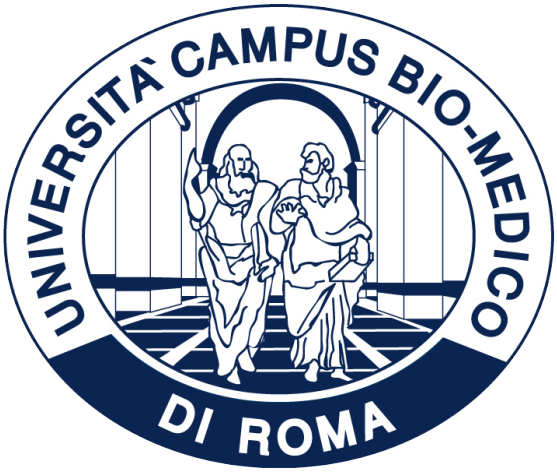
Matr. N 20100



UNIVERSITA'
CAMPUS BIO-MEDICO DI ROMA

FACOLTA' DIPARTIMENTALE DI INGEGNERIA
CORSO DI LAUREA MAGISTRALE IN INGEGNERIA
DEI SISTEMI INTELLIGENTI

DALLA LINGUA DEI SEGNI ALLA VOCE: SVILUPPO DI UNA PIPELINE PER LA SINTESI DEL PARLATO ESPRESSIVO

Relatore
Ing. Luca Bacco

Correlatori
Ing. Mario Merone
Dott. Daniele Sasso

Laureando
Ignazio Emanuele Piccichè

ANNO ACCADEMICO 2024/2025

"Sono una persona fortunata,
*ma la mia **vera fortuna** è*
*poterla **condividere**"*

Abstract

I. Introduzione

Storicamente la ricerca sulla Sign Language Translation (SLT) si è concentrata quasi esclusivamente sulla trascrizione automatica, trascurando l'espressività delle **Componenti Non-Manuali** (come *espressioni facciali e postura*) e il loro ruolo nella trasmissione naturale (tramite sintesi vocale) del contenuto semantico ed emotivo. L'adozione di tali tecnologie è inoltre ostacolata da requisiti computazionali spesso incompatibili con sistemi di produzione consumer. In questa tesi affrontiamo tali sfide effettuando un'analisi sperimentale per i modelli di SLT allo stato dell'arte, proponendo una pipeline modulare di classificazione del sentimento e la costruzione di un **dataset multimodale** per compensare la scarsità di risorse nel dominio, su cui valutiamo approcci uni- e multi-modali. Infine, proponiamo un modulo di **sintesi vocale** che integri una scala di sentimento, di cui ne valutiamo l'espressività con l'ausilio di due annotatori umani reclutati per questo studio.

II. Materiali e Metodi

In **Fig. 1** vengono illustrati sequenzialmente i moduli progettati per questa tesi.

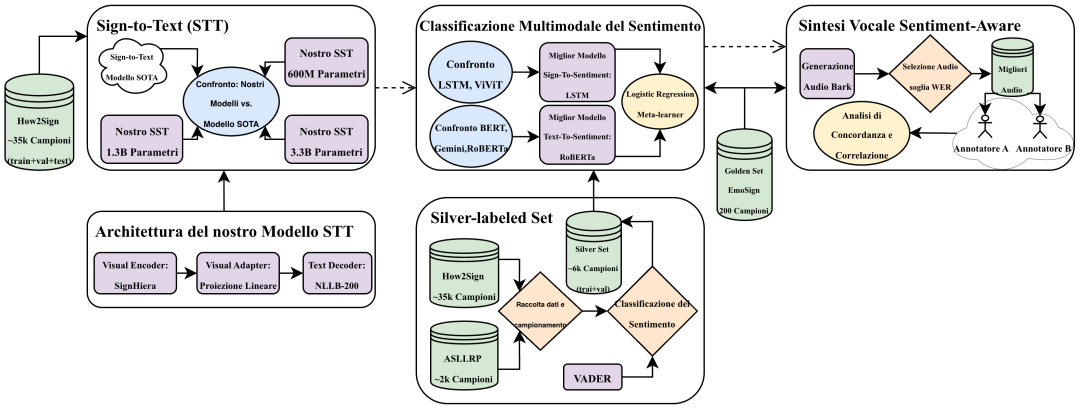


Figura 1: *Panoramica architetturale della Pipeline proposta*

Sign-to-Text Per il modulo di traduzione, addestrato su corpus **How2Sign [1]** (~35k campioni *frontal-view*), la metodologia si fonda sull'adattamento del modello multilingue **NLLB-200** al dominio visivo, seguendo l'approccio **SSVP-SLT [2] (64 GPU A100)** proposto da Meta, ad oggi stato dell'arte per tale task. Sono state confrontate diverse configurazioni dimensionali del modello (600M, 1.3B, 3.3B) e strategie di modellazione temporale, tra cui la modalità **SONAR Pooling**. Per l'addestramento (supportato da **una GPU A100**) sono stati adottati rigorosamente gli iperparametri di riferimento della letteratura. Il monitoraggio rigoroso degli esperimenti è stato affidato a **MLFlow**, tracciando puntualmente le curve di loss e le metriche di generazione (es. BLEU-1, ROUGE-L).

Classificazione Multimodale del Sentimento Il cuore del sistema è un'architettura di riconoscimento emotivo a fusione tardiva eseguita su hardware consumer (Apple Silicon M3). I modelli sono stati sviluppati su un dataset ibrido bilanciato di ~6k campioni di *train e validation* (da How2Sign [1] a ASLLRP [3]), raccolti anche tramite tecniche di web scraping ed etichettati automaticamente (con classi *positivo, neutrale, e negativo*) tramite VADER [4] (a seguito di calibrazione della soglia a 0.34) e testati su *Golden Set EmoSign* [5]. Il **ramo visivo** è frutto di una rigorosa selezione sperimentale che ha confrontato due paradigmi opposti di rappresentazione. Da un lato, un approccio *pose-based* leggero basato su LSTM, alimentato da landmark scheletrici estratti e normalizzati tramite MediaPipe [6]; dall'altro, un approccio *appearance-based* intensivo basato su Video Vision Transformer (ViViT [7]), che processa direttamente i pixel grezzi. Entrambe le architetture sono state sottoposte a *hyperparameter tuning* automatico tramite Grid Search per massimizzare il *Macro F1-Score*. Il **ramo testuale** è stato oggetto di un'analoga analisi comparativa volta a identificare il miglior interprete semantico per le trascrizioni. Sono stati valutati approcci con caratteristiche com-

plementari, contrapponendo la specializzazione sul task di *BERT-Base-Go-Emotion* e *RoBERTa* (*twitter-roberta-base-sentiment*) alla capacità di generalizzazione semantica di *Gemini 3.0 Pro*. I modelli (nelle due modalità) sono stati addestrati sul dataset di training costruito usando sia una *cross-entropy loss* (scelta classica per task di classificazione) che una *Focal loss* per mitigare lo sbilanciamento delle classi nel training set costruito (1178 positivi, 2636 neutri e 251 negativi). La convergenza delle modalità avviene nel **Meta-Learner**: i vettori di probabilità (3 valori a seguito di attivazione *softmax*) prodotti dal miglior modello visivo e testuale vengono concatenati e utilizzati come feature di input per un *classificatore di secondo livello*. Tramite una Grid Search sistematica su diversi algoritmi di Machine Learning, rispettando il vincolo di semplicità del modello di secondo livello, è stato identificato nella Logistic Regression il meta-modello ottimale per aggregare questi segnali eterogenei.

Sintesi Vocale *Sentiment-Aware* e Validazione Percettiva L'ultimo stadio impiega il modello Suno-ai Bark (su una GPU A100) per sonorizzare le caption del *Golden Set EmoSign*. Il processo di generazione arricchisce il testo iniettando **token acustici non verbali** (es. [laughs], [sighs]) selezionati in base all'etichetta emotiva originale e posizionati euristicamente in corrispondenza delle pause sintattiche per massimizzare la naturalezza. Per garantire l'intelligibilità, è stato applicato un filtro automatico basato su *Word Error Rate* (WER), trattenendo per l'analisi solo i campioni con elevata fedeltà testuale (soglia a 0.1). Il dataset filtrato è stato infine sottoposto a valutazione umana: **due annotatori** hanno classificato in modalità *blind* e randomizzata la coerenza emotiva su scala Likert [-3, +3], con un meccanismo di esclusione per i campioni affetti da artefatti audio critici. La solidità del processo di annotazione è stata verificata misurando l'accordo inter-annotatore, la coerenza interna e l'aderenza al Ground Truth originale, tramite il coefficiente *Weighted Kappa* e la correlazione di *Spearman*.

III. Risultati

Modulo Sign-To-Text La valutazione sul Test Set How2Sign evidenzia una differenza prestazionale tra le architetture. La variante **NLLB-3.3B** dimostra una superiore capacità di generalizzazione semantica, raggiungendo il picco di precisione in configurazione **Zero-shot** (BLEU-1 10.05). Di contro, il modello **1.3B** ha mostrato una maggiore recettività alle tecniche di adattamento (*Frozen Encoder*), ottenendo la migliore capacità di ricostruzione strutturale (ROUGE-L 10). Nel complesso, la configurazione **3.3B Zero-Shot** è stata identificata come il *Miglior Compromesso*, validando il modulo come *proof-of-concept* sostenibile su singola GPU, pur rimanendo distante dallo stato dell'arte (**53B**) addestrato su **64 GPU**.

Modello	Pesi	BLEU-1	ROUGE-L	Note
NLLB	1.3B	2.08 / 7.35 / 7.1	7.33 / **10** / 9.12	**Miglior ROUGE-L**
NLLB	3.3B	**10.05** / 6.69 / 7.39	9.24 / 8.12 / 9.27	**Miglior Zero-Shot**
SVVP-SLT	*53B*	*- / - / 30.2*	*- / - / 25.7*	*SOTA (Meta)*

Tabella 1: Sintesi metriche Sign-to-text (Test Set How2Sign) (Risultati per Configurazione: Zero-Shot / Frozen Encoder / Full Train)

Modulo Sign-to-Emotion (Analisi Visiva) Il benchmark condotto sul *Golden Set* di EmoSign (200 campioni) rivela criticità strutturali nelle architetture unimodali, caratterizzate da forti bias verso la classe neutra (come nel caso del Large Language Model **Qwen 2.5** in zero-shot) e negativa, come la LSTM, sebbene quest'ultima registri metriche più alte anche dello stato dell'arte (**Tab. 2**). Al contrario, il **ViViT**, pur con punteggi assoluti inferiori, mostra una distribuzione più bilanciata.

Modello	Configurazione	wAcc	wF1	Note
LSTM	Custom	**0.480**	**0.424**	**Bias**
ViViT	Fine-Tuned	0.330	0.342	**Bilanciato**
Qwen 2.5	Zero-Shot	0.105	0.059	Bias Neutro
MiniGPT	*SOTA Ref*	*0.346*	*0.400*	*Rif. EmoSign*

Tabella 2: *Benchmark Sign-to-Emotion (Scenario Ternario)*

Modulo Text-to-Emotion (Analisi Semantica) L'analisi del canale testuale, derivato dalla traduzione, conferma la maggiore robustezza dei segnali semantici rispetto a quelli puramente visivi osservata in letteratura [5]. Come mostra la **Tab. 3**, **RoBERTa Fine-Tuned** si impone come *Miglior Modello* assoluto, superando la versione pre-addestrata e dimostrando un'eccellente separazione delle classi, superando le ottime prestazioni *zero-shot* di *Gemini 3.0 Pro*, attualmente il miglior modello di reasoning di Google, che non è però possibile addestrare sul dominio e include un costo di chiamata alle API.

Modello	Configurazione	wAcc	wF1	Kappa	Note
Bert Go	Pre-trained	0.372	0.489	0.201	Modello debole
RoBERTa	Pre-trained	0.578	0.772	0.489	Modello medio
RoBERTa	Fine-tuned	**0.586**	**0.848**	**0.704**	**Migliore**
Gemini 3.0 Pro	*Zero-Shot*	*0.611*	*0.815*	*0.560*	*Best Zero-Shot*

Tabella 3: *Benchmark Text-to-Emotion*

Architettura Multimodale (Meta-Learner) Validata tramite *Grid Search*, la **Logistic Regression** si è rivelata il meta-modello ottimale contro una serie di altri validi modelli

di Machine Learning. Sul Test Set, il *meta-learner* proposto raggiunge un **wF1-Score** di 0.81 (**Tab. 4**), superando sia il riferimento generalista allo stato dell'arte (GPT-4o) sia modelli specializzati come AffectGPT. Il risultato conferma la superiorità di una fusione modulare mirata rispetto ai *Large Multimodal Models*

Modello	**Architettura**	**wAcc**	**wF1**	**Note**
Nostro Modello	Logistic Regression	**0.5395**	**0.8091**	**Best Multimodal**
GPT-4o	*LLM Generalist*	*0.5213*	*0.7672*	*SOTA EmoSign*
AffectGPT	*LLM Specialized*	*0.5618*	*0.6437*	*EmoSign Ref.*
Qwen 2.5	*LLM Generalist*	*0.4110*	*0.5429*	*EmoSign Ref.*

Tabella 4: *Confronto SOTA (Modalità Video + Testo)*

Validazione Percettiva degli audio Sentiment-Aware Sulle 200 coppie del golden set, 107 sono risultate di elevata fedeltà (soglia WER¡=0.1). È stato selezionato casualmente un pool iniziale di **54 coppie** per annotatore: da questi, gli annotatori hanno rimosso 9 file audio ritenuti non adatti, definendo un dataset di analisi di **45 coppie valide**. L'analisi, dettagliata in **Tab. 5**, rivela tre evidenze principali. In primo luogo, l'**accordo inter-annotatore globale** mostra una correlazione solida, sebbene si osservi una chiara discrepanza tra la modalità testuale, altamente convergente, e quella uditiva, soggetta a maggiore variabilità interpretativa. In secondo luogo, la **validità semantica** rispetto al *Ground Truth* è elevata, con correlazioni di *Spearman* (in media) superiori a **0.78**. Infine, la **coerenza interna** tra le valutazioni audio e testo dei singoli annotatori conferma che il sistema preserva efficacemente l'informazione affettiva durante la sintesi vocale.

Analisi	**Confronto**	**Spearman**	**wKappa**
Accordo Inter-Annotatore	Globale (Audio + Testo)	0.742	0.517
	Solo Testo	0.834	0.613
	Solo Audio	0.657	0.415
Annotatori vs Golden Set	Media Annotatori (Testo)	0.785	0.551
Coerenza Interna	Media (Audio vs Testo)	0.740	0.555

Tabella 5: *Sintesi delle metriche statistiche di validazione percettiva*

IV. Discussione e Conclusioni

L'analisi dei risultati offre una prospettiva stratificata sulle sfide della traduzione multimodale. Per quanto riguarda la **SLT**, emerge chiaramente come il dominio sia ancora

lontano dal raggiungere prestazioni operative accettabili: le criticità non riguardano solo l'usabilità dei modelli ottimizzati (necessaria per scenari *consumer*), ma si estendono anche ai macro-modelli *State-of-the-Art*, i quali, pur richiedendo risorse ingenti, mostrano metriche BLEU appena sufficienti. Nel **riconoscimento del sentimento**, il contributo più significativo risiede nella validazione della pipeline di creazione del dataset *silver-labeled*, risorsa strategica in cui la ridondanza dell'approccio **multimodale** si conferma determinante per compensare le lacune della traduzione. Sebbene la scarsità di dati attuali favorisca le performance delle LSTM, il modello ViViT ha evidenziato una distribuzione delle classi più bilanciata e robusta; tale caratteristica lo rende l'architettura tecnologicamente più promettente qualora si disponga in futuro di volumi di dati maggiori. Infine, lo studio sulla **sintesi vocale** costituisce un elemento di novità in letteratura per la metodologia adottata. I risultati ottenuti risultano molto promettenti per l'utilizzo di sistemi di generazione di voce espressiva in questo e in altri contesti, dimostrando che il condizionamento del sentimento supera efficacemente la scarsa espressività dei sistemi standard. Sviluppi futuri mireranno prioritariamente all'espansione del corpus dati per sbloccare il potenziale delle architetture Transformer e all'ottimizzazione dell'inferenza per scenari *real-time*.

Bibliografia

[1] A. Duarte et al., *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language*, 2021.

[2] Phillip R. et al., *Towards Privacy-Aware Sign Language Translation at Scale*, 2024.

[3] C. Neidle et al., *ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP)*, 2022.

[4] Hutto, C.J. et al., *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, 2014.

[5] P. Chua, C. M. Fang et al., *EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language*, 2025.

[6] Camillo L. et al., *MediaPipe: A Framework for Building Perception Pipelines*, 2019.

[7] Anurag Arnab et al., *ViViT: A Video Vision Transformer*, 2021.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

1.1 Contesto: Oltre la Traduzione Letterale delle Lingue dei Segni

La comunicazione tra persone sorde e udenti rappresenta una sfida sociale e tecnologica di primaria importanza. Negli ultimi anni, i progressi della visione artificiale e del deep learning hanno accelerato lo sviluppo di sistemi di Riconoscimento (SLR) e Traduzione (SLT) della Lingua dei Segni, aprendo prospettive concrete per servizi di *sign-to-text* e *text-to-sign* fruibili in contesti reali.

Tuttavia, una traduzione fedele richiede di andare oltre la semplice decodifica dei gesti manuali. Le Componenti Non-Manuali (Non-Manual Features, NMF) — quali espressioni facciali, postura, dinamica del capo e prosodia visiva — svolgono un ruolo cruciale nella grammatica, nella pragmatica e, soprattutto, nell'espressione emotiva, influenzando in modo determinante l'interpretazione del messaggio. Ignorare questa dimensione affettiva significa degradare la fedeltà semantica della traduzione e compromettere l'esperienza d'uso, specialmente in applicazioni assistive: un sistema che traduce una frase segnata con rabbia nello stesso modo in cui traduce una frase gioiosa non può definirsi completo.

In questo scenario, la capacità di modellare segnali multimodali in modo robusto ed efficiente è divenuta il fulcro della ricerca contemporanea [1], [2]. Abilitare una comunicazione ricca richiede di affrontare non un singolo problema, ma un ecosistema di sfide interconnesse che spaziano dalla comprensione del contenuto all'interpretazione del sentimento, fino alla generazione di un output espressivo.

1.2 Il Problema: Gap Tecnologici nella Comunicazione Aumentata

Nonostante l'importanza delle componenti non-manuali sia ampiamente riconosciuta, la ricerca si è storicamente concentrata sulla **decodifica puramente lessicale**, come il riconoscimento di segni isolati (ISLR) o la traduzione di frasi continue (CSLR). Per abilitare una filiera tecnologica completa, tuttavia, è necessario colmare tre gap interconnessi che vanno oltre la semplice trasposizione vocabolo-per-vocabolo.

La prima sfida fondamentale è la **traduzione robusta del contenuto semantico** (*Sign-to-Text*). Questo compito rimane di enorme complessità a causa della variabilità linguistica intrinseca e della necessità di vasti dataset per addestrare modelli capaci di generalizzare su frasi mai viste, garantendo una struttura sintattica coerente nella lingua di arrivo.

Tuttavia, decodificare il testo è condizione necessaria ma non sufficiente: il contenuto semantico da solo non esaurisce il messaggio. Qui emerge la seconda criticità: il **riconoscimento autentico del sentimento**. La recente pubblicazione di dataset come EmoSign [3] ha messo in luce un divario significativo: i modelli attuali mostrano performance scarse basandosi solo sui segnali visivi, diventando efficaci solo quando hanno accesso alle trascrizioni. Ciò suggerisce una mancanza di reale comprensione della prosodia visiva e una dipendenza da "scorciatoie" linguistiche. La sfida è dunque creare modelli capaci di apprendere l'intensità emotiva direttamente dai segnali visivi (NMF), indipendentemente dal testo tradotto.

Infine, l'ultimo anello mancante è la generazione di un output accessibile e naturale. Un sistema di sintesi vocale (*Text-to-Speech*) ideale non dovrebbe limitarsi a leggere asetticamente il testo tradotto, ma dovrebbe essere capace di **modularne l'audio** per riflettere l'intensità del Sentiment originale del segnante, superando la monotonia dei sintetizzatori tradizionali.

Questa tesi si colloca all'intersezione di queste tre sfide, affrontandole come aree di ricerca distinte ma sinergiche, utilizzando l'American Sign Language (ASL) come dominio di studio principale grazie alla disponibilità di risorse.

1.3 Obiettivi, Domande di Ricerca e Contributi

Per rispondere alle criticità emerse, il presente lavoro non si limita a sviluppare una singola pipeline monolitica, ma esplora tre percorsi di ricerca paralleli, i cui contributi confluiscono in un sistema unificato. Come illustrato in Figura 1, il sistema è concepito

come una cascata modulare che trasforma il segnale video grezzo in un output vocale espressivo.

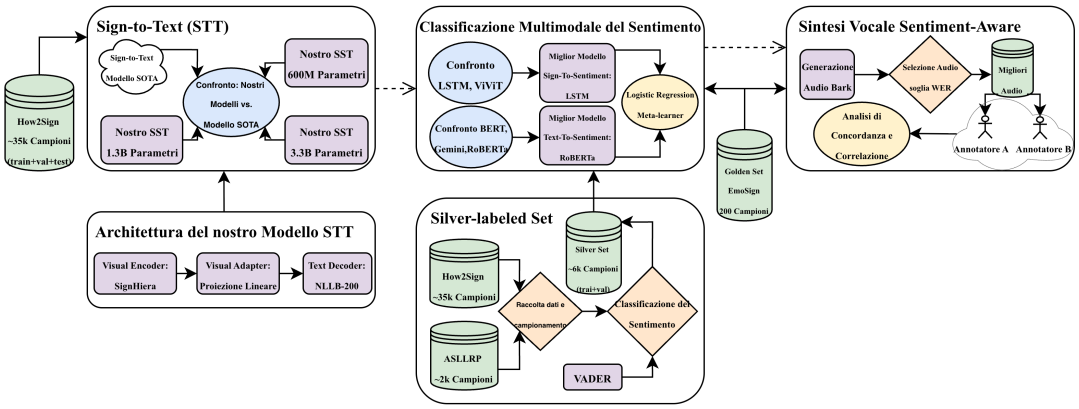


Figura 1: Panoramica architetturale della Pipeline proposta: dal video in input alla sintesi vocale validata percettivamente, passando per il riconoscimento multimodale del sentimento.

Il primo modulo costituisce la **base semantica** (*Sign-to-Text*). L'obiettivo è sviluppare un sistema di traduzione sequence-to-sequence per l'ASL, affrontando la domanda di ricerca su quali siano le prestazioni e i limiti nell'implementare un modello di traduzione utilizzando dataset pubblici (come How2Sign) in contesti a risorse vincolate. Il contributo principale in questa fase è la realizzazione di una baseline che funge da *proof-of-concept* per la validazione semantica dell'intera pipeline.

Un contributo trasversale e fondamentale per le fasi successive riguarda l' **ingegneria dei dati**. Data la scarsità di risorse annotate per il sentimento, parte integrante del lavoro è stata la costruzione di una pipeline di acquisizione e etichettatura automatica. Attraverso tecniche di web scraping e weak-labeling, è stato possibile aggregare e processare dataset eterogenei per creare un corpus *silver-labeled* essenziale per l'addestramento dei modelli supervisionati proposti.

Il secondo modulo, definito il **cuore affettivo** (*Multimodal Sign-to-Sentiment*), rappresenta il nucleo algoritmico della tesi. Superando l'approccio classico unimodale, l'obiettivo è sviluppare un'architettura di *Ensemble Learning* che unisca uno stream visivo — confrontando approcci *pose-based* e *appearance-based* — e uno stream testuale, fusi da un Meta-Learner. La ricerca indaga in che misura l'integrazione tardiva (*Late Fusion*) delle predizioni visive e testuali migliori l'accuratezza rispetto ai modelli unimodali, dimostrando quantitativamente il vantaggio della sinergia tra visione e linguaggio nel mitigare le ambiguità emotive.

Infine, il terzo modulo si occupa della **voce espressiva** (*Sentiment-Aware TTS*). L'obiettivo è realizzare un sistema di sintesi vocale generativa che utilizzi lo score di sentiment per modulare l'espressività della voce. La domanda di ricerca si focalizza

su come la mappatura dell'intensità del sentimento sui parametri del modello influenzi la percezione dell'ascoltatore e se l'audio generato introduca bias interpretativi. Il contributo finale è lo sviluppo di un algoritmo di post-processing per il controllo prosodico, validato tramite una campagna sperimentale con annotatori umani che ha confermato la coerenza interna tra l'audio sintetizzato e il *ground truth* del Sentiment.

1.4 Rilevanza e Struttura della Tesi

L'abilitazione di sistemi di mediazione tecnologica sensibili alla dimensione del sentimento ha un'immediata rilevanza scientifica e sociale, potendo migliorare drasticamente la qualità della traduzione automatica e l'accessibilità dei servizi digitali per le comunità Sorde. Un design attento a trasparenza, equità e co-progettazione è fondamentale per sviluppare tecnologie utili e rispettose [4].

La struttura dell'elaborato riflette l'approccio modulare del progetto, accompagnando il lettore attraverso le fasi logiche della ricerca. Il **Capitolo 2** analizza lo stato dell'arte, identificando i gap tecnologici attuali nella traduzione, nel riconoscimento del sentimento e nella sintesi vocale emotiva. Successivamente, il **Capitolo 3** descrive la metodologia e i materiali, spaziando dalla curatela dei dataset (ASLLRP, How2Sign) e le strategie di data engineering, alla progettazione delle architetture neurali per ciascuno dei tre moduli. Il **Capitolo 4** presenta i risultati sperimentali, offrendo una valutazione quantitativa e qualitativa delle performance di ogni sottosistema. Infine, il **Capitolo 5** propone una discussione critica delle scoperte fatte, delineando le implicazioni etiche e le prospettive future per l'integrazione di questi moduli in un sistema olistico di mediazione comunicativa.

Capitolo 2

Stato dell'arte

L'obiettivo di questo capitolo è analizzare la letteratura scientifica di riferimento per ciascuno dei tre sistemi sviluppati in questa tesi. Per riflettere la natura modulare del progetto, la trattazione non seguirà un percorso unificato, ma sarà organizzata in tre filoni di ricerca distinti, ciascuno dedicato a fornire il contesto teorico e tecnologico per uno dei moduli principali.

In prima istanza, si esaminerà lo stato dell'arte per la **Traduzione dalla Lingua dei Segni al Testo (SLT)**. Questa sezione esplorerà l'evoluzione storica del dominio, dalle prime architetture basate su gloss fino ai moderni modelli *end-to-end*, ponendo un focus specifico sull'avvento dei Transformer e sull'importanza dei dataset su larga scala.

Successivamente, l'analisi si sposterà sul **Riconoscimento del Sentimento nelle Lingue dei Segni**. Verranno approfondite le tecniche per l'interpretazione dei segnali non-manuali e analizzati i benchmark di riferimento attuali, come EmoSign, unitamente alle strategie di fusione multimodale per l'integrazione dei canali visivi e testuali.

La terza parte affronterà la letteratura relativa alla **Sintesi Vocale *Sentiment-Aware* (Emotional TTS)**, esplorando le strategie per controllare l'espressività del parlato sintetizzato, spaziando dai modelli acustici condizionati fino ai più recenti vocoder neurali. A conclusione del capitolo, verranno discusse le specificità dello scenario tecnologico per la Lingua dei Segni Italiana (LIS).

2.1 Dai Sistemi Basati su Sensori alla Visione Artificiale

I primi studi nel campo del *Sign Language Recognition* (SLR), risalenti agli anni '80 e '90, si affidavano prevalentemente a dispositivi hardware dedicati. Soluzioni basate su guanti dati (*data gloves*) e sistemi di tracciamento del movimento (come Leap Motion o Microsoft Kinect) permettevano di catturare con elevata precisione la

cinematica delle mani e delle braccia. Sebbene accurati, questi approcci presentavano limiti significativi per l'applicazione su larga scala: la loro natura invasiva e costosa alterava la naturalezza del gesto e la fluidità del linguaggio. Una limitazione ancora più critica, tuttavia, risiedeva nell'impossibilità di catturare le Componenti Non-Manuali (NMF) — quali espressioni facciali, movimenti del busto e direzione dello sguardo — che costituiscono elementi grammaticali e semantici imprescindibili nelle lingue dei segni [1].

La svolta paradigmatica è avvenuta con l'avvento dei sistemi *vision-based*. Sfruttando semplici videocamere RGB, questi approcci offrono un'interazione non invasiva e, soprattutto, permettono di catturare l'intera ricchezza espressiva del segnante nel suo ambiente naturale, aprendo la strada all'analisi computazionale olistica della performance linguistica [5].

2.2 Preprocessing e Rappresentazione dei Dati: Geometria vs. Apparenza

La scelta della rappresentazione dei dati in input è una fase cruciale che condiziona l'intera architettura a valle. In letteratura si è consolidata una dicotomia tra due filosofie principali di rappresentazione, che comportano vantaggi e svantaggi complementari in termini di efficienza, privacy e contenuto informativo.

Il primo paradigma, storicamente dominante per la sua efficienza computazionale, è l'**approccio Pose-based (Scheletrico)**. Questa metodologia si fonda su una pipeline a due stadi: inizialmente, modelli di *pose estimation* come **MediaPipe** [6] o **OpenPose** [7] estraggono frame per frame le coordinate dei punti chiave (*keypoints*) di mani, viso e corpo; successivamente, le sequenze vettoriali vengono processate da reti temporali. Oltre alla leggerezza computazionale, un vantaggio cruciale di questo approccio risiede nella *Privacy by Design*: astraendo il volto in una nuvola di punti, è possibile anonimizzare il segnante pur mantenendo l'informazione cinetica **Rust˙2024**. Tuttavia, tale astrazione comporta la perdita di dettagli di *texture* (rughe, intensità dello sguardo) fondamentali per le micro-espressioni.

In contrapposizione, l'**approccio Appearance-based (Pixel-based)** processa l'immagine nella sua interezza o tramite patch spazio-temporali. Sebbene storicamente limitato dall'alto costo computazionale delle 3D-CNN, l'avvento dei **Video Vision Transformers (ViViT)** [8] ha rinnovato l'interesse verso questa modalità. Analizzando direttamente i pixel grezzi, questi modelli riescono a catturare sfumature semantiche sottili e informazioni contestuali che la scheletrizzazione inevitabilmente scarta,

risultando particolarmente efficaci nel riconoscimento emotivo laddove la privacy non costituisca un vincolo stringente.

2.3 L'Impatto del Deep Learning nel Riconoscimento e nella Traduzione

Prima dell'avvento del Deep Learning, la modellazione statistica delle lingue dei segni era dominata dagli Hidden Markov Models (HMM) [2], efficaci nel gestire la variabilità temporale ma limitati nella capacità di estrarre feature visive complesse. L'introduzione delle architetture profonde ha rivoluzionato il campo, permettendo di affrontare con successo sia il riconoscimento di segni isolati (**ISLR**), sia la più complessa traduzione di frasi continue (**CSLR/T**), caratterizzata da fenomeni di co-articolazione.

L'evoluzione delle architetture riflette la necessità di modellare congiuntamente lo spazio e il tempo. Le **Reti Convoluzionali (CNN)** — e le loro varianti 3D come I3D o C3D — hanno costituito a lungo la spina dorsale per l'estrazione di feature spaziali, storicamente abbinate a **Reti Ricorrenti (RNN)** come LSTM e GRU per la modellazione sequenziale [2]. Recentemente, tuttavia, lo stato dell'arte si è spostato verso i **Transformer** e i modelli *attention-based*. Grazie al meccanismo di *self-attention*, queste reti superano i limiti delle RNN nel catturare dipendenze a lungo raggio all'interno della frase, gestendo meglio il contesto globale [9], [10].

Parallelamente, nel dominio scheletrico, l'esigenza di sfruttare la topologia naturale del corpo umano ha favorito l'adozione delle **Graph Neural Networks (GNNs)** [11], che modellano le articolazioni come nodi di un grafo dinamico, offrendo un'alternativa strutturata alle sequenze piatte processate dalle RNN. Questi progressi hanno abilitato lo sviluppo di moderni sistemi *end-to-end*, capaci di operare con latenze ridotte anche in scenari pratici [10].

2.4 Stato dell'Arte per la Traduzione in Testo (Sign-to-Text)

La *Sign Language Translation* (SLT) rappresenta una delle frontiere più complesse della mediazione linguistica automatica, ponendosi l'obiettivo di generare traduzioni testuali accurate e fluenti direttamente da video di lingua dei segni continua. A differenza del riconoscimento di segni isolati, questo compito deve risolvere le intricate sfide grammaticali, sintattiche e di co-articolazione tipiche del discorso naturale [2].

Storicamente, il problema è stato affrontato mediante un approccio a cascata basato su **gloss** (etichette testuali intermedie). Tale paradigma prevedeva un primo stadio di riconoscimento (*Sign-to-Gloss*) seguito da uno di traduzione (*Gloss-to-Text*). Sebbene questa scomposizione semplificasse il problema, soffriva intrinsecamente del fenomeno di **propagazione dell'errore**: un'imprecisione nel riconoscimento delle gloss comprometteva irrimediabilmente la traduzione finale. Per superare questo limite strutturale, la ricerca contemporanea ha abbracciato il paradigma **end-to-end**.

L'architettura dominante in questo contesto è il modello *sequence-to-sequence* basato su **Transformer** [9]. In tale configurazione, un **Encoder** visuale processa la sequenza di feature video creando una rappresentazione latente contestualizzata, mentre un **Decoder** genera la traduzione in modo autoregressivo. L'innovazione cruciale risiede nel meccanismo di *cross-attention*, che funge da ponte dinamico tra la modalità visiva e quella testuale. A differenza degli approcci precedenti che comprimevano l'intero video in un vettore statico, l'attenzione permette al decoder di allineare semanticamente il gesto alla parola, pesando selettivamente l'importanza dei frame video ad ogni passo della generazione. Questa capacità di allineamento temporale dinamico ha reso i modelli Transformer lo standard di riferimento, come dimostrato nel lavoro seminale *Neural Sign Language Translation* [12].

Tuttavia, l'efficacia di questi modelli *data-intensive* dipende strettamente dalla disponibilità di corpora annotati su larga scala, come **PHOENIX-Weather 2014T** [12] e **How2Sign** [10]. Le prestazioni sono valutate tramite metriche standard della traduzione automatica (BLEU, ROUGE), che quantificano la sovrapposizione di n-grammi con i riferimenti umani. Proprio la dipendenza da vasti dataset annotati e i rischi per la privacy derivanti dall'uso di volti riconoscibili rappresentano le limitazioni critiche dell'approccio supervisionato classico.

Per affrontare tali sfide, la letteratura recente ha introdotto un cambio di paradigma basato sull'**apprendimento auto-supervisionato** (*Self-Supervised Learning*). Un esempio significativo è il lavoro **SSVP-SLT** [13], che propone una metodologia in due fasi per sfruttare video non annotati. Inizialmente, il modello viene pre-addestrato su vasti corpora utilizzando tecniche di *Masked Autoencoding* (MAE), imparando a ricostruire porzioni spazio-temporali oscurate del video; questo forza la rete ad apprendere rappresentazioni robuste della dinamica segnica senza dipendere da etichette testuali. Solo successivamente il modello viene raffinato (*fine-tuning*) su dataset paralleli curati per il task di traduzione specifico. I risultati evidenziano che questo approccio non solo incrementa le performance sfruttando dati presenti nel web, ma offre anche una soluzione concreta per la privacy, validando l'efficacia dell'addestramento su video parzialmente anonimizzati.

2.4.1 Nuove Frontiere: Self-Supervised Learning e Privacy

L'approccio supervisionato classico presenta due limitazioni critiche: l'elevato costo di produzione di dataset annotati (video allineati al testo) e i rischi per la privacy derivanti dall'uso di volti riconoscibili, necessari per la grammatica non-manuale. Per affrontare tali sfide, il recente lavoro **SSVP-SLT** (*Self-Supervised Video Pretraining for Sign Language Translation*) [13] introduce un cambio di paradigma basato sull'apprendimento auto-supervisionato.

La metodologia si articola in due fasi distinte:

1. **Pre-training Auto-Supervisionato:** Il modello viene pre-addestrato su vasti corpora di video non annotati utilizzando tecniche di *Masked Autoencoding* (MAE). Il sistema impara a ricostruire porzioni spazio-temporali oscurate del video (inclusi i volti per l'anonimizzazione), forzando la rete ad apprendere rappresentazioni robuste della dinamica segnica senza dipendere da etichette testuali.

2. **Fine-tuning Supervisionato:** Solo successivamente il modello viene raffinato su dataset paralleli curati (come How2Sign) per il task di traduzione specifico.

I risultati evidenziano che questo approccio non solo incrementa le performance di traduzione sfruttando la mole di dati non etichettati presenti nel web, ma offre anche una soluzione concreta per la privacy, validando l'efficacia dell'addestramento su video parzialmente anonimizzati.

2.5 Il Riconoscimento delle Emozioni e del Sentimento nelle Lingue dei Segni

Parallelamente alla traduzione lessicale, la ricerca si è recentemente orientata verso l'analisi della componente affettiva. Un punto di svolta in questo ambito è rappresentato dalla pubblicazione del dataset **EmoSign** [3]. A differenza dei corpora precedenti, EmoSign costituisce il primo dataset multimodale su larga scala annotato specificamente per l'analisi delle emozioni da segnanti Sordi nativi, stabilendo un benchmark rigoroso che ha messo a nudo i limiti strutturali degli attuali Modelli Linguistici Multimodali (MLLMs).

Per comprendere la portata di questo benchmark, è utile analizzarne la metodologia di costruzione. Gli autori hanno selezionato 200 clip dal corpus ASLLRP [14], guidati dall'algoritmo VADER [15] per massimizzare la varianza emotiva agli estremi dello

spettro (positivo/negativo). L'etichettatura, affidata a segnanti nativi, include valutazioni su scala Likert e classificazioni *fine-grained*, evidenziando tuttavia un accordo inter-annotatore moderato (Krippendorff's Alpha 0.593) che riflette la soggettività intrinseca del task. La distribuzione delle classi, visibile in Figura 2, mostra le sfide di bilanciamento affrontate.

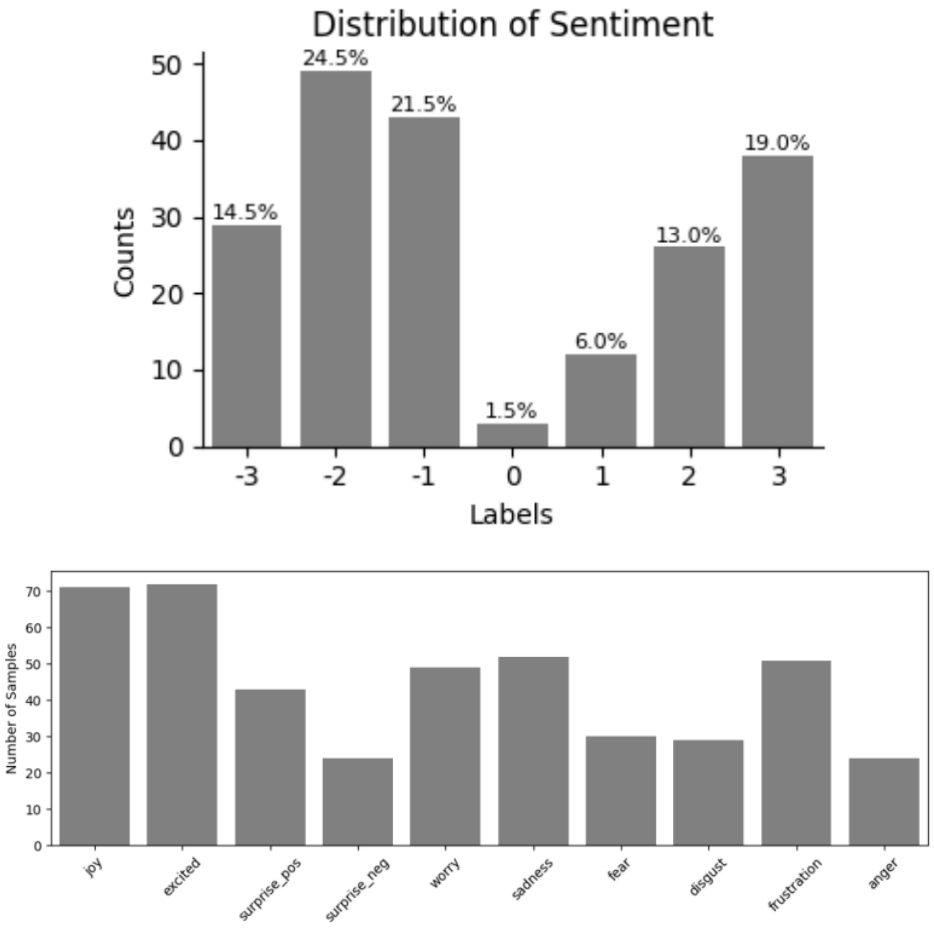


Figura 2: Sopra: distribuzione dei sentiment nel dataset EmoSign. Sotto: distribuzione delle categorie emotive binarizzate.

L'analisi delle performance dei modelli stato dell'arte (tra cui GPT-4o e AffectGPT) su questo dataset ha rivelato un fenomeno critico: il crollo delle capacità predittive in assenza del supporto testuale. Come evidenziato dai risultati riportati nelle Tabelle 1 e 2, nella condizione *video-only* anche il modello più avanzato (GPT-4o) ottiene un F1-score di appena 24.43% nel task di sentiment, risultato che sale drasticamente al 76.72% con l'aggiunta delle didascalie.

Modalità	**Modello**	**Sentiment (3-class)**		**Sentiment (7-class)**	
		wAcc	wF1	wAcc	wF1
video	MiniGPT4	34.68	**40.00**	14.46	**13.03**
	Qwen 2.5	27.34	16.47	10.26	2.44
	AffectGPT	33.33	0.04	14.29	0.04
	GPT-4o	**40.72**	24.43	**19.81**	5.97
video + *caption*	MiniGPT4	21.65	36.89	9.76	12.18
	Qwen 2.5	41.10	54.29	15.84	14.51
	AffectGPT	**56.18**	64.37	21.02	16.13
	GPT-4o	52.13	**76.72**	**22.89**	**26.35**

Tabella 1: Benchmark Sentiment Analysis su EmoSign (metriche: Weighted Accuracy e F1-score) [3].

Modalità	**Modello**	**HP**	**SP(P)**	**SP(N)**	**WR**	**SD**	**FR**	**DG**	**FS**	**AG**	NE	**Total**	**Total**
		Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	wAcc	wF1
video	MiniGPT4	**69**	**20**	14	0	0	0	0	0	0	27	13.01	**22.02**
	Qwen 2.5	35	0	**43**	0	0	0	0	5	**33**	27	**14.39**	18.53
	AffectGPT	11	0	0	**7**	**30**	0	0	5	0	**73**	12.62	11.03
	GPT-4o	35	0	0	**7**	0	0	**20**	**53**	0	0	11.50	20.76
video + *caption*	MiniGPT4	**89**	0	14	7	**30**	**43**	10	0	**33**	9	23.56	35.89
	Qwen 2.5	63	**40**	**29**	64	0	14	20	32	**33**	**55**	34.96	44.67
	AffectGPT	85	20	0	50	**30**	14	10	32	**33**	27	30.17	47.77
	GPT-4o	**89**	20	0	**79**	20	29	**50**	**74**	0	0	**35.97**	**55.09**

Tabella 2: Risultati del benchmark per la classificazione delle emozioni [3].

Approfondendo l'analisi qualitativa (*Emotion Cue Grounding*), emerge che i modelli spesso costruiscono spiegazioni *post-hoc* per giustificare un sentiment derivato dal testo, piuttosto che eseguire un'autentica comprensione visiva. Casi emblematici, come quello del modello Qwen che lamenta l'"assenza di audio" per giustificare l'incertezza, confermano una comprensione superficiale del dominio. Inoltre, si osserva un forte bias verso categorie emotive generiche e una frequente confusione tra stati visivamente simili ma semanticamente distinti, come "preoccupazione" (*worry*) e "tristezza" (*sadness*).

In sintesi, la letteratura evidenzia tre limitazioni strutturali che motivano la presente ricerca: un'estrema dipendenza dal testo che compensa le lacune visive; la difficoltà nel disambiguare la doppia funzione grammaticale ed emotiva delle espressioni facciali; e infine un marcato bias di generalizzazione che porta i modelli a collassare su categorie neutre o positive in assenza di input semantico esplicito.

2.5.1 Strategie di Fusione Multimodale

Per affrontare queste sfide e integrare efficacemente i flussi dati, la letteratura propone principalmente tre paradigmi di fusione. L'**Early Fusion** opera a livello di dati, aggregando le feature di basso livello in un'unica rappresentazione vettoriale prima

dell'elaborazione; questo metodo cattura correlazioni precoci ma richiede sincronizzazione temporale e aumenta la complessità dimensionale. Un approccio intermedio è la **Hybrid Fusion**, che introduce stadi di interazione tra le reti (come la co-attenzione) durante l'elaborazione. Infine, la **Late Fusion** — strategia adottata in questo lavoro — opera a livello decisionale: ogni modalità viene elaborata da modelli "esperti" indipendenti e le predizioni vengono combinate solo nello stadio finale. Ciò consente di sfruttare architetture specializzate per ogni tipo di dato (es. LSTM per lo scheletro, Transformer per il testo), gestendo con maggiore robustezza le discrepanze informative.

2.6 Stato dell'Arte per la Sintesi Espressiva: Voce e Animazione

Parallelamente ai progressi nel riconoscimento, la sintesi vocale (*Text-to-Speech*, TTS) ha subito un'evoluzione radicale con l'avvento del Deep Learning. Sebbene la qualità timbrica abbia raggiunto livelli di naturalezza quasi umani, la maggior parte dei sistemi commerciali produce ancora un parlato prosodicamente piatto ("neutro"), incapace di veicolare la pragmatica e le sfumature emotive essenziali nella comunicazione. La sintesi vocale emotiva (*Emotional TTS*) nasce per colmare questo divario, introducendo meccanismi di controllo esplicito sull'espressività dell'audio generato.

Architetturalmente, le pipeline neurali classiche si sono a lungo basate su una struttura a due stadi: un **modello acustico** (come **Tacotron 2** [16] o **FastSpeech 2** [17]) deputato a convertire il testo in una rappresentazione intermedia (tipicamente uno spettrogramma Mel), seguito da un **vocoder neurale** (es. **HiFi-GAN** [18]) per la sintesi della forma d'onda ad alta fedeltà. Recentemente, modelli come **VITS** [19] hanno superato questa dicotomia proponendo architetture *end-to-end* che migliorano l'efficienza e la coerenza del segnale.

Per infondere emozione in questo processo, la strategia dominante in letteratura è il condizionamento supervisionato. Utilizzando dataset etichettati emotivamente (es. *ESD Dataset* [20]), il modello viene addestrato a ricevere in input, oltre al testo, un **embedding emotivo** — un vettore latente che codifica lo stile target (gioia, tristezza, rabbia) — imparando ad associare a esso specifiche variazioni di intonazione e ritmo [21]. Approcci alternativi non supervisionati, come i **Global Style Tokens (GST)** [22], permettono di apprendere stili latenti senza etichette esplicite, offrendo maggiore flessibilità ma minore controllabilità. Il terzo modulo di questa tesi adotta il primo approccio, sfruttando il controllo esplicito per mappare il sentiment riconosciuto sulla prosodia vocale.

La Svolta Generativa: AudioLM e Bark. Un cambio di paradigma decisivo è rappresentato dalla recente transizione dai modelli acustici regressivi ai *Codec Language Models* generativi (es. VALL-E [23], AudioLM [24]). A differenza dei predecessori, questi modelli trattano la sintesi audio non come un problema di regressione del segnale, ma come un compito di *sequence modeling*, sfruttando l'architettura Transformer (decoder-only) per predire token audio discreti, analogamente a come i LLM predicono le parole.

In questo filone si colloca **Bark** [25], il modello open-source adottato in questo lavoro. Bark utilizza il quantizzatore vettoriale neurale **EnCodec** [26] per discretizzare l'audio in una gerarchia di token processati sequenzialmente: il modello genera prima dei *Semantic Tokens* di alto livello (che catturano il significato e la prosodia macroscopica), per poi raffinarli progressivamente in *Coarse* e *Fine Acoustic Tokens* che definiscono i dettagli ad alta frequenza [27]. Questa natura generativa abilita una caratteristica unica: la capacità di produrre segnali paralinguistici complessi (risate, esitazioni, respiri) semplicemente inserendo specifici *tag* testuali (es. `[laugh]`, `[sighs]`) nel prompt. Tale controllabilità sui fenomeni non verbali rende Bark il candidato ideale per restituire la ricchezza emotiva spesso persa nei sistemi TTS standard.

2.7 Lo Scenario Tecnologico e di Ricerca per la Lingua dei Segni Italiana (LIS)

La ricerca accademica sulla Lingua dei Segni Italiana sconta un ritardo strutturale dovuto alla carenza di dataset pubblici su larga scala, specialmente per compiti avanzati come il riconoscimento di frasi continue (CSLR/T) e l'analisi emotiva. Tale lacuna rappresenta un collo di bottiglia critico per l'addestramento di modelli di *deep learning* performanti, costringendo spesso la comunità scientifica a validare le metodologie su lingue più ricche di risorse, come l'American Sign Language (ASL). Nonostante le difficoltà sul fronte dei dati, il panorama commerciale italiano dimostra un notevole dinamismo, con un ecosistema di startup e aziende che propongono soluzioni innovative per l'accessibilità.

2.7.1 Panoramica dei Dataset Esistenti per la LIS

Le risorse attualmente disponibili per la LIS sono frutto di sforzi accademici e progetti internazionali, ma risultano spesso frammentarie o vincolate a scopi specifici (lessicografici o didattici), mancando della scala e della strutturazione necessarie per la traduzione automatica moderna.

Un primo pilastro è rappresentato dai progetti lessicografici come **Spread the Sign**, un'iniziativa di cooperazione europea che ha creato un vasto dizionario video multilingue. Sebbene preziosa per la documentazione e la didattica, questa risorsa manca delle annotazioni temporali e della continuità frasale richieste per il machine learning.

Sul fronte dei dataset computazionali, le risorse sono focalizzate su domini ristretti o specifici task di riconoscimento:

- **MultiMedaLIS (2024)**: Un recente contributo accademico focalizzato sul dominio medico. Si tratta di un dataset multimodale che sfrutta una ricca varietà di canali (RGB fino a 1080p, Depth, Optical Flow, scheletri) per circa 26.000 istanze. Sebbene i segni siano isolati, sono stati selezionati per la componibilità in frasi, ma l'accesso ai dati è attualmente vincolato a scopi di ricerca proprietaria.
- **Dataset dell'Alfabeto LIS (2021)**: Una risorsa *open source* progettata specificamente per il riconoscimento statico delle configurazioni manuali (*fingerspelling*). Il dataset offre centinaia di immagini RGB (64x64 px) per le 22 lettere dell'alfabeto manuale, risultando ideale per il training di classificatori di base ma insufficiente per la traduzione del discorso continuo.
- **A3LIS-147 (2015)**: Uno dei primi esperimenti italiani per il riconoscimento automatico, comprendente 1.470 clip video (RGB-D) di 147 segni del lessico quotidiano eseguiti da 10 segnanti. Pur essendo un punto di riferimento storico, la sua dimensione limitata ne restringe l'applicabilità alle moderne architetture profonde.
- **LIS Corpus Project (2011)**: Un progetto di grande valore linguistico che raccoglie conversazioni spontanee e narrazioni di 165 segnanti in 10 città, catturando le varianti regionali. Tuttavia, la natura non strutturata dei video e le restrizioni di licenza (accessibile solo tramite accordi con istituzioni come l'ISTC-CNR) ne rendono complesso l'utilizzo per l'addestramento massivo di reti neurali.

Questa disamina conferma l'assenza critica di un corpus LIS *gold-standard* per il parlato continuo, analogo all'How2Sign americano, che abiliti la ricerca sulla traduzione end-to-end.

2.7.2 Soluzioni Commerciali e Applicative in Italia

In contrasto con la frammentazione della ricerca di base, il settore privato italiano offre una vetrina di soluzioni applicative mature, che spaziano dall'intervento umano mediato alla completa automazione.

L'approccio più consolidato è il **Video-Interpretariato a Distanza (VRI)**, dove la tecnologia funge da ponte per l'intervento umano. Piattaforme come **e-LISIR** offrono servizi on-demand tramite app, permettendo a privati e Pubbliche Amministrazioni di accedere a interpreti professionisti da remoto, abbattendo le barriere logistiche.

Parallelamente, si assiste all'ascesa di soluzioni basate sull'Intelligenza Artificiale. Nel campo della traduzione automatica (*Sign-to-Text*), startup come **Handy Signs** propongono "interpreti digitali" basati su visione artificiale, capaci di tradurre bidirezionalmente (segno-voce e voce-testo) tramite tablet, con un focus specifico sui servizi di sportello (B2B). Sul versante opposto, la generazione (*Text-to-Sign*) è presidiata da attori come **Evodeaf**, che sviluppano ecosistemi basati su avatar 3D animati da algoritmi di machine learning, integrando dizionari collaborativi e strumenti di scansione OCR.

Infine, l'approccio *sensor-based* continua a evolversi in dispositivi indossabili innovativi. Progetti come **Limix** con il braccialetto **Talking Hands** rappresentano l'evoluzione moderna dei data gloves, puntando su design e portabilità per offrire uno strumento di traduzione personale che non dipende da telecamere esterne.

2.8 Il Posizionamento della Ricerca nel Contesto Attuale

L'analisi dello stato dell'arte ha evidenziato come la ricerca si sia concentrata prevalentemente sulla traduzione dal segno al testo (SLT). Tuttavia, un quadro completo della comunicazione inclusiva non può prescindere dalle frontiere della produzione visiva (*Sign Language Production*) e dalle implicazioni etiche legate all'uso dei dataset e all'automazione di una lingua minoritaria. Tali tematiche, per la loro rilevanza trasversale e l'impatto sui risultati sperimentali, verranno discusse approfonditamente nel **Capitolo 5**, alla luce delle evidenze emerse durante lo sviluppo della pipeline proposta.

Capitolo 3

Materiali e Metodi

Il presente capitolo illustra nel dettaglio i materiali e le metodologie adottate per la realizzazione della pipeline ”Sign-to-Speech”. Per garantire chiarezza espositiva e rigore scientifico, la trattazione è strutturata distinguendo la fase di ingegneria dei dati dalla descrizione delle architetture di Deep Learning.
L'esposizione si articola in quattro aree fondamentali. Inizialmente viene presentata la raccolta e preparazione dei dati, offrendo una disamina trasversale dei dataset utilizzati (ASLLRP, How2Sign, EmoSign) e delle procedure di acquisizione (web scraping) e preprocessing. Successivamente, vengono descritti i tre moduli che compongono l'architettura: il sistema di traduzione automatica (Modulo 1 - Sign-to-Text), il sistema ibrido di riconoscimento del sentimento tramite Meta-Learner (Modulo 2) e, infine, la metodologia per la generazione di parlato espressivo (Modulo 3 - Sentiment-Aware).

3.1 Dataset e Ingegneria dei Dati

La validazione empirica di un modello di deep learning, specialmente in un dominio complesso come il riconoscimento del *Sentiment* nella lingua dei segni, dipende in modo critico dalla qualità, dalla pertinenza e dalla scala dei dati utilizzati. La fase di acquisizione e preparazione dei dati ha rappresentato una componente fondamentale di questo lavoro di tesi (*Data Engineering*), richiedendo lo sviluppo di soluzioni tecniche ad-hoc per superare le limitazioni di accesso e formattazione dei corpus esistenti.

In questa sezione descriviamo i dataset selezionati (**ASLLRP** e **How2Sign**) e le metodologie impiegate per la loro raccolta, etichettatura e processamento.

3.1.1 Panoramica dei Corpus Selezionati

La strategia adottata combina due tra i più importanti corpus per la Lingua dei Segni Americana (ASL), con l'obiettivo di costruire un insieme di dati eterogeneo che supporti

efficacemente sia l'approccio *pose-based* che quello *appearance-based*.

Il primo pilastro è l'**American Sign Language Linguistic Research Project (ASLLRP)**, accessibile tramite il portale del Data Analysis Institute (DAI) della Rutgers University[1]. Questo dataset è di inestimabile valore per l'alta qualità linguistica, essendo stato prodotto con segnanti nativi e annotato meticolosamente. Tuttavia, la sua fruizione presentava significative sfide tecniche — quali l'assenza di download massivo e il formato video non segmentato — che hanno reso necessario lo sviluppo di una pipeline di scraping dedicata.

Il secondo pilastro è costituito da **How2Sign**[2], un corpus su larga scala che si distingue per accessibilità e ricchezza modale, offrendo pacchetti dati già organizzati in split standard (circa 31.000 campioni per il training). Per sostenere la modularità della nostra sperimentazione, sono state selezionate due risorse specifiche: le **B-F-H 2D Keypoints clips**, contenenti coordinate scheletriche pre-estratte utili per abbattere i tempi di pre-processing dei modelli leggeri, e le **Green Screen RGB clips**. L'acquisizione di queste ultime in formato grezzo è risultata indispensabile sia per il task di trascrizione, sia per l'addestramento del modello *Video Vision Transformer* (ViViT), che necessita di analizzare micro-espressioni facciali e texture della pelle perse nella scheletrizzazione.

3.1.2 Pipeline di Elaborazione del Corpus ASLLRP

Diversamente da How2Sign, il corpus ASLLRP non era immediatamente fruibile per pipeline di deep learning. È stato quindi necessario implementare una catena di elaborazione articolata in acquisizione, segmentazione, etichettatura ed estrazione delle feature.

Acquisizione tramite Web Scraping e Segmentazione. L'assenza di API sul portale DAI rendeva impraticabile l'acquisizione manuale delle clip. Per ovviare al problema, è stato sviluppato uno script di **web scraping** in Python basato sulla libreria Selenium[3]. Lo sviluppo di questo tool rappresenta un contributo metodologico utile alla comunità di ricerca, in quanto automatizza l'autenticazione e la gestione del download sequenziale, rendendo finalmente accessibile questo dataset per studi su larga scala. Una volta acquisiti i video grezzi, script dedicati hanno isolato le singole frasi (*utterance*) e riconciliato ogni clip con la relativa traduzione XML, generando un indice unificato che funge da ponte tra segnale visivo e contenuto semantico.

[1] https://dai.cs.rutgers.edu/dai/s/dai

[2] https://how2sign.github.io/index.html

[3] Codice disponibile su: **GitHub - WebScraping_ASLDataset**

Analisi del Sentiment e Creazione del Dataset (Weak Labeling). Poiché il corpus ASLLRP è nativamente privo di etichette emotive, è stata implementata una strategia di *Weak Labeling* basata sul testo. L'obiettivo era partizionare i dati in un **Golden Set** di alta affidabilità per il test e un **Silver Set** su larga scala per il training. Come *Ground Truth*, è stato adottato il dataset **EmoSign** [3], composto da circa 200 clip di ASLLRP annotate manualmente. Per etichettare automaticamente il resto del corpus, abbiamo applicato l'algoritmo **VADER** alle traduzioni testuali, calibrando la soglia di taglio tramite un approccio iterativo "Top-Down" sul Golden Set. L'iterazione si è arrestata a una soglia $\tau = 0.34$, valore che massimizzava l'intersezione tra predizioni automatiche e classi reali. Le clip etichettate con questa soglia costituiscono il *Silver Set*, garantendo una distribuzione delle classi coerente con la percezione umana.

Infine, per alimentare i modelli geometrici (LSTM), è stato utilizzato **MediaPipe** per estrarre le coordinate 3D, formattando l'output per replicare la struttura dati di How2Sign e creare un dataset unificato.

3.1.3 Elaborazione di How2Sign e Composizione Finale

Il processamento di How2Sign si è concentrato sull'omogeneizzazione dei metadati, applicando la medesima pipeline VADER ($\tau = 0.34$) per permettere la fusione dei due corpus in un unico vasto dataset di addestramento.

La composizione finale dei dati per i tre task sperimentali è la seguente:

- **Task 1 (Sign-to-Text):** È stato impiegato esclusivamente il dataset How2Sign nella sua interezza, suddiviso negli split ufficiali (31.165 campioni per il training, 1.741 per la validazione e 2.357 per il testing).
- **Task 2 (Sign-to-Sentiment):** È stato costruito un dataset ibrido. Il *Training Set* (oltre 4.000 istanze) unisce il Silver Set di ASLLRP e una selezione bilanciata di How2Sign. Il *Test Set* è costituito esclusivamente dai 200 campioni del Golden Set di EmoSign, garantendo una valutazione basata sulla verità umana.
- **Task 3 (Sintesi Vocale):** I test di generazione (inferenza) sono stati condotti sui 200 campioni del dataset EmoSign, utilizzando le trascrizioni e le etichette originali.

3.2 Modulo 1: Metodologia per la Traduzione in Testo (Sign-to-Text)

Il primo modulo della pipeline affronta il compito della *Sign Language Translation* (SLT), con l'obiettivo di tradurre sequenze video continue di American Sign Language (ASL) in frasi di testo in lingua inglese, garantendo correttezza grammaticale e fedeltà semantica.

La metodologia sviluppata costituisce un adattamento originale dell'architettura allo stato dell'arte **SSVP-SLT** (*Self-Supervised Video Pretraining for Sign Language Translation*) [13]. Il framework originale propone un approccio modulare a due stadi che combina un encoder visivo (SignHiera) con un decoder testuale massivamente multilingue, specificamente **NLLB-200** (*No Language Left Behind*). Tale architettura sfrutta lo spazio di embedding condiviso (noto come SONAR) per migliorare l'allineamento semantico tra le feature visive e le rappresentazioni testuali.

Tuttavia, la replicazione fedele di tale lavoro pone barriere d'ingresso proibitive per la ricerca accademica standard, dipendendo da pre-addestramenti massivi su dati proprietari e risorse computazionali ingenti (64 GPU NVIDIA A100). Nell'ottica di democratizzare questo approccio e renderlo sostenibile in un ambiente a risorse vincolate (singola GPU), la presente tesi adotta l'architettura basata su NLLB introducendo una strategia sperimentale differenziata. In primo luogo, abbiamo condotto un'**Esplorazione Dimensionale Scalare**, addestrando diverse varianti del modello (da 600M a 3.3B parametri) per isolare l'impatto della dimensione del modello rispetto al pre-training. In secondo luogo, abbiamo adottato un approccio di **Transfer Learning Adattivo**: in luogo del costoso pre-training video *from scratch*, il sistema è stato inizializzato sfruttando esclusivamente la conoscenza linguistica pregressa di NLLB, focalizzando il training sull'allineamento tra le feature visive e lo spazio di embedding del traduttore.

3.2.1 Preprocessing e Estrazione delle Feature Visive

L'input del sistema è costituito dalle clip video del dataset **How2Sign**. A differenza degli approcci real-time che operano in streaming, la nostra architettura processa l'intera sequenza video simultaneamente per catturare il contesto globale dell'utterance.

Ogni video è sottoposto a un rigoroso pre-processamento volto a standardizzare l'input: i frame vengono ridimensionati a 224×224 pixel e campionati uniformemente per estrarre una sequenza fissa di $T = 128$ frame. Tale valore rappresenta il trade-off necessario per gestire i vincoli di memoria senza sacrificare la dinamica dei segni

veloci. Infine, i valori dei pixel vengono normalizzati utilizzando le statistiche standard di ImageNet (μ, σ).

Per l'estrazione delle feature è stata adottata l'architettura **Hiera** (*Hierarchical Vision Transformer*), nella variante `hiera_base_16x224`. A differenza dei Vision Transformer (ViT) standard, Hiera impiega una struttura gerarchica che riduce progressivamente la risoluzione spaziale aumentando il numero di canali, permettendo l'apprendimento di feature multi-scala robuste. Il modello è stato inizializzato con i pesi pubblici `dm_70h_ub_signhiera.pth`; l'output dell'encoder è un tensore di feature $F \in \mathbb{R}^{T \times D}$ (con $D = 768$), che viene serializzato su disco per ottimizzare l'efficienza degli esperimenti.

3.2.2 Architettura del Modulo di Traduzione e Strategie di Training

Il cuore del sistema è un modello Sequence-to-Sequence incaricato di trasformare le feature visive continue in token testuali discreti. Abbiamo implementato l'architettura `SonarSignModel`, che funge da ponte tra la modalità visiva e quella linguistica.

Poiché il decoder NLLB è progettato per ricevere in input token testuali, è stato necessario progettare un **Visual Adapter**: una rete neurale leggera (MLP) che proietta lo spazio delle feature visive ($D_{vis} = 768$) nello spazio latente del modello linguistico ($D_{txt} = 1024$), trasformando i frame video in "soft-prompts" vettoriali processabili nativamente. Per bilanciare le prestazioni, sono state esplorate diverse configurazioni del backbone NLLB, dalla variante Distilled (600M) fino a quella da 3.3B parametri gestita tramite *Gradient Checkpointing*.

Per validare l'efficacia del modello, la sperimentazione ha confrontato diverse strategie. Relativamente all'aggregazione temporale, abbiamo contrapposto l'approccio **Standard Seq2Seq** (basato su Cross-Attention frame-by-frame) alla variante **SONAR Pooling (Vector-to-Text)**, che comprime l'intero video in un singolo vettore denso tramite Mean Pooling. Parallelamente, sono state valutate due modalità di aggiornamento dei parametri: la configurazione **Frozen Encoder**, che congela i pesi dell'Encoder di NLLB per prevenire il "catastrophic forgetting", e la configurazione **Full Fine-Tuning**, che aggiorna tutti i parametri per massimizzare l'adattamento alla sintassi spaziale della ASL.

3.2.3 Setup Sperimentale e Iperparametri

La notevole disparità delle risorse computazionali rispetto alla letteratura (singola GPU A100 vs cluster da 64 GPU) ha imposto scelte ingegneristiche rigorose per garantire

la convergenza. Per assicurare la riproducibilità, è stata adottata una configurazione standardizzata (iperparametri presi dal paper di riferimento): il training è stato eseguito sul dataset How2Sign con batch size di 16, utilizzando la *Cross-Entropy Loss* con Label Smoothing a 0.2. L'ottimizzazione è stata affidata all'algoritmo AdamW ($lr = 1e^{-3}$, weight decay 0.1) con uno scheduler ibrido che combina Linear Warmup e Cosine Annealing. Per stabilizzare l'addestramento in regime di *Mixed Precision* (bfloat16), è stato applicato sistematicamente il *Gradient Clipping* (norma 1.0). Infine, per evitare l'overfitting, è stato implementato l'*Early Stopping* con pazienza di 15 epoche, mentre la generazione finale sfrutta una *Beam Search* a 5 raggi.

3.3 Modulo 2: Architettura Sign-to-Sentiment Multimodale (Il Meta-Learner)

Il secondo modulo costituisce il cuore "affettivo" della tesi. Superando l'approccio classico unimodale focalizzato esclusivamente sul video, proponiamo un'architettura a **Ensemble Multimodale** che fonde le informazioni visive (prosodia) con quelle semantiche (testo) tramite una strategia di *Late Fusion* governata da un Meta-Learner.

L'architettura si articola in tre blocchi funzionali distinti: lo Stream Visivo, lo Stream Testuale e il Modulo di Fusione. Di seguito, approfondiamo la metodologia sviluppata per il sottosistema visivo, la cui definizione ha richiesto un'indagine comparativa tra diverse architetture neurali.

3.3.1 Sottosistema Visivo: Confronto tra Paradigmi

Per affrontare il compito di riconoscimento delle emozioni da dati sequenziali, l'indagine sperimentale non ha seguito un percorso lineare, ma ha messo a confronto due paradigmi rappresentativi dello stato dell'arte, ciascuno portatore di specifici vantaggi: l'approccio *Pose-based* (geometrico) e l'approccio *Appearance-based* (visivo puro).

Approccio Geometrico e Privacy by Design (LSTM e ST-GCN). La prima famiglia di modelli opera su dati scheletrici. Questa scelta implementa nativamente il principio di **Privacy by Design**: astraendo il volto in una nuvola di punti (landmark), è possibile analizzare la prosodia e l'intensità del gesto mantenendo l'anonimato del segnante, un requisito cruciale in ambito assistivo. Come modello di riferimento è stata adottata una rete **Long Short-Term Memory (LSTM)**, implementata nel modulo `EmotionLSTM`. Questa architettura processa sequenze di vettori di landmark (input size 1404) utilizzando meccanismi di *gating* per catturare le dipendenze temporali a lungo termine,

proiettando infine l'ultimo stato nascosto tramite un layer lineare per la classificazione. Parallelamente, è stata valutata l'architettura **Spatial-Temporal Graph Convolutional Network (ST-GCN)**. La logica di questa scelta risiede nel trattare i landmark non come vettori piatti, ma come nodi di un grafo strutturato anatomicamente. Il modello alterna convoluzioni spaziali (postura) e temporali (movimento) basate su una matrice di adiacenza predefinita. Sebbene teoricamente valida, questa rappresentazione topologica si è rivelata rigida nel catturare le micro-espressioni rispetto alla fluidità delle RNN.

Approccio Appearance-based (ViViT). In contrapposizione alla geometria pura, è stato implementato un approccio basato sui pixel utilizzando il **Video Vision Transformer (ViViT)**. Questo modello processa direttamente i frame grezzi, permettendo di catturare olisticamente sia le informazioni spaziali che l'evoluzione temporale, includendo dettagli di texture (es. rughe d'espressione) che la scheletrizzazione inevitabilmente perde. Abbiamo adottato una strategia di *transfer learning* partendo dal checkpoint `vivit-b-16x2` pre-addestrato su Kinetics-400. Il video viene suddiviso in "tubelet" spaziotemporali processati come token; la strategia di implementazione ha previsto un *Fine-Tuning* mirato, congelando il backbone per preservare le feature generali di movimento e addestrando esclusivamente la testa di classificazione ("Classifier Head") inizializzata da zero.

Setup Sperimentale e Ottimizzazione Il training e la validazione sono stati gestiti attraverso un protocollo rigoroso volto a garantire riproducibilità su hardware Apple Silicon M3.

La gestione dei dati ha richiesto strategie specifiche per mitigare lo sbilanciamento delle classi. Sono state adottate tecniche di *Downsampling* della classe maggioritaria e l'utilizzo di un `WeightedRandomSampler` nei DataLoader. Le pipeline di preprocessing differiscono per architettura: mentre per i modelli geometrici è stata applicata una normalizzazione degli scaler numerici fittati sul training set, per il modello ViViT la classe `VideoDataset` gestisce il campionamento uniforme dei frame e la normalizzazione dei pixel tramite l'`ImageProcessor` di Hugging Face.

Il ciclo di addestramento, orchestrato tramite la libreria *Ignite*, utilizza funzioni di costo adattive: una **Focal Loss** per i modelli geometrici (per concentrarsi sui campioni "difficili") e una `CrossEntropyLoss` pesata per il ViViT. L'ottimizzazione è affidata ad `AdamW` con scheduler `ReduceLROnPlateau`, integrando regolarizzazioni sistematiche come Weight Decay, Dropout ed *Early Stopping* basato sull'F1-Macro di validazione.

La ricerca degli iperparametri ottimali è stata automatizzata tramite **Optuna**. Per il ViViT, lo studio si è concentrato sui parametri di fine-tuning, esplorando su scala logaritmica il Learning Rate (10^{-5} – 10^{-3}) e il Weight Decay, utilizzando pruners per

interrompere precocemente i trial non promettenti. La valutazione finale monitora un set diversificato di metriche tramite *MLflow*, tra cui l'Accuratezza, l'**F1-Score Macro** (metrica target), e metriche granulari come la Precision per classe, visualizzate tramite Matrici di Confusione per diagnosticare bias specifici verso le emozioni negative o neutre.

3.3.2 Stream Testuale: Analisi del Sentiment da Trascrizioni

Il secondo pilastro dell'architettura multimodale è rappresentato dallo Stream Testuale (*Text-to-Sentiment*). Se il modello visivo si occupa di interpretare *come* viene espresso un concetto (prosodia, espressione), il modello testuale analizza *cosa* viene detto (semantica), elaborando le trascrizioni in lingua inglese generate dal modulo di traduzione o provenienti dal ground truth. L'obiettivo di questo sottosistema è generare un vettore di probabilità (*logits*) per le tre classi di sentiment (Positive, Negative, Neutral) basandosi esclusivamente sul contenuto lessicale e sintattico della frase. La metodologia si è sviluppata attraverso una fase preliminare di selezione del modello (*Model Selection*) su architetture pre-addestrate, seguita dal consolidamento tramite *fine-tuning* supervisionato.

Selezione del Modello (Model Selection). Per identificare l'architettura più idonea, è stata condotta un'analisi comparativa su tre modelli rappresentativi dello stato dell'arte NLP. Il primo candidato, **BERT-Base (GoEmotions)**[4], offre un output granulare su 28 classi emotive (es. *joy, remorse*); per renderlo compatibile con il nostro target ternario, è stato necessario implementare un algoritmo di aggregazione tassonomica che somma le probabilità delle sottomodulazioni (es. aggregando *admiration, joy, love* nella classe Positive). Il secondo candidato, **RoBERTa-Base (Twitter Sentiment)**[5], nasce specificamente per la *Sentiment Analysis* su un corpus di 124 milioni di tweet. Essendo nativo a 3 classi e ottimizzato per testi brevi e informali, presenta un allineamento intrinseco con la sintassi essenziale delle traduzioni ASL. Infine, è stato testato **Gemini 3.0 Pro**[6] come benchmark zero-shot per valutare se le capacità di ragionamento di un LLM generalista potessero superare l'efficienza dei modelli specializzati.

L'analisi sperimentale ha decretato la superiorità di **RoBERTa-Base** in termini di Weighted F1-score sul dataset di test. La sua pre-ottimizzazione su testi provenienti dai social media si è rivelata un vantaggio decisivo rispetto alla complessità di aggregazione richiesta da BERT. Di conseguenza, RoBERTa è stato selezionato come backbone per la fase successiva.

[4]https://huggingface.co/bhadresh-savani/bert-base-go-emotion
[5]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
[6]https://deepmind.google/models/gemini/pro/

Fine-Tuning e Configurazione. Per massimizzare l'accuratezza rispetto allo scenario zero-shot, è stato eseguito un **Full Fine-Tuning** aggiornando tutti i 125M di parametri del modello sul dominio specifico. L'addestramento ha sfruttato il *Silver Set* (circa 6.000 campioni ibridi ASLLRP/How2Sign) per il training, riservando il *Golden Set* di EmoSign per il test. Il processo, implementato tramite la libreria Hugging Face Transformers, ha adottato una configurazione conservativa per preservare la conoscenza pregressa: è stato utilizzato un Learning Rate di $1e^{-5}$ con un Batch Size di 256 per 10 epoche. Per gestire eventuali sbilanciamenti residui nelle classi, la funzione di costo standard è stata sostituita con una **Focal Loss**. Il risultato finale è un modello che produce un vettore di confidenza tridimensionale pronto per l'integrazione nel Meta-Learner.

3.3.3 Il Meta-Learner: Strategia di Fusione Tardiva (Late Fusion)

Il terzo e ultimo stadio del modulo multimodale è costituito dal **Meta-Learner**, un classificatore di livello superiore progettato per orchestrare e sintetizzare le informazioni provenienti dai due stream unimodali paralleli. L'architettura adotta una strategia di *Late Fusion* (fusione tardiva), operando nello spazio decisionale piuttosto che in quello delle feature. A differenza degli approcci di *Early Fusion*, che concatenano vettori ad alta dimensionalità provenienti da domini eterogenei (feature geometriche e semantiche), il nostro approccio fonde le stime probabilistiche (*soft-labels*) generate dai modelli esperti a valle. Questa scelta metodologica disaccoppia la complessità dei singoli stream, permettendo a ciascun modello (LSTM e RoBERTa) di specializzarsi sul proprio dominio e demandando al Meta-Learner il compito di apprendere l'affidabilità relativa di ciascuna modalità in base al contesto dell'utterance.
Il workflow sperimentale si è articolato nella costruzione del *Meta-Dataset* e nella successiva ottimizzazione del modello di fusione.

Ingegneria delle Feature e Meta-Dataset. Per addestrare il classificatore di fusione, è stato costruito un dataset tabulare intermedio in cui ogni istanza rappresenta un segmento video e le feature corrispondono ai livelli di confidenza espressi dai modelli unimodali. Sfruttando i modelli precedentemente addestrati e congelati (*frozen*), sono state eseguite procedure di inferenza sugli split del dataset. Nello specifico, dallo stream visivo (`EmotionLSTM`) sono stati estratti i vettori post-Softmax (previa normalizzazione dei landmark), mentre dallo stream testuale (`RoBERTa`) sono stati derivati i *logits* convertiti in probabilità.

Formalmente, per ogni campione i-esimo, i due modelli esperti producono due vettori di probabilità $\mathbf{v} \in [0, 1]^3$, normalizzati in modo che la somma delle componenti

sia unitaria:

$$\mathbf{v}_{vis}^{(i)} = [P(Neg)_{vis}, P(Neu)_{vis}, P(Pos)_{vis}]$$
$$\mathbf{v}_{txt}^{(i)} = [P(Neg)_{txt}, P(Neu)_{txt}, P(Pos)_{txt}]$$

Il vettore di feature $\mathbf{x}^{(i)}$ fornito in input al Meta-Learner è ottenuto tramite la concatenazione diretta di questi due vettori, definendo uno spazio $\mathbb{R}^6$:

$$\mathbf{x}^{(i)} = \mathbf{v}_{vis}^{(i)} \oplus \mathbf{v}_{txt}^{(i)}$$

La coerenza tra le modalità è stata garantita mediante un'operazione di *Inner Join* sulle chiavi univoche dei file video, assicurando l'allineamento dei campioni.

Protocollo di Addestramento e Selezione del Modello. Data la ridotta dimensionalità dello spazio di input (solo 6 feature), è stato possibile ed opportuno esplorare famiglie di classificatori classici (*Shallow Learning*) piuttosto che reti neurali profonde, in ossequio al principio di parsimonia per evitare l'overfitting.

La procedura di *Model Selection* si è articolata in due fasi distinte per garantire una valutazione non distorta. Nella prima fase, dedicata all'ottimizzazione degli iperparametri, è stata eseguita una *Grid Search* con **Stratified 5-Fold Cross-Validation** esclusivamente sul **Training Set**. Questa operazione ha coinvolto quattro famiglie algoritmiche principali (**Logistic Regression**, **Random Forest**, **SVM**, **Decision Tree**), permettendo di identificare la configurazione ottimale per ciascuna di esse e gestendo al contempo il forte sbilanciamento delle classi tramite l'assegnazione di pesi inversi alla frequenza (`class_weight='balanced'`).

Successivamente, i quattro stimatori ottimizzati sono stati sottoposti a una valutazione comparativa su un **Validation Set** dedicato e mai osservato durante il tuning. Il confronto è stato guidato dalla massimizzazione dell'*F1-Score Weighted* e del *Cohen's Kappa Score*, metriche ritenute più robuste dell'accuratezza in questo contesto. Il modello risultato vincitore in questa fase è stato infine "congelato" e promosso direttamente alla fase di Test finale, senza ulteriori ri-addestramenti sui dati di validazione, al fine di misurarne la capacità di generalizzazione su dati totalmente inediti.

3.4 Modulo 3: Sintesi Vocale Espressiva *"Sentiment Aware"*

Il terzo e ultimo stadio della pipeline affronta la sfida della generazione di un segnale audio che coniughi l'intelligibilità semantica con la coerenza affettiva rispetto al messaggio originale in Lingua dei Segni. Mentre i moduli precedenti si sono focalizzati sulla traduzione del contenuto e sull'identificazione del Sentiment e la quantificazione dell'intensità del Sentiment, questo modulo ha il compito di modulare la prosodia del parlato sintetico per riflettere fedelmente tale intensità.

La metodologia adottata si articola in tre fasi sequenziali che spaziano dalla generazione condizionata, implementata tramite una strategia di *Prompt Engineering* dinamico per il modello Bark, fino alla validazione percettiva umana, passando per un rigoroso controllo qualità automatico basato su metriche ASR (*Automatic Speech Recognition*) per filtrare gli artefatti generativi.

3.4.1 Architettura Generativa e Mappatura Acustica

Come motore di sintesi è stato selezionato **Bark**, un modello *transformer-based text-to-audio*. A differenza dei sistemi TTS tradizionali come Tacotron [16] o FastSpeech [17], Bark opera generando token audio discreti direttamente dal testo. Questa architettura generativa permette di modellare non solo i fonemi, ma anche i tratti paralinguistici non verbali (risate, esitazioni, respiri, sospiri), elementi essenziali per veicolare la naturalezza del sentimento umano.

Al fine di tradurre lo score di Sentiment del dataset EmoSign (variabile discreta nell'intervallo $[-3, +3]$) in un prompt acustico efficace, è stata sviluppata una strategia di mapping deterministico (implementata in `emotion_mapper.py`) che agisce su tre livelli di controllo del modello.

In primo luogo, la **Selezione dello Speaker (History Prompt)** assegna a ogni macrocategoria emotiva (Positive, Negative, Neutral) un pool di voci specifiche, selezionate empiricamente per la loro predisposizione timbrica (es. voci brillanti per il positivo, voci profonde o "breathless" per il negativo). La scelta dello speaker all'interno del pool avviene deterministicamente tramite l'hash del nome del video, garantendo la totale riproducibilità dell'esperimento.

Parallelamente, la **Modulazione della Temperatura** (T) regola la stocasticità nella predizione dei token per bilanciare espressività e stabilità. Per le emozioni *Positive*, è stata adottata una temperatura più elevata ($T = 0.75$) per massimizzare la varianza prosodica e generare inflessioni dinamiche tipiche della gioia; per le emozioni *Negative*, si utilizza un valore intermedio ($T = 0.65$) che consente artefatti come i sospiri

mantenendo il controllo strutturale; per lo stato *Neutral*, la temperatura è ridotta al minimo ($T = 0.55$) per garantire un parlato piano e privo di allucinazioni acustiche.

Infine, l'**Inserimento Contestuale dei Tag Paralinguistici** sfrutta i token semantici speciali di Bark. Per evitare inserimenti meccanici, un modulo di ottimizzazione sintattica (`emotion_tag_optimizer.py`) analizza l'utterance individuando i *Natural Breaks*. La strategia distingue tra bassa intensità (Score ± 1), dove si adotta un approccio conservativo posizionando i tag (es. `[chuckles]` o esitazioni) all'inizio o alle pause, e alta intensità (Score $\pm 2, \pm 3$), dove si privilegia l'inserimento intra-frasale (*mid-sentence*) per massimizzare l'impatto emotivo. La Tabella 3 riassume lo schema adottato.

Emozione	**Score**	**Tag Inserito**	**Descrizione**	**Temp (T)**
Positive	+3	`[laughter]`	Risata piena	0.75
	+2	`[laughs]`	Risata moderata	0.75
	+1	`[chuckles]`	Risatina sottile	0.75
Neutral	0	*Nessuno*	Parlato neutro	0.55
Negative	-1	`...`	Pausa esitante	0.65
	-2	`[gasps]`	Sorpresa negativa/Shock	0.65
	-3	`[sighs]`	Sospiro/Disperazione	0.65

Tabella 3: Schema di mappatura tra intensità emotiva EmoSign e configurazione di generazione Bark.

3.4.2 Controllo Qualità e Validazione Percettiva

I modelli audio generativi sono intrinsecamente soggetti a fenomeni di "allucinazione" e degrado dell'intelligibilità quando forzati da temperature elevate. Per garantire la validità scientifica dei risultati, la pipeline integra una fase di filtraggio automatico seguita dalla valutazione umana.

Filtraggio Automatico (WER). Il controllo qualità automatico si ispira alla metodologia di Wang e Székely [27] per verificare l'integrità semantica. La pipeline opera sequenzialmente: l'audio generato viene trascritto tramite il modello **Whisper** e confrontato con la caption originale normalizzata. Sulla base delle raccomandazioni per applicazioni ad alta fedeltà, è stata imposta una soglia di accettazione rigorosa del 10% sul *Word Error Rate* ($WER \leq 0.1$). I campioni che superano tale soglia vengono scartati a priori, assicurando che gli annotatori valutino solo file perfettamente intelligibili.

Protocollo di Valutazione Umana (Blind Cross Evaluation). Per misurare la congruenza emotiva, è stata condotta una campagna di annotazione soggettiva tramite un'applicazione dedicata sviluppata in Streamlit. Il protocollo *blind* ha adottato una **doppia modalità di valutazione (Cross-Modal)**: per ogni video, gli annotatori hanno valutato in momenti distinti e randomizzati il solo testo (*TEXT_ONLY*) e il solo audio (*AUDIO_ONLY*). Questa separazione è cruciale per isolare il contributo della prosodia da quello semantico. La classificazione è avvenuta su scala Likert a 7 punti (da -3 a $+3$), includendo un meccanismo di esclusione per segnalare eventuali difetti non linguistici (es. rumori spuri) sfuggiti al filtro automatico.

3.5 Metriche di Valutazione e Benchmark

La natura ibrida della pipeline, che attraversa i domini del *Deep Learning*, della *Computer Vision*, del *Natural Language Processing* e della *Speech Synthesis*, richiede l'adozione di protocolli di valutazione specifici per ciascun task. In questa sezione definiamo il perimetro di valutazione, descrivendo i dataset di test, le metriche quantitative adottate e le baseline di riferimento derivate dallo stato dell'arte.

Modulo 1: Traduzione (Sign-to-Text). Per il modulo di traduzione, l'obiettivo è misurare la fedeltà semantica e la qualità grammaticale delle frasi generate rispetto al riferimento umano. La validazione è stata condotta sul **Test Set ufficiale di How2Sign** (2.357 campioni), uno dei benchmark più sfidanti per la traduzione ASL in dominio aperto. Sono state adottate le metriche standard per la *Machine Translation*:

- **BLEU ($n = 1 \dots 4$):** Misura la precisione degli n-grammi tra l'ipotesi e il riferimento. L'analisi include tutte le metriche da BLEU-1 a BLEU-4: mentre BLEU-1 valuta la correttezza lessicale (parole singole), le metriche di ordine superiore (fino a BLEU-4) sono indicatori fondamentali della fluidità sintattica e della coerenza locale della frase.
- **ROUGE-L:** Valuta la capacità del modello di recuperare la più lunga sottosequenza comune, premiando la coerenza strutturale della frase.

Come baseline di riferimento è stato selezionato il modello **SSVP-SLT** (*Self-Supervised Video Pretraining for Sign Language Translation*) [13], nella configurazione addestrata su dati How2Sign, che rappresenta l'attuale "Upper Bound" per questo dominio specifico.

Modulo 2: Riconoscimento del Sentimento (Sign-to-Sentiment). Per la valutazione del riconoscimento del sentimento, il riferimento principale è costituito dal benchmark pubblicato nel recente lavoro *EmoSign* [3]. Questo studio ha testato le capacità dei moderni *Large Multimodal Models* (LMMs) come **MiniGPT4**, **Qwen 2.5**, **AffectGPT** e **GPT-4o** nel classificare il sentiment a partire da input video. I test sono stati condotti in due configurazioni distinte, cruciali per comprendere le dipendenze dei modelli: **Video-Only** (solo flusso visivo) e **Video + Caption** (video e trascrizione testuale). Le metriche adottate per il confronto sono l'*Accuracy* e il *Weighted F1-Score*, essenziali per valutare la capacità del nostro Meta-Learner di superare le limitazioni delle architetture generaliste. A queste si affiancano metriche avanzate per sondare la robustezza del modello: il **Class Gap** (deviazione standard degli F1-score), introdotto per quantificare l'equità delle performance tra le diverse emozioni, e la **Mean Probability Confidence**, utilizzata per analizzare la sicurezza decisionale delle predizioni corrette.

Modulo 3: Sintesi Vocale (Sentiment-Aware). A differenza dei moduli precedenti, per la sintesi vocale da testo emotivo non esiste un benchmark standardizzato universale. La valutazione della qualità si basa sulla metodologia proposta da Wang e Székely (2024) [27], che identifica nel **Word Error Rate (WER)** il correlato più affidabile per la stabilità semantica dei modelli generativi. È stata stabilita una soglia di scarto rigorosa del **10%** ($WER > 0.1$): questo parametro funge da *Baseline di Qualità*, garantendo che qualsiasi audio che superi tale soglia venga considerato "tecnicamente fallito" ed escluso dalla valutazione umana, indipendentemente dalla sua espressività.

Capitolo 4

Risultati

In questo capitolo vengono presentati e discussi i risultati sperimentali ottenuti per i tre moduli che compongono la pipeline *Sign-to-Speech* proposta. L'esposizione segue rigorosamente l'architettura modulare definita nei capitoli precedenti: si inizierà dall'analisi della traduzione automatica (Modulo 1), per poi approfondire il cuore algoritmico del riconoscimento del sentimento multimodale (Modulo 2) e concludere con la validazione della sintesi vocale espressiva (Modulo 3).

4.1 Risultati Modulo 1: Sistema di Traduzione in Testo (Sign-to-Text)

In questa sezione vengono presentati e analizzati i risultati quantitativi ottenuti addestrando l'architettura *NLLB-Adapter* sul dataset How2Sign. La sperimentazione ha perseguito un duplice obiettivo: da un lato, identificare il miglior compromesso tra capacità parametrica del modello e risorse computazionali disponibili, confrontando backbone da 600M, 1.3B e 3.3B parametri; dall'altro, isolare l'impatto della strategia di aggiornamento dei pesi, confrontando sistematicamente l'approccio *Full Fine-Tuning* (aggiornamento di tutti i parametri) con l'approccio *Frozen Encoder* (aggiornamento del solo adattatore e decoder).

4.1.1 Analisi della Convergenza e Metriche di Validazione

Il monitoraggio delle prestazioni sul *Validation Set* durante la fase di addestramento offre indicazioni cruciali sulla stabilità della convergenza e sul rischio di overfitting. La Tabella 4 illustra le metriche registrate nel punto di minima Loss (*Best Validation Loss*), corrispondente alla configurazione ottimale salvata tramite Early Stopping.

Modello	Configurazione	BLEU-1	BLEU-4	ROUGE-L	Best Val Loss
NLLB-600M	Full Train	**14.60**	**1.32**	10.37	5.53
NLLB-1.3B	Full Train	9.50	0.80	10.92	5.46
NLLB-1.3B	Frozen Encoder	8.30	0.79	8.41	5.46
NLLB-3.3B	Full Train	12.08	0.84	**11.23**	**5.41**
NLLB-3.3B	Frozen Encoder	12.15	1.11	10.82	**5.41**

Tabella 4: Metriche di performance sul Validation Set all'epoca di minima Loss. Si osserva che la Loss si stabilizza in funzione della dimensione del modello (5.46 per 1.3B, 5.41 per 3.3B) indipendentemente dalla strategia di training. Tuttavia, il modello 1.3B Full Train mostra un ROUGE-L sospettosamente alto (10.92) rispetto alla versione Frozen, un potenziale indizio di overfitting sui dati di validazione.

Dall'analisi della Tabella 4 emergono dinamiche contrastanti. In primo luogo, si nota una correlazione inversa tra Loss e metriche BLEU nel modello più leggero (**NLLB-600M Full Train**): esso esibisce i punteggi BLEU-1 (14.60) e BLEU-4 (1.32) più alti in assoluto sul set di validazione, ma registra contestualmente la *Validation Loss* peggiore (5.53). Questo apparente paradosso suggerisce che il modello piccolo tende a "memorizzare" sequenze lessicali frequenti, gonfiando il BLEU, ma fatica a modellare la distribuzione di probabilità complessa della lingua target, indicando una generalizzazione fragile.

Un secondo dato significativo riguarda l'invarianza della Loss nei modelli più grandi (1.3B e 3.3B), dove la strategia di training (Full vs Frozen) non impatta sul valore finale della funzione di costo (stabile rispettivamente a 5.46 e 5.41). Ciò dimostra che il "tetto" di apprendimento statistico è determinato primariamente dalla capacità del backbone pre-addestrato, confermando la robustezza delle rappresentazioni latenti di NLLB. Infine, confrontando le due varianti del modello intermedio, la versione *Full Train* appare superiore in validazione rispetto alla Frozen (ROUGE-L 10.92 vs 8.41), ma come verrà discusso nella sezione successiva, tale divario non si riflette nelle prestazioni sul Test Set, suggerendo un potenziale overfitting.

4.1.2 Performance sul Test Set e Confronto con le Baseline

La valutazione sul *Test Set* ufficiale, costituito da dati mai osservati dal modello, rappresenta la verifica definitiva delle capacità di generalizzazione. La Tabella 5 presenta il quadro completo dei risultati, mettendo a confronto le baseline *Zero-shot*, le configurazioni addestrate e il riferimento dello Stato dell'Arte.

Modello	Configurazione	B-1	B-2	B-3	B-4	ROUGE-L	Note
NLLB-600M	Inference (No F.T.)	6.91	1.58	0.28	0.06	7.64	Baseline Zero-shot
NLLB-600M	Full Train	7.21	2.50	0.95	0.47	8.94	Miglioramento limitato
NLLB-1.3B	Inference (No F.T.)	2.08	0.42	0.09	0.02	7.33	Baseline Zero-shot
NLLB-1.3B	Full Train	7.10	2.39	0.98	0.49	9.12	Recupero semantico
NLLB-1.3B	**Frozen Encoder**	7.35	**2.89**	**1.19**	0.57	**10.00**	**Sweet Spot: Miglior Bilanciamento**
NLLB-3.3B	Inference (No F.T.)	**10.05**	1.55	0.30	0.08	9.24	Baseline robusta
NLLB-3.3B	Full Train	7.39	2.47	1.10	0.52	9.27	Recupero semantico
NLLB-3.3B	Frozen Encoder	6.69	2.56	1.18	**0.58**	8.12	Miglior BLEU-4 assoluto
SSVP-SLT	*H2S dataset*	*30.2*	*16.7*	*10.5*	*7.0*	*25.7*	*SOTA (64 GPUs A100)*

Tabella 5: Risultati comparativi completi sul Test Set. B-n indica BLEU-n. Il confronto diretto sul modello 1.3B rivela che la strategia *Frozen Encoder* supera nettamente la *Full Train* (ROUGE-L 10.00 vs 9.12), confermando che il congelamento dei pesi agisce come un regolarizzatore essenziale in contesti *low-resource*.

Il confronto tra le configurazioni *Zero-shot* e quelle addestrate sancisce l'imprescindibilità della fase di training. Indipendentemente dalla dimensione, i modelli non addestrati ottengono un punteggio BLEU-4 prossimo allo zero (< 0.1), indicando che, pur riconoscendo occasionalmente singole parole, mancano della capacità di articolare una sintassi grammaticale coerente a partire dalle sole feature visive. Dopo il training, il BLEU-4 sale a valori tangibili (0.47 - 0.58): sebbene bassi in termini assoluti, rappresentano un salto qualitativo significativo, testimoniando l'apprendimento della mappatura non lineare tra la sintassi spaziale dell'ASL e la grammatica inglese.

Di particolare interesse è l'analisi del modello **NLLB-1.3B**, che permette un confronto diretto tra le strategie di aggiornamento dei pesi. Inversamente a quanto osservato in validazione, sul Test Set la versione **Full Train** degrada le prestazioni (ROUGE-L 9.12), mentre la versione **Frozen Encoder** mantiene la sua robustezza, ottenendo il miglior ROUGE-L del lotto (10.00) e un BLEU-2 superiore. Questo fenomeno si spiega con il *Catastrophic Forgetting*: in un regime di dati limitati e training breve, il *Full Fine-Tuning* modifica troppo aggressivamente i parametri, "dimenticando" parte della struttura linguistica appresa durante il pre-training. La strategia *Frozen Encoder* agisce quindi come un regolarizzatore strutturale, preservando il Language Model e forzando l'adattatore a colmare esclusivamente il gap modale. Per questo motivo, la configurazione 1.3B Frozen viene identificata come lo "Sweet Spot" architetturale per questo task.

Infine, il confronto con lo Stato dell'Arte rappresentato da *SSVP-SLT* evidenzia un divario prestazionale netto (BLEU-4: 0.58 vs 7.00). Tale differenza non è imputabile all'architettura proposta, bensì alla profonda disparità di risorse. Il modello SOTA beneficia infatti di un pre-training auto-supervisionato su milioni di video YouTube, assente nella nostra pipeline, e dell'utilizzo di un cluster di 64 GPU A100, contro la singola GPU utilizzata in questo studio. Tali vincoli computazionali hanno imposto batch size ridotti, limitando la stabilità dei gradienti necessaria per la convergenza ottimale.

Ciononostante, i risultati validano il modulo come *Proof-of-Concept*, dimostrando che l'approccio *Frozen Encoder + Adapter* è una strategia percorribile per estrarre struttura grammaticale e coerenza semantica anche in ambienti a risorse vincolate.

4.2 Risultati Modulo 2: Architettura Sign-to-Sentiment Multimodale (Il Meta-Learner)

L'architettura proposta per il riconoscimento del sentimento si fonda su una strategia di fusione tardiva (*Late Fusion*) che integra tre macro-componenti distinti: un sottosistema visivo (*Sign-to-Sentiment*), un sottosistema testuale (*Text-to-Sentiment*) e un classificatore di livello superiore (*Meta-Learner*). In questa sezione vengono analizzate le prestazioni di ciascun componente, partendo dai singoli stream unimodali per arrivare alla valutazione della sinergia multimodale.

4.2.1 Analisi dello Stream Visivo (Sign-to-Sentiment)

Il primo blocco funzionale affronta il compito più complesso: inferire il sentimento basandosi esclusivamente sul segnale video, senza alcun supporto semantico. L'obiettivo di questa fase sperimentale è determinare quale rappresentazione dei dati — geometrica (*Keypoints*) o olistica (*Pixel*) — risulti più efficace per cogliere la prosodia visiva della LIS/ASL.

A tal fine, non si è seguito un percorso evolutivo lineare, bensì si sono confrontate diverse famiglie di modelli che rappresentano approcci filosofici distinti allo stato dell'arte. Da un lato, come baseline geometrica, è stata adottata una rete **LSTM (Pose-based)** operante su scheletri estratti con MediaPipe, confrontata anche con un tentativo esplorativo basato su *Spatial-Temporal Graph Convolutional Networks* (ST-GCN) per trattare i landmark come nodi di un grafo. Dall'altro, si è valutato l'approccio **ViViT (Appearance-based)**, basato su Video Vision Transformer fine-tunato sui pixel grezzi, più complesso computazionalmente ma capace di cogliere dettagli di texture. Infine, per completezza, è stato testato **Qwen2.5-VL**, un modello multimodale generalista interrogato in modalità *zero-shot*.

La validazione è stata condotta in due scenari di complessità crescente: classificazione Binaria (Positive vs Negative) e Ternaria (con l'aggiunta della classe Neutral).

Scenario Binario (Positive vs Negative)

Nel primo scenario, il compito è stato semplificato rimuovendo i campioni neutri per valutare la capacità dei modelli di distinguere polarità del Sentiment opposte. La Tabella 6 riassume le metriche quantitative ottenute sul Golden Set.

Modello	Configurazione	wAcc	wF1	Analisi Sintetica
LSTM	Custom	0.5150	0.4703	Performance media, limitata dalla geometria.
ViViT	Fine-Tuned	**0.5650**	**0.5218**	**Miglior risultato complessivo.**
Qwen2.5-VL	Zero-Shot	0.5150	0.3689	Bias sistematico (Mode Collapse).

Tabella 6: Riepilogo delle metriche quantitative nello scenario binario. Il modello ViViT supera la baseline LSTM di oltre 5 punti percentuali in F1-Score (0.5218 vs 0.4703), dimostrando la superiorità dell'approccio appearance-based nel cogliere le micro-espressioni.

Analisi delle Matrici di Confusione Per comprendere la natura degli errori, analizziamo la distribuzione delle predizioni riportata in Tabella 7. L'analisi disaggregata rivela subito il fallimento strutturale del modello generalista Qwen: esso soffre di un bias totale ("Mode Collapse"), classificando il 100% dei campioni come *Positivi*. Questo indica che, senza un fine-tuning specifico, il modello non riesce ad ancorare le feature visive ai concetti emotivi, ripiegando sulla classe statisticamente più probabile.

Al contrario, emerge la superiorità del ViViT sui campioni negativi. Il modello appearance-based identifica correttamente 26 campioni negativi contro i 22 dell'LSTM. Sebbene il numero di falsi positivi rimanga alto per tutti i modelli — sintomo della difficoltà intrinseca del task senza audio o testo — la capacità del ViViT di estrarre segnale dalla classe minoritaria conferma che l'analisi dei pixel è necessaria per decodificare prosodie negative sottili, come il corrugamento della fronte, che la sola geometria a volte perde.

Modello	Classe Reale (Supporto)	Pred: Neg	Pred: Pos	Diagnosi del Comportamento
ViViT	Negative (99)	**26**	73	**Miglior Sensibilità.**
	Positive (101)	14	**87**	Unico modello a distinguere pattern negativi.
LSTM	Negative (99)	22	77	Bias verso la classe positiva.
	Positive (101)	20	81	Inferiore a ViViT nel riconoscimento dei Negativi.
Qwen2.5	Negative (99)	0	99	**Bias Totale (100% Positivo).**
	Positive (101)	0	101	Il modello ignora completamente la classe negativa.

Tabella 7: Dettaglio delle matrici di confusione per lo scenario binario. Si noti il "Mode Collapse" di Qwen, che predice esclusivamente la classe Positiva. Al contrario, ViViT dimostra la migliore capacità di generalizzazione identificando il maggior numero di campioni Negativi (26).

Scenario Ternario (Positive vs Neutral vs Negative)

L'introduzione della classe **Neutral** incrementa drasticamente la complessità dello spazio decisionale. Come si osserva confrontando le Tabelle 6 e 8, le prestazioni generali subiscono un calo fisiologico: l'aggiunta di un'area di ambiguità visiva (il neutro) rende più difficile la separazione delle classi per modelli addestrati su dataset limitati, evidenziando le difficoltà di generalizzazione su dati mai visti.

I risultati comparativi offrono comunque una panoramica chiara, posizionando i nostri modelli rispetto all'intero spettro delle baseline dello stato dell'arte. Sorprendentemente, l'LSTM sviluppato in questa tesi ottiene il Weighted F1-Score più alto in assoluto (**0.4243**), superando non solo le architetture visive pure, ma anche i Large Multimodal Models più avanzati come GPT-4o e MiniGPT4 quando operano in modalità *video-only*. Questo risultato si spiega con la natura del segnale: mentre gli LLM generalisti cercano invano concetti semantici in un video muto, l'LSTM si focalizza esclusivamente sulla cinematica dei keypoints facciali e corporei (velocità del segno, postura), che sono correlati diretti dell'intensità emotiva.

Modello	**Configurazione**	**wAcc**	**wF1**	**Analisi Comparativa**
		Nostri Esperimenti		
LSTM	Custom	**0.4800**	**0.4243**	**SOTA (Video-Only).**
ViViT	Fine-Tuned	0.3300	0.3424	Performance bilanciate ma inferiori a LSTM.
Qwen2.5-VL-3B	Zero-Shot	0.1050	0.0597	Nostra riproduzione locale (Fallimento).
		Benchmark EmoSign (Video-Only)		
MiniGPT4	Zero-Shot	0.3468	0.4000	Competitivo, ma inferiore al nostro LSTM.
GPT-4o	Zero-Shot	0.4072	0.2443	Soffre senza input testuale.
Qwen 2.5	Zero-Shot	0.2734	0.1647	Performance scarse.
AffectGPT	Zero-Shot	0.3333	0.0004	Collasso totale (predice solo una classe).

Tabella 8: Benchmark comparativo esteso nello scenario ternario. L'LSTM sviluppato in questa tesi ottiene il Weighted F1-Score più alto in assoluto (0.4243), superando anche i Large Multimodal Models (incluso GPT-4o).

Il Ruolo della Specializzazione: Analisi degli Errori Un wF1 alto non implica perfezione. Disaggregando le prestazioni tramite la matrice di confusione (Tabella 9), emergono due profili di specializzazione distinti. L'LSTM agisce come uno "Specialista del Negativo": mostra un'aggressività predittiva che, pur generando falsi positivi, garantisce un Recall eccezionale sulla classe negativa (94 su 121 identificati). Ha imparato a riconoscere efficacemente i marcatori cinetici forti, come movimenti bruschi, tipici della rabbia o frustrazione. Al contrario, il ViViT si comporta da generalista "debole": tenta di modellare la distribuzione reale delle tre classi ma fallisce nella sensibilità, identificando solo 18 campioni negativi su 121 e confondendo spesso la

negatività con la neutralità. Qwen2.5-VL conferma invece il trend negativo, collassando quasi interamente sulla classe neutra.

Modello	Classe Reale	Pred: Neg	Pred: Neu	Pred: Pos	Profilo di Specializzazione
LSTM	Negative (121)	**94**	21	6	**Altissima Sensibilità al Negativo.**
	Neutral (3)	3	**0**	0	Tende a interpretare il neutro come negativo.
	Positive (76)	62	12	**2**	Bias conservativo verso il negativo.
ViViT	Negative (121)	**18**	47	56	**Debolezza critica sui Negativi.**
	Neutral (3)	1	**2**	0	Migliore equilibrio distributivo,
	Positive (76)	10	20	**46**	ma troppi falsi positivi/neutri.
Qwen2.5-VL	Negative (83)	0	83	0	**Mode Collapse (Neutro).**
	Neutral (16)	0	16	0	Predice quasi sempre Neutro (194/200).
	Positive (101)	0	95	6	Inutile per la classificazione.

Tabella 9: Analisi disaggregata delle predizioni. Si evidenzia la specializzazione complementare: l'LSTM agisce come un potente "Rilevatore di Negatività" (recuperando 94/121 casi), mentre il ViViT perde la stragrande maggioranza dei segnali negativi.

Selezione del Modello per il Meta-Learner Alla luce dei risultati sperimentali, il modello **LSTM** è stato selezionato come componente visivo definitivo per l'architettura multimodale. La scelta non si basa solo sulla dominanza prestazionale nello scenario ternario (Best-in-Class con wF1 0.4243), ma anche su fattori architetturali strategici. In primis, l'uso di dati geometrici implementa nativamente il principio di *Privacy by Design*: analizzando solo le coordinate scheletriche, l'LSTM permette di riconoscere la prosodia senza processare (o trasmettere) le immagini del volto del segnante, un vantaggio cruciale in contesti assistivi sensibili. Inoltre, questo modello offre la migliore complementarietà informativa: mentre il testo (analizzato nel prossimo modulo) veicola il "cosa" viene detto, l'LSTM cattura efficacemente il "come" (velocità, intensità), fornendo al Meta-Learner un segnale ortogonale essenziale per modulare l'espressività della sintesi vocale.

4.2.2 Analisi dello Stream Testuale (Text-to-Sentiment)

Il secondo pilastro dell'architettura multimodale si concentra sull'analisi semantica. A differenza del modulo visivo che interpreta la prosodia, questo sottosistema elabora le trascrizioni testuali (*Ground Truth*) per inferire la polarità prosodica intrinseca del messaggio lessicale. L'obiettivo sperimentale è identificare il *Language Model* più adatto a gestire la sintassi peculiare delle traduzioni da Lingua dei Segni, confrontando approcci pre-addestrati (*Zero-Shot*) con approcci specializzati (*Fine-Tuned*).

Per coprire lo spazio delle soluzioni possibili, sono state valutate quattro configurazioni distinte. Come baseline granulare è stato testato **BERT-GoEmotions**, addestrato su 28 classi e rimappato sul target ternario. Successivamente, si è passati a **RoBERTa-Base** (nella variante Twitter), scelto per la sua natura nativa a 3 classi, e alla sua versione

Fine-Tuned, sottoposta a un ulteriore stadio di addestramento supervisionato sul nostro *Silver Set*. Infine, è stato interrogato **Gemini 3.0 Pro** come riferimento per valutare le capacità di ragionamento semantico avanzato di un LLM generalista.

Benchmark Quantitativo delle Architetture Testuali

La Tabella 10 presenta il confronto delle metriche globali ottenute sul Golden Set.

Modello	Configurazione	wAcc	wF1	Kappa	Analisi Sintetica
Bert Go	Pre-trained	0.3719	0.4886	0.2012	Performance scarse (Disallineamento).
RoBERTa	Pre-trained	0.5777	0.7715	0.4899	Baseline solida ma non eccellente.
RoBERTa	**Fine-Tuned**	0.5858	**0.8475**	**0.7035**	**Best Model.**
Gemini 3.0	Zero-Shot	**0.6107**	0.8145	0.5603	Ottima generalizzazione senza training.

Tabella 10: Confronto delle metriche di performance sul task Text-to-Sentiment a 3 classi. Il modello RoBERTa Fine-Tuned domina la classifica per Weighted F1 (0.8475) e Kappa (0.70), metriche cruciali per l'affidabilità, sebbene Gemini 3.0 ottenga una Balanced Accuracy leggermente superiore grazie a una migliore gestione della classe neutra.

I dati evidenziano una progressione prestazionale netta. Il modello **Bert-GoEmotions** mostra evidenti limiti, con una Balanced Accuracy di appena 0.37, suggerendo che l'aggregazione di 28 micro-emozioni in 3 macro-classi introduce un rumore eccessivo. Il passaggio da RoBERTa Pre-trained a **RoBERTa Fine-Tuned** segna invece un incremento sostanziale nella robustezza, portando il Kappa di Cohen da 0.48 (accordo moderato) a 0.70 (accordo sostanziale): questo dimostra che l'addestramento sul dominio specifico delle traduzioni ASL è determinante per disambiguare il sentiment. Infine, l'uso di **Gemini 3.0 Pro** conferma che la capacità di ragionamento generalista è molto efficace nel gestire i casi limite (miglior Accuracy 0.61), anche se il modello specializzato (RoBERTa FT) rimane superiore nella metrica globale wF1.

Analisi della Distribuzione degli Errori e Complementarietà

Per validare la scelta del modello per il Meta-Learner, è essenziale analizzare dove i modelli falliscono. La Tabella 11 mostra l'evoluzione del comportamento predittivo, evidenziando come RoBERTa Fine-Tuned riesca a separare nettamente le polarità.

Modello	Classe Reale	Pred: Neg	Pred: Neu	Pred: Pos	Profilo di Errore
Bert Go	Negative	31	**82**	8	**Bias Critico verso il Neutro.**
	Neutral	1	1	1	Interpreta l'incertezza come neutralità
	Positive	0	36	40	nella maggior parte dei casi.
RoBERTa Base	Negative	85	25	11	**Alta Confusione.**
	Neutral	1	1	1	Migliora sui negativi ma confonde
	Positive	5	18	53	spesso i positivi con i neutri.
RoBERTa FT	Negative	**98**	1	22	**Eccellenza Bimodale.**
	Neutral	1	0	2	Recupera quasi tutti i Negativi (98)
	Positive	4	0	**72**	e quasi tutti i Positivi (72).

Tabella 11: Dettaglio delle matrici di confusione. Si osserva l'evoluzione del decision boundary: Bert Go (sopra) collassa sul Neutro; RoBERTa Base (centro) inizia a separare le classi ma mantiene alta incertezza; RoBERTa Fine-Tuned (sotto) mostra una separazione netta e decisa tra le polarità, minimizzando la classe neutra.

L'analisi incrociata tra i risultati di **RoBERTa Fine-Tuned** e quelli dell'LSTM (discusso nella Sezione 4.2.1) rivela la logica fondante dell'architettura proposta. Il contributo più critico del modello testuale emerge sulla classe **Positive**: mentre l'LSTM falliva quasi totalmente (identificando solo 2 campioni su 76), RoBERTa FT ne identifica correttamente 72 su 76. Questo conferma che la gioia e l'approvazione sono veicolate primariamente dalla semantica esplicita (parole come "happy", "good", "love") piuttosto che da marcatori cinetici corporei univoci.

Parallelamente, RoBERTa mostra un recall eccezionale anche sui negativi (98 identificati), leggermente superiore all'LSTM. Tuttavia, i 22 falsi positivi rappresentano casi insidiosi di ironia o negazioni complesse che il testo da solo non risolve. Qui entra in gioco la complementarietà con l'LSTM: un corpo agitato su una frase apparentemente positiva sarà il segnale per il Meta-Learner di correggere la predizione verso il negativo. Infine, va notato il fallimento generalizzato sulla classe *Neutral* (0/3 identificati), un risultato statisticamente poco significativo data la scarsità di campioni, ma che suggerisce una tendenza del sistema a polarizzare le decisioni.

In conclusione, **RoBERTa Fine-Tuned** viene selezionato come stream testuale definitivo per la sua capacità quasi chirurgica di riconoscere i Positivi, colmando la lacuna principale dell'LSTM e creando una coppia di "esperti" ideale per la fusione nel Meta-Learner.

4.2.3 Performance del Meta-Learner (Integrazione Multimodale)

L'ultimo stadio della pipeline affronta la sfida dell'integrazione decisionale tramite il *Meta-Learner*, un classificatore di livello superiore progettato per orchestrare e pesare dinamicamente i segnali provenienti dai due "esperti" unimodali selezionati: lo stream visivo basato su *LSTM*, scelto per la sensibilità ai marker cinetici della negatività, e lo stream testuale basato su *RoBERTa Fine-Tuned*, robusto nel distinguere

le polarità positive. L'input del Meta-Learner non è costituito dai dati grezzi, bensì dallo spazio latente dei *logits* generati dai modelli a monte. Questo approccio di *Late Fusion* permette al meta-modello di apprendere l'*affidabilità relativa* dei due esperti in funzione del contesto, risolvendo i conflitti decisionali (ad esempio, privilegiando RoBERTa quando l'LSTM mostra incertezza su segnali ambigui).

Selezione dell'Architettura di Fusione (Model Selection)

Per individuare la strategia di fusione ottimale, è stata condotta una *Grid Search* sul Training Set, confrontando diverse famiglie di classificatori per mappare gli input probabilistici nell'etichetta finale. La Tabella 12 riporta le performance medie in validazione.

Meta-Modello	Architettura	Balanced Acc.	wF1	Kappa	Note
Logistic Regression	*Lineare*	**0.7539**	**0.7546**	**0.5075**	**Miglior Modello (Best Fit).**
Random Forest	*Ensemble*	0.7499	0.7514	0.5070	Competitivo ma più complesso.
Decision Tree	*Albero*	0.7499	0.7496	0.4954	Performance identica a RF.
SVM	*Kernel*	0.7343	0.7393	0.4900	Inferiore nella ricerca iperparametri.

Tabella 12: Risultati della Model Selection per il Meta-Learner. La Regressione Logistica ottiene le performance migliori (wF1 0.7546). Il fatto che un modello lineare superi approcci non lineari suggerisce che i modelli a monte abbiano già linearizzato efficacemente lo spazio delle feature.

I risultati indicano la **Regressione Logistica** come architettura ottimale. La sua superiorità rispetto a modelli più complessi come Random Forest o SVM implica che la relazione tra le predizioni degli esperti e la verità sia lineare e diretta. Il Meta-Learner agisce di fatto come un calibratore di fiducia, assegnando pesi scalari alle probabilità di LSTM e RoBERTa. In virtù del principio di parsimonia e della maggiore interpretabilità, questo modello è stato promosso alla fase di test.

Validazione Finale e Analisi degli Errori

La validazione conclusiva sul Test Set (Golden) sancisce l'efficacia della pipeline. Come evidenziato nella Tabella 13, il nostro sistema supera lo Stato dell'Arte rappresentato da GPT-4o nella metrica chiave Weighted F1-Score (0.8091 vs 0.7672). Questo risultato conferma la tesi di fondo del lavoro: un sistema composto da "piccoli" modelli specializzati e finemente addestrati batte un modello generalista massivo nel dominio specifico delle lingue dei segni.

Modello	Architettura	Modalità	wAcc	wF1	Note
Meta-Learner (Ours)	*Hybrid Ensemble*	**Video + Text**	**0.5395**	**0.8091**	**Nuovo SOTA su EmoSign**
GPT-4o	*LMM (OpenAI)*	Video + Text	0.5213	0.7672	Benchmark Paper EmoSign
AffectGPT	*LMM*	Video + Text	0.5618	0.6437	Ref. Paper EmoSign
Qwen 2.5	*LMM*	Video + Text	0.4110	0.5429	Ref. Paper EmoSign
MiniGPT4	*LMM*	Video + Text	0.2165	0.3689	Ref. Paper EmoSign

Tabella 13: Confronto finale sul Test Set (Golden). Il Meta-Learner supera lo Stato dell'Arte (GPT-4o), dimostrando l'efficacia dell'architettura modulare.

Per comprendere la dinamica interna della fusione, analizziamo la matrice di confusione finale (Tabella 14). L'analisi dei flussi di errore rivela il successo della strategia di integrazione, evidenziando un comportamento robusto sulle classi polari. Sulla classe **Positive**, il sistema eredita la precisione di RoBERTa, raggiungendo un Recall del 90.8% (69 su 76 campioni corretti) e minimizzando drasticamente i falsi positivi che affliggevano l'LSTM. Parallelamente, mantiene una solida sensibilità sui **Negativi** (Recall 71.1%), identificando correttamente 86 campioni su 121. I casi di errore residui rappresentano lo "zoccolo duro" dell'ambiguità multimodale, dove probabilmente ironia o espressioni facciali contrastanti (es. sorrisi di imbarazzo) generano segnali ingannevoli.

GT \Pred	Negative	Neutral	Positive	Recall per Classe
Negative	**86**	16	19	71.1%
Neutral	1	**0**	2	0.0%
Positive	3	4	**69**	90.8%

Tabella 14: Matrice di Confusione del Meta-Learner (Logistic Regression) sul Test Set. Il sistema mostra un comportamento bilanciato sulle due classi dominanti.

Infine, si conferma la tendenza a polarizzare le decisioni, con zero predizioni corrette per la classe **Neutra**. Questo comportamento suggerisce che il Meta-Learner, addestrato su dati fortemente sbilanciati, ha appreso un'ipotesi del mondo essenzialmente bipolare, forzando le decisioni verso le emozioni forti. In sintesi, il Meta-Learner realizza con successo la sintesi tra i due stream: stabilizza la predizione sui positivi riducendo il rumore visivo e preserva la capacità di rilevare i negativi, risultando complessivamente più accurato e bilanciato di qualsiasi sua singola parte.

4.3 Risultati Modulo 3: Sintesi Vocale Espressiva "Sentiment Aware"

L'ultimo stadio della pipeline sperimentale riguarda la validazione della sintesi vocale prosodica generata dal modello Bark. A differenza dei moduli precedenti, valutabili

tramite metriche oggettive, la qualità della sintesi prosodica è una grandezza intrinsecamente soggettiva, legata alla percezione uditiva umana. L'analisi mira a rispondere a tre quesiti fondamentali: verificare la *stabilità tecnica* del generatore (intelligibilità), misurare la *validità* della percezione umana rispetto al Ground Truth e, soprattutto, quantificare l'*efficacia espressiva*, intesa come la coerenza tra la prosodia generata e il contenuto semantico.

La validazione ha seguito un rigido protocollo a imbuto (*funnel*), partendo dall'intero dataset EmoSign per arrivare a un sottoinsieme curato di coppie audio-testuali sottoposte al giudizio di due valutatori indipendenti.

4.3.1 Generazione e Selezione dei Campioni (Data Selection)

Il processo di selezione è stato progettato per isolare la variabile "espressività", garantendo che gli annotatori valutassero solo materiale tecnicamente valido e privo di difetti di pronuncia o allucinazioni acustiche.

La pipeline è iniziata con la generazione dei file audio per l'intero dataset EmoSign (200 campioni), utilizzando le trascrizioni e le etichette di sentimento originali. Su questi è stato applicato un filtro automatico di intelligibilità basato su *Whisper ASR*, imponendo una soglia rigorosa di **Word Error Rate (WER)** ≤ 0.1. Il filtro ha trattenuto 107 campioni validi, scartandone il 46.5%. Questo alto tasso di rifiuto conferma l'instabilità delle attuali architetture TTS generative basate su token discreti: quando forzato a esprimere sentimenti intensi, il modello tende talvolta a compromettere la struttura fonetica. Il filtro WER si è dunque rivelato essenziale per scremare gli errori semantici a monte.

Dal bacino dei 107 audio validi, è stato estratto un sottoinsieme bilanciato di 54 coppie (Audio + Testo) da sottoporre agli annotatori. Durante la fase di ascolto (eseguita in modalità *blind* e disaccoppiata), i valutatori hanno applicato un ulteriore filtro manuale, flaggando 9 campioni affetti da artefatti non linguistici (rumori metallici, respiri eccessivi) sfuggiti all'ASR. L'analisi statistica presentata di seguito è calcolata sul dataset finale di **45 coppie** (90 valutazioni totali per annotatore) doppiamente validate.

4.3.2 Validazione Percettiva e Analisi Psicometrica

La valutazione soggettiva si articola in quattro livelli di analisi: la definizione della baseline umana, la misurazione della coerenza interna, l'identificazione dei bias e la verifica dell'affidabilità inter-annotatore.

Validazione rispetto al Gold Standard (Human Baseline)

Prima di valutare l'audio, è fondamentale stabilire una *Human Baseline*: quanto sono abili gli esseri umani a dedurre il sentimento leggendo esclusivamente la trascrizione? La Tabella 15 mostra le metriche di concordanza tra i voti *Text-Only* e il Ground Truth.

Annotatore	**Modalità**	**Pearson (r)**	**Spearman (ρ)**	**wKappa**	**MAE (Punti)**
Annotatore 1	Text Only	0.790	0.805	0.556	1.000
Annotatore 2	Text Only	0.772	0.765	0.546	1.133

Tabella 15: Validazione Umana vs Ground Standard. L'alta correlazione di Pearson (≈ 0.78) conferma che il testo è un predittore forte. Il Weighted Kappa (≈ 0.55) indica un accordo "Moderato", mentre il MAE definisce una "soglia fisiologica" di errore (≈ 1.0) intrinseca alla soggettività umana.

I valori di correlazione ($r \approx 0.78$) validano la scelta di utilizzare le trascrizioni come base per la sintesi, confermando che il testo veicola gran parte dell'informazione emotiva. Tuttavia, il Weighted Kappa (0.55) rivela un margine di interpretazione ineliminabile: diverse persone percepiscono sfumature di intensità diverse nello stesso testo, stabilendo un limite superiore realistico per le prestazioni attese dalla macchina.

Coerenza Interna e Analisi dei Bias

Il cuore della validazione risiede nella **Coerenza Interna**, ovvero la misura in cui il sentimento percepito dall'audio correla con quello del testo. Come mostrato in Tabella 16, la correlazione di Pearson si attesta stabilmente sopra 0.73, indicando un forte legame lineare: il sistema ha tradotto correttamente la direzione del sentimento (positivo/negativo) in prosodia senza introdurre dissonanze.

Annotatore	**Coppie (N)**	**Kappa (Weighted)**	**Pearson (r)**	**Spearman (ρ)**
Annotatore 1	45	0.529	0.722	0.743
Annotatore 2	45	0.581	0.751	0.737
Media	-	0.555	0.736	0.740

Tabella 16: Analisi della Coerenza Interna (Audio vs Testo). La forte correlazione lineare conferma l'efficacia del condizionamento prosodico.

Tuttavia, il valore moderato del Kappa suggerisce discrepanze sull'intensità esatta. Per approfondire, è stato analizzato il **Trend del Bias** (Figura 3). La curva sigmoidale evidenzia un fenomeno di compressione della gamma dinamica (*Central Tendency Bias*): quando il sentimento originale è estremo (+3/-3), l'audio generato viene percepito come più moderato (+1.5/-2). Questo limite è attribuibile all'architettura di Bark che, pur espressiva, non raggiunge ancora l'escursione drammatica di un attore umano.

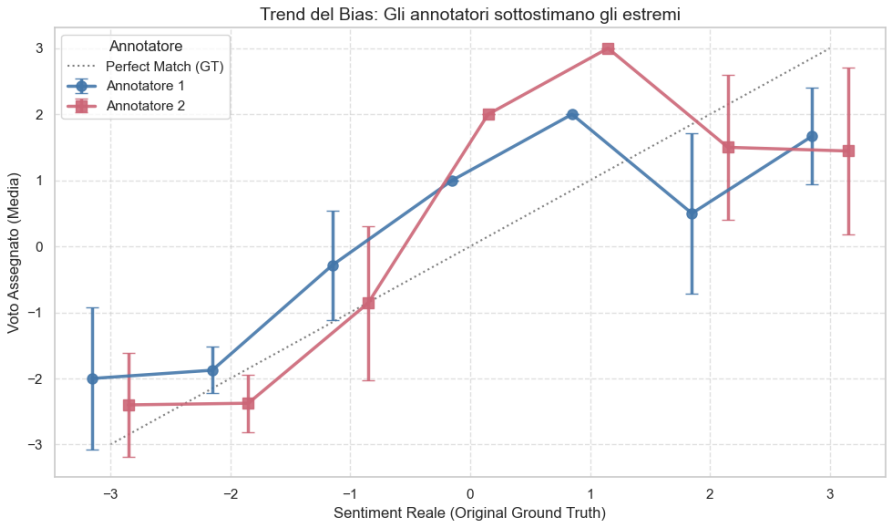


Figura 3: Trend del Bias: Confronto tra Sentiment Reale e Percepito. Si nota un appiattimento agli estremi: i sentimenti molto forti vengono percepiti come più moderati nell'audio sintetico.

Affidabilità dell'Annotazione (Inter-Rater Agreement)

Infine, per garantire la solidità scientifica dei risultati, è stato calcolato l'accordo tra i due annotatori (Tabella 17). L'accordo globale elevato ($r = 0.74$) conferma la robustezza del protocollo. Si osserva un fisiologico calo nell'accordo *Audio Only* ($r = 0.66$) rispetto al *Text Only* ($r = 0.83$), confermando che la percezione della prosodia è intrinsecamente più soggettiva della comprensione semantica, ma comunque statisticamente significativa per validare il sistema.

Modalità	**N. Campioni**	**Pearson (r)**	**Spearman (ρ)**	**Kappa (w)**
GLOBALE	**90**	**0.743**	**0.742**	**0.517**
Text Only	45	0.836	0.834	0.613
Audio Only	45	0.664	0.657	0.415

Tabella 17: Confronto dell'accordo tra annotatori. L'accordo Globale elevato valida il protocollo, pur evidenziando la maggiore soggettività del canale audio.

Capitolo 5

Discussione e Conclusioni

L'analisi complessiva dei risultati conseguiti in questo lavoro di tesi offre una prospettiva stratificata e complessa sulle sfide odierne della traduzione multimodale. Ripercorrendo il percorso di ricerca, emerge chiaramente come il tentativo di colmare il divario comunicativo tra la Comunità Sorda e il mondo udente richieda un approccio che vada ben oltre la semplice trasposizione lessicale, abbracciando la dimensione affettiva del linguaggio.

Per quanto riguarda il primo modulo, dedicato alla *Sign Language Translation* (SLT), i dati raccolti indicano inequivocabilmente che il dominio è ancora distante dal raggiungere prestazioni operative accettabili per scenari di utilizzo reale. Le criticità riscontrate non riguardano solamente l'usabilità dei modelli ottimizzati in regime di risorse vincolate, ma si estendono strutturalmente anche ai macro-modelli considerati lo stato dell'arte. È emerso, infatti, un paradosso tecnologico: anche architetture che richiedono risorse computazionali ingenti mostrano metriche (come il BLEU-4) appena sufficienti a garantire una minima coerenza sintattica. Nel nostro specifico contesto sperimentale, l'adozione della strategia *Frozen Encoder* si è rivelata un compromesso più efficace rispetto al *Full Fine-Tuning*, permettendo di preservare le conoscenze linguistiche preesistenti del modello. Tuttavia, resta evidente una discrepanza tra l'obiettivo ideale di una traduzione fluida e la realtà attuale, in cui la scarsità di dataset annotati su larga scala rappresenta il principale collo di bottiglia per l'apprendimento delle complesse grammatiche spaziali.

Spostando l'attenzione sul riconoscimento del sentimento, il contributo più significativo del presente lavoro risiede nella validazione della pipeline di creazione del dataset *silver-labeled*. Questa risorsa si è dimostrata strategica per mitigare la cronica scarsità di dati annotati, dimostrando la fattibilità di addestrare modelli performanti su etichette generate automaticamente. La validazione su trascrizioni *Ground Truth* ha evidenziato la complementarietà intrinseca tra i canali: il testo fornisce la polarità,

mentre il video ne modula l'intensità. Questa sinergia è fondamentale non solo per l'arricchimento semantico in condizioni ideali, ma si rivela determinante per garantire la robustezza del sistema in futuri scenari operativi, dove l'analisi visiva potrà sopperire a potenziali imprecisioni della traduzione automatica. In questo ambito, la valutazione delle architetture visive ha offerto spunti interessanti: sebbene la scarsità di dati attuali favorisca le performance delle reti LSTM — che, basandosi sulla cinematica scheletrica, agiscono come eccellenti rilevatori di negatività garantendo al contempo la privacy — il modello ViViT ha evidenziato una distribuzione delle classi più bilanciata e robusta. Tale caratteristica rende l'approccio *pixel-based* l'architettura tecnologicamente più promettente per il futuro, qualora si disponga di volumi di dati sufficienti a cogliere le micro-espressioni facciali che la scheletrizzazione inevitabilmente ignora. Il successo del *Meta-Learner* sviluppato, capace di superare modelli generalisti come GPT-4o sul benchmark di riferimento, valida la scelta di un'architettura modulare composta da esperti specializzati.

Un elemento di forte novità in letteratura è costituito dallo studio sulla sintesi vocale espressiva. I risultati ottenuti risultano molto promettenti per l'integrazione di sistemi di generazione vocale *Sentiment-Aware* in tecnologie assistive. L'analisi psicometrica condotta ha dimostrato che, sebbene esista un fisiologico bias di "smorzamento" per cui i sentimenti estremi vengono percepiti come più moderati nell'audio sintetico, la direzione emotiva viene preservata con coerenza. Questo conferma che il condizionamento paralinguistico supera efficacemente la monotonia dei sistemi *Text-to-Speech* standard, restituendo una dimensione empatica alla comunicazione mediata dalla tecnologia.

Riflettendo sulle discrepanze tra gli obiettivi pianificati e i risultati raggiunti, è doveroso notare come i vincoli hardware e la citata immaturità dei sistemi di traduzione abbiano imposto un adattamento metodologico, portando a disaccoppiare parzialmente i moduli per valutarne le potenzialità teoriche al netto degli errori di propagazione. Ciononostante, il lavoro conferma la validità dei moduli *"Classificazione Multimodale del Sentimento"* e *"Sintesi Vocale Sentiment-Aware"* nel suo complesso.

5.1 Oltre la Traduzione: Frontiere e Implicazioni Etiche

La realizzazione di un ecosistema di comunicazione inclusivo impone di guardare oltre la semplice interpretazione del segnale oggetto di questa tesi. I risultati ottenuti aprono la strada verso la chiusura del cerchio comunicativo, la produzione o *Sign Language Production* (SLP), ma richiedono contestualmente una riflessione critica sull'impatto sociale di queste tecnologie.

Dalla Voce all'Avatar: Le Sfide del Text-to-Sign

Mentre questo lavoro si è concentrato sulla sintesi vocale emotiva (*Sign-to-Speech*), la controparte generativa visiva (*Text-to-Sign*) tramite avatar 3D rappresenta la naturale evoluzione per restituire un feedback all'utente Sordo. Tuttavia, analogamente a quanto osservato per la sintesi vocale, anche la SLP deve affrontare sfide cinematiche uniche per evitare l'effetto "robotico".

La prima criticità è il **problema della co-articolazione**. Un segno non è un'entità statica isolata; la sua esecuzione varia drasticamente in base al contesto. I futuri modelli dovranno apprendere a generare transizioni fluide (*movement epenthesis*) tra i segni, una complessità simile a quella affrontata nel nostro modulo audio per legare i token prosodici. La sfida maggiore risiede però nella **sincronizzazione delle Componenti Non-Manuali (NMF)**. Generare un avatar significa orchestrare flussi paralleli: mani, volto e postura. Come evidenziato da Saunders et al. [28], la mancanza di espressività facciale è spesso causa di rifiuto della tecnologia. Il nostro studio sull'intensità emotiva potrebbe fungere da base per controllare anche l'espressività degli avatar, garantendo una corretta **prosodia visiva**, un'area ancora inesplorata rispetto alla controparte vocale.

Implicazioni Etiche e Co-progettazione

L'esperienza maturata con i dataset utilizzati in questa tesi (come How2Sign e ASLLRP) impone una riflessione etica che non rappresenta un'appendice, ma un pilastro centrale. La letteratura recente e i nostri stessi esperimenti confermano la criticità dei **bias nei dataset**. Risorse come How2Sign soffrono spesso di una scarsa rappresentatività in termini di età, etnia, varianti dialettali e condizioni di illuminazione. Allenare modelli su dati così omogenei rischia di creare sistemi meno robusti con utenti reali appartenenti a minoranze, perpetuando involontariamente disuguaglianze digitali.

Per mitigare questi rischi e garantire uno sviluppo tecnologico equo, l'approccio del *co-design* appare oggi la via maestra. Come sottolineato da Mugele et al. [4], il coinvolgimento attivo delle persone Sorde non deve limitarsi alla valutazione finale, ma deve permeare l'intero ciclo di sviluppo. Lo slogan "*Nothing About Us Without Us*" guida verso la creazione di strumenti che siano non solo tecnicamente validi, ma socialmente accettati e realmente utili alla comunità di riferimento.

5.2 Conclusioni e Sviluppi Futuri

Alla luce di queste considerazioni etiche e tecniche, e guardando agli sviluppi futuri, la ricerca dovrà mirare prioritariamente all'espansione del corpus dati, sfruttando tecniche di *Self-Supervised Learning* su video non etichettati per sbloccare il pieno potenziale

delle architetture Transformer. Parallelamente, sarà cruciale ottimizzare i tempi di inferenza per abilitare scenari *real-time*.

Le ricadute applicative di questi avanzamenti sono vaste e spaziano dagli assistenti vocali inclusivi, capaci di dare una voce ”umana” e coerente allo stato d'animo dell'utente, fino a strumenti educativi per l'apprendimento della prosodia visiva. In conclusione, sebbene la strada verso una traduzione perfetta sia ancora lunga, questo studio dimostra che l'Intelligenza Artificiale può evolvere da semplice strumento di decodifica a ponte empatico, capace di interpretare e trasmettere non solo il contenuto, ma anche l'intenzione emotiva della comunicazione umana.

Ringraziamenti

Desidero ringraziare in primis il mio Relatore, **Luca Bacco**, non solo per aver sostenuto la mia idea e per la sua indiscussa competenza, ma anche per essersi prestato a lunghe serate di lavoro in laboratorio senza mai battere ciglio, instaurando un rapporto di collaborazione che è andato ben oltre la formalità accademica, diventando autenticamente amichevole. Ringrazio doverosamente i miei Correlatori, l'Ing. **Mario Merone** e il Dott. **Daniele Sasso**, il cui contributo è stato determinante per navigare le complessità tecniche di questo lavoro e per avermi spinto a migliorare costantemente la qualità della tesi.

Un ringraziamento sincero va a tutto il **corpo docente** incontrato in questo percorso. Grazie per gli insegnamenti trasmessi e per aver lasciato, ognuno a suo modo, un segno nella mia formazione professionale e personale.

Ora voglio dedicare questo spazio a parole che, forse, non avrei la forza o il carattere di pronunciare a voce. Potrà sembrare lungo, o forse noioso, ma come scriverò più avanti: *"Se no, che senso ha?"*

Il mio grazie più sentito va ai miei **nonni**, per le persone uniche che sono e per la **famiglia** che hanno costruito.
Grazie a **Nonno Peppe**, da cui ho ereditato la tranquillità e quella pazienza che ormai ci contraddistingue.
Grazie a **Nonna Pierina**, per avermi trasmesso la manualità, il sorriso e la forza di rialzarsi sempre, anche nei momenti più bui.
Grazie a **Nonno Ignazio**, per avermi trasmesso l'instancabilità nel lavoro e quella capacità unica di combattere la noia inventandosi imprese fuori dal comune, talvolta al limite della pazzia.
E grazie a **Nonna Maria** (*"u chiù grossu fattu è"*), da cui ho preso il pregio, o forse il difetto, di preoccuparmi per ogni cosa e di voler bene a tutti indistintamente.

Mi chiedo: *"Alcuni di questi sono dei difetti?"*
Mi rispondo: *"Cu' si ni futti"*.
"Quali sono i valori a cui scegliamo di dare importanza? Io ho scelto di esaltare questi."

Il ringraziamento più profondo va alla mia **famiglia**.
A **Mamma e Papà**, per avermi sempre lasciato libero di scegliere la mia strada e per i valori che mi avete trasmesso attraverso il vostro esempio. Grazie soprattutto per aver accettato il mio "telefono silenzioso". Grazie per aver compreso i miei silenzi senza farmene una colpa e per avermi aspettato sempre, con pazienza.
Alle mie sorelle, **Cristina e Sophia**, le mie prime compagne di viaggio. Grazie per aver condiviso con me la crescita, i fallimenti, i successi, e anche i litigi furiosi ma necessari che ci hanno reso più forti. Siete parte essenziale di chi sono oggi.
Un pensiero speciale va anche a **Rosario**, futuro marito di Cristina. Grazie per la persona meravigliosa che sei e per le cose splendide che state costruendo insieme.

Un abbraccio va a tutta la mia famiglia allargata, **zii e cugini**. Grazie per la famiglia che siamo. Riconosco la mia grande colpa, quella di essere spesso troppo silenzioso e distante. Ma ho capito che, quando siamo **tutti insieme**, la mia vera felicità sta proprio lì: nell'ascoltarvi e nel vederci uniti.

Un pensiero affettuoso va a **Ezio e Bruna**, che abitano ancora lì, a due passi dalla mia vecchia scuola. Grazie a Ezio, per la gioia di tutte le campanelle che mi faceva suonare, e a Bruna, per le cene squisite che sanno di casa. Ogni volta che siedo alla vostra tavola si sblocca un ricordo e, per un attimo, ritorno *quel* bambino.

Un grazie speciale a **Vincenzo**, per la sua infinita pazienza, per il supporto costante e per la forte amicizia che abbiamo cementato in questi anni. Grazie per avermi insegnato una lezione fondamentale: *non sempre possiamo controllare ciò che ci accade, ma abbiamo sempre il controllo su come interpretarlo e su come decidiamo di reagire.*

Grazie a **Tanisha**, conosciuta durante quei giorni di trekking sulle Dolomiti. Grazie per avermi fatto scoprire la solarità dell'essere umano e per la dolcezza e la sopportazione che mi hai riservato in questo periodo, dandomi la prova concreta che **l'amore non ha limiti**.

Un pensiero lo dedico ai miei compagni di viaggio universitari: **Fabio, Michela, Virginia, Mattia, Lorenzo, Alessia e Debora**. Grazie per essere le persone meravigliose che siete. Grazie per aver condiviso con me ogni singolo momento, dai sacrifici e l'ansia prima degli esami alle esperienze indimenticabili vissute insieme. Ognuno di voi ha lasciato un pezzettino di sé nel mio carattere.

Ringrazio i miei **coinquilini**, con cui ho avuto il piacere di condividere mille avventure; grazie per avermi spinto a uscire dalla zona di comfort e per avermi insegnato ad accettare e rispettare le necessità e gli spazi altrui.

Un pensiero va al gruppo di **Share the Light**: nonostante il mio essere agnostico, mi avete fatto riscoprire sotto una luce diversa la capacità che un credo comune ha di creare comunità e di stringere amicizie così forti e vere.

Forse ho citato troppe persone. Probabilmente anche persone che, per un'amicizia finita, per la distanza o per l'ineluttabilità della vita, un giorno non faranno più parte del mio quotidiano. Ma scrivo queste righe proprio per non dimenticare chi ho incontrato. Ognuno di voi ha plasmato, in piccola o grande parte, l'Ignazio del presente e del futuro. È con voi che sono cresciuto in questo percorso. E d'altronde, *se no, che senso ha?*

Tutto questo percorso, e l'obiettivo stesso di abbattere certe barriere, nasce per onorare una responsabilità che sento mia:

"Se uno può fare delle cose buone per gli altri, ha l'obbligo morale di farle tutte. Non è una scelta, ma una responsabilità."

Infine, voglio lasciare un promemoria al me stesso del futuro: qualsiasi cosa accadrà, ricorda che *"poteva andare peggio"*.

E un grazie, l'ultimo, lo devo **a me stesso**. Perché alla fine, **un senso l'abbiamo trovato**.

Bibliografia

[1] S. Sharma e S. Singh, «Vision-based sign language recognition system: A Comprehensive Review,» in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 140–144. DOI: 10.1109/ICICT48043.2020.9112409.

[2] A. Núñez-Marcos, O. Perez-de-Viñaspre e G. Labaka, «A survey on Sign Language machine translation,» *Expert Systems with Applications*, vol. 213, p. 118 993, 2023, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118993. indirizzo: https://www.sciencedirect.com/science/article/pii/S0957417422020115.

[3] P. Chua, C. M. Fang, T. Ohkawa, R. Kushalnagar, S. Nanayakkara e P. Maes, *EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language*, 2025. arXiv: 2505.17090 [cs.CV]. indirizzo: https://arxiv.org/abs/2505.17090.

[4] A. Mugele, L. De Greve, M. Van der Heyden e C. Schmit, «Nothing about us without us–co-designing a sign language translation application with the deaf community,» *Journal of Science Communication*, vol. 23, n. 01, A04, 2024. DOI: 10.22323/2.23010204.

[5] H. Park, Y. Lee e J. Ko, «Enabling Real-time Sign Language Translation on Mobile Platforms with On-board Depth Cameras,» *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, n. 2, giu. 2021. DOI: 10.1145/3463498. indirizzo: https://doi.org/10.1145/3463498.

[6] C. Lugaresi et al., *MediaPipe: A Framework for Building Perception Pipelines*, 2019. arXiv: 1906.08172 [cs.DC]. indirizzo: https://arxiv.org/abs/1906.08172.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei e Y. Sheikh, «OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, n. 1, pp. 172–186, 2021. DOI: 10.1109/TPAMI.2019.2929257.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić e C. Schmid, «ViViT: A Video Vision Transformer,» in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, ott. 2021, pp. 6836–6846.

[9] A. Vaswani et al., «Attention is All you Need,» in *Advances in Neural Information Processing Systems*, I. Guyon et al., cur., vol. 30, Curran Associates, Inc., 2017. indirizzo: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[10] A. Duarte et al., «How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language,» in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, giu. 2021, pp. 2735–2744.

[11] S. Yan, Y. Xiong e D. Lin, «Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,» *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, n. 1, apr. 2018. DOI: 10.1609/aaai.v32i1.12328. indirizzo: https://ojs.aaai.org/index.php/AAAI/article/view/12328.

[12] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney e R. Bowden, «Neural Sign Language Translation,» in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, giu. 2018.

[13] P. Rust, B. Shi, S. Wang, N. C. Camgöz e J. Maillard, *Towards Privacy-Aware Sign Language Translation at Scale*, 2024. arXiv: 2402.09611 [cs.CL]. indirizzo: https://arxiv.org/abs/2402.09611.

[14] C. Neidle, A. Opoku e D. Metaxas, *ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP)*, 2022. arXiv: 2201.07899 [cs.CL]. indirizzo: https://arxiv.org/abs/2201.07899.

[15] C. Hutto e E. Gilbert, «VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,» in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, vol. 8, 2014, pp. 216–225.

[16] J. Shen et al., «Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,» in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783. DOI: 10.1109/ICASSP.2018.8461368.

[17] Y. Ren et al., *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*, 2022. arXiv: 2006.04558 [eess.AS]. indirizzo: https://arxiv.org/abs/2006.04558.

[18] J. Kong, J. Kim e J. Bae, «HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,» in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan e H. Lin, cur., vol. 33, Curran Associates, Inc., 2020, pp. 17 022–17 033. indirizzo: https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.

[19] J. Kim, J. Kong e J. Son, «Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,» in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila e T. Zhang, cur., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 5530–5540. indirizzo: https://proceedings.mlr.press/v139/kim21f.html.

[20] K. Zhou, B. Sisman, R. Liu e H. Li, «Emotional voice conversion: Theory, databases and ESD,» *Speech Communication*, vol. 137, pp. 1–18, 2022, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2021.11.006. indirizzo: https://www.sciencedirect.com/science/article/pii/S0167639321001308.

[21] S. Cho e S.-Y. Lee, *Multi-speaker Emotional Text-to-speech Synthesizer*, 2021. arXiv: 2112.03557 [cs.CL]. indirizzo: https://arxiv.org/abs/2112.03557.

[22] Y. Wang et al., «Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis,» in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy e A. Krause, cur., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, ott. 2018, pp. 5180–5189. indirizzo: https://proceedings.mlr.press/v80/wang18h.html.

[23] C. Wang et al., *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*, 2023. arXiv: 2301.02111 [cs.CL]. indirizzo: https://arxiv.org/abs/2301.02111.

[24] Z. Borsos et al., *AudioLM: a Language Modeling Approach to Audio Generation*, 2023. arXiv: 2209.03143 [cs.SD]. indirizzo: https://arxiv.org/abs/2209.03143.

[25] S. AI, *Bark: A Transformer-based Text-to-Audio Model*, https://github.com/suno-ai/bark, 2023.

[26] A. Défossez, J. Copet, G. Synnaeve e Y. Adi, *High Fidelity Neural Audio Compression*, 2022. arXiv: 2210.13438 [eess.AS]. indirizzo: https://arxiv.org/abs/2210.13438.

[27] S. Wang e É. Székely, *Evaluating Text-to-Speech Synthesis from a Large Discrete Token-based Speech Language Model*, 2024. arXiv: 2405.09768 [eess.AS]. indirizzo: https://arxiv.org/abs/2405.09768.

[28] B. Saunders, N. C. Camgöz e R. Bowden, «Signing Avatars: A Review of the State of the Art in Sign Language Production,» *Computer Graphics Forum*, vol. 41, n. 2, pp. 615–640, 2022. DOI: 10.1111/cgf.14504.