacità di misurare l'associazione monotona tra due variabili ordinali o continue, senza assumere linearità o distribuzione normale dei dati (Spearman, 1904). Questo coefficiente, indicato con ρ, varia tra -1 e 1: valori positivi indicano che le due variabili tendono a crescere insieme, valori negativi che una aumenta quando l'altra diminuisce, mentre valori prossimi a zero indicano assenza di correlazione monotona.

I risultati, riportati in Tabella 6.3, mostrano una correlazione negativa significativa ($\rho = -0.2977,\ p < 0.001$), indicando che all'aumentare della gravità degli errori la similarità semantica tende a diminuire.

Tabella 6.3: Risultati della correlazione di Spearman tra gravità degli errori e similarità semantica

Test	$\boldsymbol{\rho}$	**p-value**	**Interpretazione**
Spearman	-0.2977	<0.001	Correlazione negativa significativa: all'aumentare della gravità, la similarità semantica tende a diminuire.

Successivamente, per verificare se la similarità semantica differisse significativamente tra i diversi livelli di gravità, è stato condotto il test non parametrico di Kruskal–Wallis. Questo test è particolarmente adeguato per confrontare più gruppi indipendenti senza assumere la normalità delle distribuzioni e valuta se almeno uno dei gruppi differisce dagli altri in termini di mediana.

Il test restituisce due parametri principali: il valore H, che rappresenta la statistica del test e misura la somma dei ranghi tra i gruppi confrontati rispetto alla variabilità complessiva, e il p-value, che indica la probabilità di osservare un valore di H almeno così estremo assumendo che l'ipotesi nulla (assenza di differenze tra i gruppi) sia vera.

I risultati, mostrati in Tabella 6.4, indicano la presenza di differenze significative tra almeno due livelli di gravità ($H = 35.921,\ p < 0.001$), suggerendo che l'intensità degli errori fonetici influisce sulla similarità semantica tra trascrizione e frase di riferimento.

Tabella 6.4: Risultati del test per confrontare i livelli di gravità sulla similarità semantica

Test	**H**	**p-value**	**Interpretazione**
Kruskal–Wallis	35.921	<0.001	Differenze significative tra almeno due livelli di gravità.

In aggiunta ai test precedenti, è stata condotta un'analisi di regressione robusta (*Robust Linear Model*, RLM) per stimare la relazione lineare tra gravità e similarità semantica. La RLM è una variante della regressione lineare che limita l'influenza di outlier o osservazioni anomale, fornendo stime più stabili dei coefficienti rispetto alla regressione lineare classica.

I principali parametri di valutazione della RLM includono i coefficienti $\boldsymbol{\beta}$, che rappresentano l'effetto stimato di ciascuna variabile indipendente sulla variabile dipendente, l'intercetta, che indica il valore previsto della variabile dipendente quando tutte le variabili indipendenti sono nulle, e il *p-value* associato a ciascun coefficiente, che fornisce una misura della significatività statistica dell'effetto stimato. In alcuni casi possono essere considerati anche gli intervalli di confidenza dei coefficienti, che indicano l'incertezza associata alla stima. I risultati, riportati in Tabella 6.5, evidenziano un coefficiente negativo significativo per la gravità ($\boldsymbol{\beta} = -0.0047,\ p < 0.002$), confermando che l'aumento della gravità degli errori riduce la similarità semantica delle trascrizioni.

Tabella 6.5: Risultati della regressione robusta (RLM) tra gravità e similarità semantica

Variabile	$\boldsymbol{\beta}$	**Std. Err**	**z**	**p-value**	**95% CI**
Intercetta	0.9444	0.006	166.74	<0.001	[0.933, 0.956]
Gravità	-0.0047	0.001	-3.136	0.002	[-0.008, -0.002]

Per facilitare la comprensione visiva dei risultati, sono stati realizzati due grafici complementari. La Figura 6.6 mostra uno scatterplot tra gravità e similarità semantica con sovrapposta la linea di regressione robusta (RLM). La linea rossa tratteggiata rappresenta la tendenza stimata, mentre i punti blu indicano le osservazioni individuali. Come evidenziato dalla pendenza negativa della linea RLM, la similarità semantica tende a diminuire progressivamente all'aumentare del numero di errori, suggerendo che una maggiore gravità comporti

una perdita di coerenza semantica. Tuttavia, la dispersione dei punti mostra che la relazione non è perfettamente lineare: per livelli intermedi di gravità (tra 3 e 6 errori) si osserva una variabilità considerevole, con valori di similarità anche elevati. Questo fenomeno indica la presenza di margini di sovrapposizione e suggerisce che l'impatto semantico degli errori non dipenda unicamente dal loro numero, ma anche dalla loro natura linguistica o dalla posizione all'interno del testo.

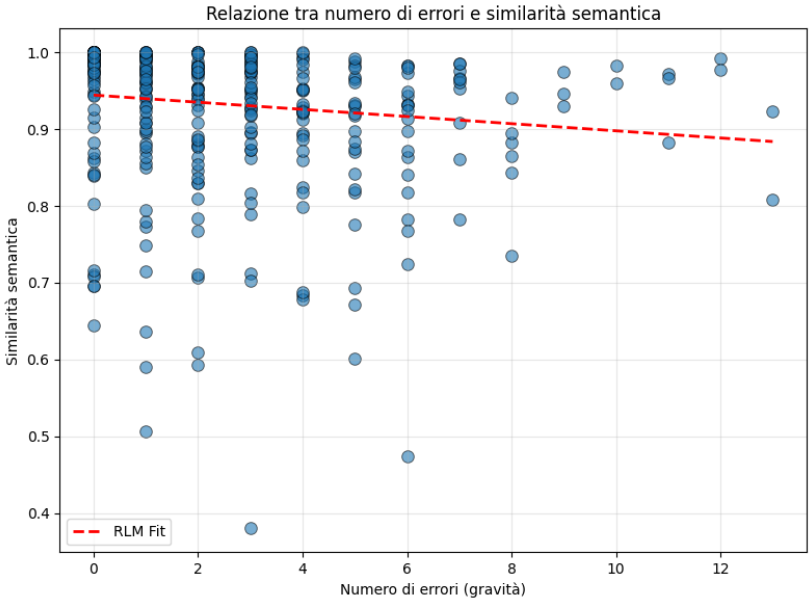

Figura 6.6: Scatterplot di gravità vs similarità semantica con linea di regressione robusta (RLM).

La Figura 6.7 mostra invece la distribuzione della similarità semantica per ciascun livello qualitativo di gravità. I punti neri rappresentano osservazioni individuali, i punti rossi la media, mentre le scatole mostrano l'intervallo interquartile e la linea centrale la mediana. Si osserva che la similarità è più elevata in assenza di errori e decresce progressivamente con l'aumentare della gravità. Il livello *grave* presenta la variabilità più ampia e la media più bassa, confermando l'effetto deleterio degli errori severi sul mantenimento del significato originario. Al tempo stesso, la parziale sovrapposizione tra le distribuzioni dei livelli *lieve*, *medio* e *grave* evidenzia che non tutti gli errori classificati come gravi producono un deterioramento semantico marcato, mentre alcuni errori minori possono comunque incidere in modo significativo. Tale fenomeno evidenzia margini interpretativi utili per un affinamento della classificazione degli errori e delle metriche di similarità, al fine di cogliere con maggiore precisione le sfumature linguistiche e contestuali che influenzano la conservazione del significato.

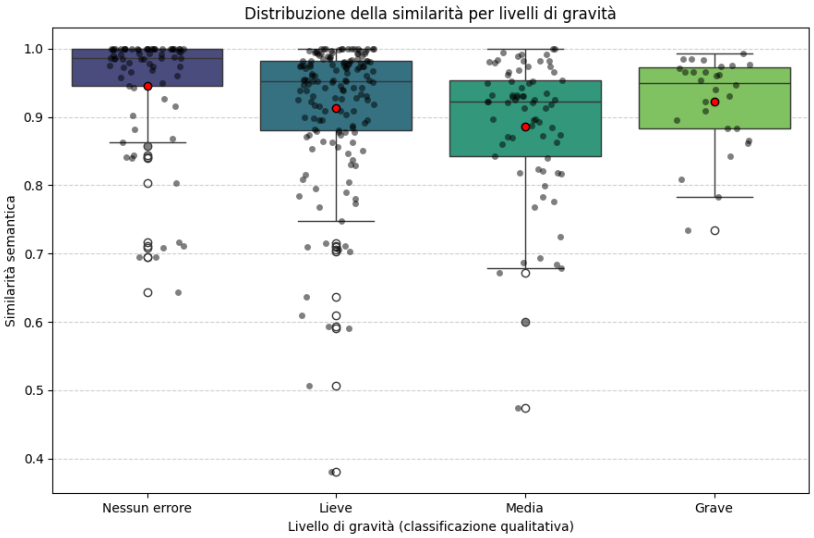

Figura 6.7: Boxplot della similarità semantica per livelli di gravità.

Nel complesso, l'analisi per la **RQ3** evidenzia una chiara relazione negativa tra la gravità degli errori e la similarità semantica delle trascrizioni, confermata dalla correlazione di Spearman, dal test di Kruskal–Wallis e dalla regressione robusta. I grafici delle Figure 6.6 e 6.7 mostrano come errori lievi o moderati tendano a preservare la coerenza semantica del testo, mentre errori di maggiore gravità ne compromettono la fedeltà in misura più consistente. La presenza di margini e sovrapposizioni tra i livelli di gravità suggerisce tuttavia che il fenomeno non sia rigidamente deterministico, ma dipenda da una combinazione di fattori strutturali e linguistici. Queste evidenze sottolineano non solo l'importanza di strategie di correzione mirate agli errori più gravi, ma anche la necessità di approcci più flessibili e graduati nella valutazione dell'impatto semantico, così da garantire una rappresentazione più accurata e affidabile del significato testuale.

6.4.4 Sintesi delle risposte alle domande di ricerca

In questa sezione si fornisce una sintesi dei risultati emersi dall'analisi presentata per le tre domande di ricerca, con l'obiettivo di chiarire le risposte finali e preparare il terreno per la discussione critica nel capitolo successivo.

Per quanto riguarda la **RQ1** (*quanto influisce la tipologia di errore sulla fedeltà della trascrizione?*), i risultati indicano che le differenze tra le categorie di errore fonetico (*toni*, *vocali*, *consonanti*) non sono statisticamente significative. Il modello analizzato mostra quindi una robustezza relativamente uniforme rispetto ai diversi tipi di errore, suggerendo che inter-

venti di ottimizzazione mirati a categorie specifiche di errori fonetici non siano strettamente necessari.

In merito alla **RQ2** (*quanto incide la quantità di errori sulla precisione della trascrizione?*), l'analisi evidenzia una correlazione positiva significativa tra numero di errori presenti in ciascuna frase e *Character Error Rate*. Il modello mantiene prestazioni stabili fino a un certo livello di perturbazioni (fino a 7 errori per frase), oltre il quale si osserva un incremento marcato del tasso di errore, identificando un punto critico nella capacità del sistema di gestire errori multipli.

Infine, per la **RQ3** (*quanto gli errori di trascrizione alterano il significato semantico della frase?*), i risultati mostrano che, sebbene la presenza di errori fonetici possa ridurre leggermente la coerenza semantica delle trascrizioni, la maggior parte delle frasi conserva un significato sostanzialmente corretto. In particolare, le metriche di similarità semantica indicano che il modello riesce a preservare la comprensibilità del contenuto anche in presenza di errori ortografici o fonetici, confermando una certa resilienza dal punto di vista semantico.

Questa sintesi dei risultati fornisce dunque risposte chiare e quantitative alle tre domande di ricerca. Nel capitolo successivo, la **Discussione** approfondirà le implicazioni di questi risultati, valutando le motivazioni teoriche e pratiche alla base delle osservazioni, confrontandole con lavori precedenti e delineando possibili strategie di miglioramento del modello.

6.5 Discussione dei risultati

La discussione dei risultati ottenuti mira a integrare le evidenze quantitative emerse dalle analisi precedenti con una riflessione critica sui comportamenti osservati del sistema di riconoscimento vocale (ASR) impiegato. In particolare, si analizza l' **Eval Model**, corrispondente alla versione *fine-tuned* del modello `wav2vec2-large-xlsr-53-th`, una variante multilingue di Wav2Vec 2.0 ottimizzata per la lingua thailandese. Questo modello è stato ulteriormente addestrato su un corpus contenente pronunce con errori fonetici artificiali, generati intenzionalmente su tre dimensioni principali: toni, consonanti e vocali. Lo scopo del *fine-tuning* non è la correzione automatica, bensì l'apprendimento di trascrizioni fedeli alla pronuncia errata, ossia la capacità di rappresentare ciò che viene effettivamente pronunciato senza applicare correzioni o normalizzazioni linguistiche.

L'obiettivo generale di questa fase di studio è valutare fino a che punto il modello riesca a mantenere robustezza fonetica e coerenza semantica nella trascrizione di pronunce non

standard. A tale scopo sono state formulate tre domande di ricerca (RQ1–RQ3), che guidano la seguente discussione.

RQ1 – Impatto della tipologia di errore

I risultati relativi alla RQ1 indicano che la tipologia di errore fonetico (toni, consonanti o vocali) non produce differenze statisticamente significative nel Character Error Rate (CER). Ciò suggerisce che l' **Eval Model** (`wav2vec2-large-xlsr-53-th` *fine-tuned*) possiede una buona invarianza rispetto alla natura dell'errore, riuscendo a mantenere prestazioni stabili anche in presenza di alterazioni fonetiche di diversa natura.

Dal punto di vista linguistico, questo comportamento può essere interpretato come un effetto della rappresentazione acustico-fonetica appresa durante il *fine-tuning*: il modello sembra aver sviluppato una sensibilità generalizzata alle variazioni di pronuncia, piuttosto che a specifiche categorie di errore. Dal punto di vista computazionale, tale risultato conferma la validità dell'approccio adottato: l'esposizione del modello a un corpus contenente errori intenzionali ha permesso di migliorare la sua robustezza fonetica, ossia la capacità di rappresentare fedelmente pronunce deviate senza tentare di "correggerle".

Tuttavia, è importante sottolineare che l'assenza di significatività statistica non implica che gli errori di diversa natura abbiano lo stesso impatto comunicativo. Gli errori tonali, ad esempio, pur non aumentando significativamente il CER, possono avere effetti molto più marcati sul significato semantico percepito da un ascoltatore umano. Questo evidenzia un limite delle metriche puramente fonetiche e suggerisce la necessità di integrare valutazioni semantiche o percettive nelle fasi successive di analisi.

6.5.1 RQ2 – Impatto della quantità di errori

Per la RQ2, i risultati mostrano una forte correlazione positiva ($\rho = 0.891$, $p = 0.001$) tra numero di errori fonetici introdotti nella frase e tasso di errore di trascrizione (CER). Il modello mantiene prestazioni relativamente stabili fino a circa sette errori per frase, ma oltre questa soglia il CER cresce in modo esponenziale, suggerendo la presenza di una soglia critica di degradazione.

Questo comportamento è coerente con la natura del modello ASR *fine-tuned* (`wav2vec2-large-xlsr-53-th`): fino a un certo punto, il modello riesce a generalizzare sulla base delle regolarità acustiche apprese durante il *fine-tuning*, ma quando la quantità di errori supera una soglia, la frase perde coerenza acustico–fonetica rispetto ai pattern appresi, compromettendo la corretta decodifica.

Tuttavia, è importante osservare che l'aumento del CER nelle frasi con otto, nove e dieci errori non dipende unicamente dalla difficoltà intrinseca di tali casi, ma anche da una limitazione del campione disponibile. Queste categorie contengono un numero ridotto di esempi nel dataset, il che riduce la capacità del modello di adattarsi o di generalizzare efficacemente a tali configurazioni. In altre parole, il peggioramento delle prestazioni oltre la soglia dei sette errori riflette non solo una maggiore complessità acustica, ma anche una limitazione statistica dovuta alla scarsità di dati in quella fascia, che impedisce al modello di apprendere rappresentazioni affidabili per quei livelli di distorsione.

Dal punto di vista applicativo, questo risultato suggerisce che:

- il modello è tollerante a un numero moderato di errori, tipico delle pronunce di parlanti principianti o intermedi;
- per pronunce fortemente deviate (oltre otto errori per frase) sarebbe utile integrare moduli di supporto, come sistemi di *error recovery* o *forced alignment* fonetico Billa and Hermansky (2019); McAuliffe et al. (2017), per migliorare la stabilità della trascrizione;
- ulteriori esperimenti dovrebbero includere un campionamento più bilanciato del numero di errori, al fine di consentire al modello di apprendere in modo più omogeneo anche dai casi estremi.

Inoltre, la variabilità osservata per frasi con cinque–sette errori può essere legata non solo alla quantità di errori, ma anche alla posizione e alla distribuzione degli stessi all'interno della frase. Errori concentrati su sillabe toniche o in posizione iniziale tendono a disturbare maggiormente il riconoscimento rispetto a quelli in posizioni neutre o ridondanti. Questo aspetto rappresenta un punto di approfondimento utile per studi futuri sull'impatto locale e contestuale degli errori fonetici.

6.5.2 RQ3 – Similarità semantica e gravità degli errori

Per la RQ3, l'analisi ha confrontato la frase errata (cioè la versione pronunciata e usata come *label* durante l'addestramento del modello) con la sua versione corretta, calcolandone la similarità semantica in funzione del numero di errori introdotti. Questa misura consente di valutare fino a che punto la frase con errori mantenga il suo significato originario, indipendentemente dalla correttezza fonetica.

I risultati mostrano che, anche in presenza di un numero moderato di errori (fino a sette), la similarità semantica rimane elevata, con valori medi tra 0.75 e 0.85. Solo oltre questa soglia

si osserva un calo più marcato (similarità < 0.6), che corrisponde ai casi di gravità elevata, in cui le distorsioni fonetiche compromettono la comprensione del contenuto.

Questo comportamento suggerisce che l'Eval Model, pur essendo addestrato per trascrivere fedelmente gli errori, preserva in larga misura la coerenza semantica delle frasi. Ciò è probabilmente dovuto alla capacità dell'encoder Wav2Vec2 di sfruttare contesti acustici più ampi per inferire la struttura lessicale della frase, anche quando singole unità fonetiche risultano alterate.

Da un punto di vista applicativo, questi risultati sono rilevanti per i sistemi di feedback automatico sulla pronuncia, in quanto mostrano che un ASR addestrato in questo modo è in grado di fornire trascrizioni fedeli ma semanticamente coerenti, consentendo di stimare la gravità dell'errore non solo in termini fonetici ma anche semantici.

6.5.3 Sintesi e implicazioni complessive

La discussione complessiva dei risultati mette in evidenza alcuni aspetti centrali del comportamento del modello e delle implicazioni metodologiche del presente studio. Innanzitutto, l'Eval Model (`wav2vec2-large-xlsr-53-th` *fine-tuned*) ha dimostrato una notevole robustezza rispetto alla tipologia di errore fonetico introdotto, segno che il processo di *fine-tuning* basato su un corpus contenente pronunce deviate ha consentito al sistema di sviluppare una rappresentazione fonetica generalizzata. Il modello, infatti, non mostra variazioni significative di prestazione tra errori di tono, consonante o vocale, evidenziando la capacità di adattarsi a diverse forme di distorsione acustica senza degradare drasticamente la qualità della trascrizione.

Un secondo elemento rilevante riguarda la relazione tra quantità di errori e prestazioni. Il modello mantiene un livello di accuratezza accettabile fino a circa sette errori per frase, soglia oltre la quale il tasso di errore di trascrizione (CER) aumenta in modo più marcato. Questo comportamento riflette l'esistenza di un limite di tolleranza oltre il quale la quantità di alterazioni fonetiche supera la capacità del modello di generalizzare rispetto ai pattern acustici appresi. Va tuttavia sottolineato che l'incremento del CER osservato nelle frasi con otto o più errori è in parte attribuibile alla scarsità di esempi in tale fascia, che riduce la possibilità per il modello di apprendere rappresentazioni robuste e statisticamente affidabili. In prospettiva, un bilanciamento maggiore del dataset e l'integrazione di strumenti di supporto come moduli di *error recovery* o *forced alignment* fonetico Billa and Hermansky (2019); McAuliffe et al. (2017) potrebbero contribuire a migliorare la stabilità e la coerenza delle trascrizioni in presenza di pronunce fortemente deviate.

Infine, l'analisi condotta sulla similarità semantica tra le frasi errate e le loro versioni corrette mostra come, nonostante le deviazioni fonetiche, il modello riesca a preservare in buona misura il contenuto semantico delle frasi. Anche con un numero moderato di errori, la similarità semantica rimane elevata, indicando che l'encoder di Wav2Vec2 sfrutta efficacemente il contesto acustico e linguistico per mantenere la coerenza globale del messaggio. Questo risultato suggerisce che un modello ASR addestrato in modo da riprodurre fedelmente la pronuncia errata può, al tempo stesso, conservare informazioni semantiche utili per la valutazione qualitativa degli errori.

Dal punto di vista metodologico, i risultati ottenuti confermano l'efficacia dell'approccio di injection controllata di errori fonetici nel *fine-tuning* di modelli ASR destinati all'ambito didattico. L'esposizione del modello a dati imperfetti lo rende più realistico e adattivo, poiché consente di gestire in maniera naturale le pronunce effettive dei parlanti non madrelingua senza applicare processi di normalizzazione automatica. L'integrazione di analisi fonetiche e semantiche rappresenta inoltre un passo importante verso sistemi di riconoscimento vocale capaci non solo di trascrivere il parlato, ma anche di stimare la gravità e l'impatto semantico degli errori di pronuncia. In questa prospettiva, l'approccio adottato offre un contributo diretto allo sviluppo di strumenti intelligenti per il feedback automatico e il supporto all'apprendimento della lingua thai.

CAPITOLO 7

Conclusioni e Lavori futuri

Il presente capitolo ha lo scopo di riassumere i principali risultati ottenuti nel corso di questa ricerca, discutendo le implicazioni metodologiche ed empiriche emerse dallo studio condotto sui modelli di riconoscimento automatico del parlato per la lingua Thai. Dopo una sintesi complessiva del lavoro svolto e delle evidenze sperimentali più rilevanti, vengono delineate alcune prospettive di ricerca futura finalizzate a consolidare, estendere e approfondire i risultati raggiunti, sia dal punto di vista tecnico sia da quello linguistico.

7.1 Conclusioni

Il lavoro presentato in questa tesi ha proposto e valutato una pipeline metodologica integrata per l'analisi del parlato in lingua Thai, finalizzata alla produzione di trascrizioni foneticamente fedeli e alla valutazione dell'impatto degli errori di pronuncia sul significato. La pipeline combina tre componenti principali: (i) un modello di riconoscimento automatico del parlato (*Automatic Speech Recognition*, ASR) fine-tuned sviluppato in questo lavoro, basato sull'architettura *wav2vec2-large-xlsr-53-th* come modello di partenza; (ii) tecniche di tokenization e normalizzazione del testo (Phatthiyaphaibun et al., 2023); e (iii) un componente di analisi semantica basata sul modello `paraphrase-multilingual-MiniLM-L12-v2` (Mahmoud et al., 2025), che utilizza rappresentazioni distribuzionali di frasi per misurare la similarità semantica tra trascrizione e riferimento.

Il dataset utilizzato per l'addestramento e la valutazione del modello è stato costruito a partire dal corpus LOTUS (Sertsi et al., 2016), successivamente arricchito con frasi contenenti errori fonetici iniettati in modo controllato. Il corpus complessivo è costituito da **3.255 esempi audio-testo**, suddivisi secondo una proporzione standard: **80% per il training set**, **10% per il validation set** e **10% per il test set**. Tale suddivisione ha consentito di mantenere un equilibrio tra la fase di apprendimento e quella di verifica sperimentale. Tutte le analisi quantitative e qualitative condotte per rispondere alle tre domande di ricerca definite nel lavoro (**RQ1**, **RQ2** e **RQ3**) sono state eseguite esclusivamente sui dati del *test set*, al fine di garantire l'indipendenza dei risultati rispetto ai dati utilizzati per l'addestramento.

I risultati sperimentali mostrano che il modello proposto mantiene un buon livello di robustezza rispetto alla natura dell'errore. L'analisi statistica del *Character Error Rate* (CER) (Baevski et al., 2020) non ha infatti evidenziato differenze significative tra le tre principali categorie di errore fonetico (toni, vocali e consonanti), suggerendo che il sistema è in grado di gestire in modo equilibrato le diverse fonti di variabilità fonetica. Inoltre, è stata osservata una correlazione positiva tra il numero di errori iniettati e il CER, con una soglia di stabilità fino a circa sette errori per frase, oltre la quale le prestazioni del modello degradano in modo più marcato. Tale risultato consente di identificare un limite pratico oltre il quale la capacità di trascrizione accurata viene compromessa.

Dal punto di vista semantico, la similarità tra la trascrizione prodotta e la frase di riferimento rimane elevata per errori di gravità lieve o moderata, con valori medi compresi tra 0.75 e 0.85, mentre tende a diminuire sensibilmente per livelli di errore più gravi. L'analisi statistica condotta mediante correlazioni di *Spearman*, test di *Kruskal–Wallis* e regressione robusta conferma che la pipeline è in grado di preservare, entro certi limiti, la comprensibilità e il contenuto informativo del parlato, anche in presenza di distorsioni fonetiche.

Questi risultati complessivi confermano la validità dell'approccio adottato: il fine-tuning del modello di base su dati contenenti pronunce deviate consente di ottenere trascrizioni più aderenti alla pronuncia effettiva senza sacrificare eccessivamente la coerenza semantica. Le scelte implementative, tra cui la configurazione di addestramento con venti epoche, un *learning rate* pari a 1×10^{-4} e una dimensione di batch di 16, hanno contribuito a ottimizzare l'equilibrio tra accuratezza fonetica e preservazione del significato.

Nonostante i risultati promettenti, alcune limitazioni meritano di essere sottolineate. In primo luogo, il dataset utilizzato si basa su errori fonetici iniettati artificialmente e su un numero limitato di registrazioni di test; pertanto, sarà necessario validare i risultati su dati reali prodotti da parlanti L2 di diversa provenienza e livello di competenza. In secondo luogo,

la valutazione delle prestazioni si è basata esclusivamente su metriche automatiche, come il CER e la similarità semantica calcolata tramite *cosine similarity* (Mahmoud et al., 2025). Sarà quindi importante confrontare tali misure con giudizi umani per verificare la corrispondenza tra errore automatico e percezione fonetica. Infine, il modello è stato valutato unicamente su Thai standard; la sua generalizzazione a varietà dialettali e a condizioni acustiche più complesse richiederà ulteriori indagini sperimentali.

7.2 Lavori futuri

I risultati ottenuti aprono diverse prospettive di ricerca per approfondire e consolidare il lavoro svolto. Una prima direzione riguarda **la raccolta e l'integrazione di dati reali provenienti da studenti L2**. L'utilizzo di pronunce autentiche consentirà di validare la robustezza del modello in scenari naturali e di verificare la capacità di generalizzazione rispetto a variabilità individuali, prosodiche e articolatorie. Tale estensione permetterà inoltre di analizzare la distribuzione effettiva degli errori di pronuncia, fornendo un quadro più realistico delle difficoltà incontrate dai parlanti non nativi e costituendo la base per strumenti avanzati di feedback linguistico.

Un ulteriore sviluppo riguarda **la validazione percettiva tramite giudizi umani**. Il confronto tra le metriche automatiche adottate e le valutazioni soggettive della comprensibilità del parlato costituirebbe un passo fondamentale per garantire che le misure proposte riflettano effettivamente la percezione umana delle differenze fonetiche e semantiche. Studi di questo tipo permetterebbero di calibrare meglio la soglia di significatività semantica rispetto al numero e alla gravità degli errori di pronuncia, fornendo indicazioni preziose per la progettazione di strumenti didattici personalizzati.

Sarà inoltre utile approfondire **l'analisi della localizzazione e della tipologia degli errori** per individuare quali posizioni o categorie fonetiche abbiano un impatto maggiore sulle prestazioni di riconoscimento e sulla comprensibilità. Questa analisi consentirebbe di comprendere se determinati contesti fonetici, come le consonanti finali o i toni in sillabe accentate, influenzino maggiormente l'accuratezza del modello e, di conseguenza, la progettazione di interventi correttivi mirati.

Un'altra linea di ricerca di rilievo riguarda **l'integrazione di caratteristiche prosodiche e tonali**. L'aggiunta di informazioni quali intonazione, durata e variazione di frequenza fondamentale potrebbe migliorare la sensibilità del modello agli errori di tono, particolarmente critici nella lingua Thai. L'inclusione di tali feature permetterebbe di estendere il sistema

verso una rappresentazione più ricca e multidimensionale del segnale vocale, fornendo dati utili anche per strumenti didattici basati su feedback prosodico.

Dal punto di vista computazionale, sarà interessante esplorare **strategie di ottimizzazione per l'inferenza efficiente**, come la distillation e la quantization del modello, al fine di ridurne la complessità e renderlo più adatto all'esecuzione in contesti a risorse limitate o in sistemi embedded. Ciò aprirebbe la possibilità di impieghi pratici senza compromettere significativamente la qualità del riconoscimento, rendendo i sistemi accessibili anche in contesti educativi o applicazioni mobile.

In coerenza con queste prospettive, risulta naturale considerare **lo sviluppo di strumenti integrati di supporto alla pronuncia per studenti L2**, che combinino la generazione automatica di audio con pronunce alterate, la trascrizione fonetica degli errori, l'analisi semantica e la valutazione mediante metriche come il CER. Un simile tool potrebbe offrire feedback interattivi e suggerimenti correttivi personalizzati, fungendo sia da strumento didattico concreto sia da piattaforma di test avanzata per monitorare e migliorare ulteriormente i modelli ASR per parlato non nativo. L'integrazione di dati reali e di giudizi percettivi umani all'interno dello strumento garantirebbe inoltre una validazione continua e una maggiore efficacia nell'apprendimento della pronuncia.

Un'ulteriore estensione riguarda **la valutazione del modello su varietà dialettali del Thai e su fenomeni di *code-switching***, che rappresentano situazioni comuni nel parlato spontaneo. Analizzare la capacità del sistema di adattarsi a tali contesti consentirebbe di testare la sua reale robustezza linguistica e la trasferibilità dell'approccio proposto.

Infine, un miglioramento significativo potrà derivare da **l'affinamento dei modelli di embedding semantico per la lingua Thai**. L'addestramento di modelli di rappresentazione del significato specificamente ottimizzati su coppie di frasi corrette ed errate permetterebbe di ottenere una misura di similarità semantica più aderente alla percezione linguistica dei parlanti nativi. Con una prospettiva più ampia, la pipeline sviluppata potrà essere applicata ad altre lingue tonali, come il vietnamita o il cantonese, per verificarne la generalizzabilità e l'efficacia in contesti linguistici affini.

In conclusione, la ricerca condotta ha mostrato come sia possibile adattare modelli ASR di ultima generazione alla lingua Thai e analizzare sistematicamente l'effetto degli errori di pronuncia sulla trascrizione e sulla comprensibilità. Le prospettive future delineate mirano a consolidare tali risultati, estendendone la validità empirica e metodologica, e a contribuire allo sviluppo di modelli di riconoscimento del parlato sempre più accurati, robusti e linguisticamente consapevoli.

Bibliografia

Megan Ardila, Gabriel Branson, Kelly Davis, Michael Henretty, Josh Kohler, Reuben Meyer, Lindsay Morais, Gregor Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2019. URL `https://arxiv.org/abs/1912.06670`. arXiv preprint, arXiv:1912.06670.

Zaw Htet Aung, Thanachot Thavornmongkol, Atirut Boribalburephan, Vittavas Tangsriworakan, Knot Pipatsrisawat, and Titipat Achakulvisut. Thonburian whisper: Robust fine-tuned and distilled whisper for thai. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 149–156, October 2024. URL `https://aclanthology.org/2024.icnlsp-1.17/`.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`. arXiv preprint, arXiv:2006.11477.

John Billa and Hynek Hermansky. Error handling in speech recognition systems: A review. *Computer Speech & Language*, 56:80–102, 2019. doi: 10.1016/j.csl.2018.12.003.

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1061–1072, 2008. URL `https://ieeexplore.ieee.org/document/4634091`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, 2017. URL `https://arxiv.org/abs/1607.04606`.

Thodsaporn Chay-intr and Manabu Okumura Hidetaka Kamigaito. Character-based thai word segmentation with multiple attentions. In *Proceedings of RANLP 2021*, pages 277–286, 2021. URL https://aclanthology.org/2021.ranlp-1.31.pdf.

Pattarawat Chormai, Ponrawee Prasertsom, and Attapol Rutherford. Attacut: A fast and accurate neural thai word segmenter. *arXiv preprint*, arXiv:1911.07056, 2019. URL https://arxiv.org/abs/1911.07056. Dilated CNN-based model with syllable embeddings for Thai word segmentation.

Ekapol Chuangsuwanich, Atiwong Suchato, Korrawe Karunratanakul, Burin Naowarat, Chompakorn Chaichot, Penpicha Sangsa-nga, Thunyathon Anutarases, Nitchakran Chaipojjana, and Yuatyong Chaichana. Gowajee corpus. Repository / technical report, version 0.9.3, Chulalongkorn University, 2020. URL https://github.com/ekapolc/gowajee_corpus.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *arXiv preprint arXiv:1911.02116*, 2019. URL https://arxiv.org/abs/1911.02116.

Pierre Nicolas Durette. gtts: Google text-to-speech python library, 2025. URL https://github.com/pndurette/gTTS. GitHub Repository.

Imad M. El-Emary and Fatma S. Khafaga. Hidden markov model and gaussian mixture model for speech recognition. In *2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 1–6, 2015. doi: 10.1109/I4CT.2015.7310138. URL https://d1wqtxts1xzle7.cloudfront.net/106646696/article1380631495_El-emary_et_al-libre.pdf.

Fuhao Feng and et al. Language-agnostic bert sentence embedding (labse). In *arXiv preprint arXiv:2007.01852*, 2020. URL https://arxiv.org/abs/2007.01852.

Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2008. doi: 10.1561/2000000004. URL https://ieeexplore.ieee.org/document/8187420.

Vishal Gupta, Kundan Krishna, Emma Strubell, and Andrew McCallum. Low-resource speech recognition with transfer learning and data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages

1234–1245, 2023. doi: 10.18653/v1/2023.acl-main.123. URL https://aclanthology.org/2023.acl-main.123/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Chutima Lowphansirikul and et al. Wangchanberta: Pre-trained language model for thai. In *arXiv preprint arXiv:2104.05155*, 2021. URL https://arxiv.org/abs/2104.05155.

Hatem Mahmoud, Waleed Alshangiti, Mohamed Othman, Ibrahim Alfarhood, and Abdulrahman Alarifi. Fine-tuning qursim on monolingual and multilingual models for semantic search. *Information*, 16(2):84, 2025. doi: 10.3390/info16020084. URL https://www.mdpi.com/2078-2489/16/2/84.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pages 498–502, 2017. URL https://www.isca-archive.org/interspeech_2017/mcauliffe17_interspeech.html.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*, 2013. URL https://arxiv.org/abs/1301.3781.

Mozilla Foundation / Common Voice team. Common voice dataset (release metadata, language hours). Dataset pages / release blogs (varie release: v9, v12, v19, ...). URL https://commonvoice.mozilla.org.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech, 2023. URL https://arxiv.org/abs/2305.19709. arXiv preprint, arXiv:2305.19709.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014. URL https://aclanthology.org/D14-1162/.

Wannaphong Phatthiyaphaibun, Chompakorn Chaksangchaichot, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. Thai wav2vec2.0 with commonvoice v8, 2022. URL `https://arxiv.org/abs/2208.04799`. arXiv preprint, arXiv:2208.04799.

Wannaphong Phatthiyaphaibun, Ekapol Chaovavanich, Charin Polpanumas, Lalita Lowphansirikul, Pattarawat Chormai, and Peerat Limkonchotiwat. Pythainlp: Thai natural language processing in python, 2023. URL `https://arxiv.org/abs/2312.04649`. arXiv preprint, arXiv:2312.04649.

Siripong Potisuk. Acoustic description of successive non-identical nasal sounds for automatic segmentation of continuous thai speech. In *Proceedings of the Speech Science and Technology Conference (SST)*, 2010. URL `https://assta.org/proceedings/sst/SST-10/SST2010/PDF/AUTHOR/ST100031.PDF`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`. arXiv preprint, arXiv:2212.04356.

Titipat Achakulvisut Rakpong Kittinaradorn, Kittinan Srithaworn Korakot Chaovavanich, Chanwit Kaewkasi Pattarawat Chormai, and Krichkorn Oparad Tulakan Ruangrong. DeepCut: A Thai word tokenization library using Deep Neural Network, September 2019. URL `http://doi.org/10.5281/zenodo.3457707`.

Markéta Řezáčková, Jan Švec, and Daniel Tihelka. T5g2p: Using text-to-text transfer transformer for grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2021*, pages 2062–2066, 2021. URL `https://www.isca-archive.org/interspeech_2021/rezackova21_interspeech.pdf`.

Phuttapong Sertsi, Vataya Chunwijitra, et al. A cloud-based framework for thai large vocabulary speech recognition, 2016. survey / conference paper che riporta dettagli su LOTUS e sottocorpus; v. articoli e resoconti tecnici NECTEC.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904. URL `https://www.jstor.org/stable/pdf/1412159.pdf`.

Jalinee Sriwirote, Vasan Thapiang, and Rutherford Timtong. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords, 2023. URL `https://arxiv.org/abs/2311.12475`. arXiv preprint, arXiv:2311.12475.

Panyut Sriwirote and et al. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords. In *arXiv preprint arXiv:2311.12475*, 2023. URL `https://arxiv.org/abs/2311.12475`.

Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *Proceedings of Interspeech 2023*, pages 4069–4073, 2023. doi: 10.21437/Interspeech.2023-1828. URL `https://www.isca-speech.org/archive/interspeech_2023/suwanbandit23_interspeech.html`.

Sumonmas Thatphithakkul, Kwanchiva Thangthai, and Vataya Chunwijitra. The development of lotus-trd: A thai regional dialect speech corpus. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6, 2024. doi: 10.1109/O-COCOSDA64382.2024.10800335.

Wannaphong. Khanomtan tts: Open-source thai text-to-speech model, 2022. URL `https://github.com/wannaphong/KhanomTan-TTS-v1.0`. GitHub Repository.

Chai Wutiwiwatchai and Sadaoki Furui. Thai speech processing technology: A review. *Computer Speech & Language*, 20(4):694–716, 2006. URL `https://www.sciencedirect.com/science/article/abs/pii/S0167639306001397`.

Chai Wutiwiwatchai and Chalermpol Hansakunbuntheung. Analysis of segmental duration for thai speech synthesis. In *Proceedings of Speech Prosody*, pages 447–450, 2004. URL `https://www.isca-archive.org/speechprosody_2004/hansakunbuntheung04_speechprosody.pdf`. Introduces the Thai Speech Synthesis Corpus (TSynC-1), later extended to TSynC-2 by NECTEC for TTS research.

Jian Zhu, Cong Zhang, and David Jurgens. Charsiug2p: A multilingual grapheme-to-phoneme toolkit. GitHub repository, 2022. URL `https://github.com/lingjzhu/CharsiuG2P`.