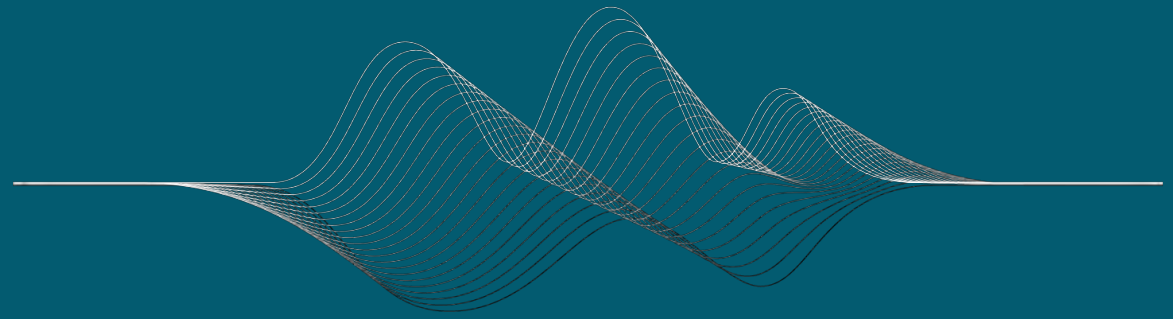


inspire

Inspiring Signal Intelligence with AI

inspire

Sintesi generativa di parlato patologico per la valutazione clinica automatica della disartria

Davide Lillini · Dottorato di ricerca, XXXIX ciclo

Tutor: Stefano Squartini · Co-tutor: Lucia Migliorelli, Leonardo Gabrielli

Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche

Le persone dietro la ricerca



Stefano Squartini

Professore ordinario

Università Politecnica delle Marche



Lucia Migliorelli

Tenure-Track Researcher

Università di Teramo



Leonardo Gabrielli

Professore associato

Università Politecnica delle Marche



Carlo Aironi

Postdoctoral researcher

Università Politecnica delle Marche

inspire

Cosa facciamo

Sistemi di supporto alle valutazioni cliniche tramite analisi di segnali audio

Segnali audio
voce e respiro

→

DSP e IA
elaborazione del segnale e apprendimento automatico

→

Supporto clinico
alla valutazione del paziente

inspire

I due temi di ricerca

Apnee notturne

Analisi acustica del respiro nel sonno

Disartria

Analisi del parlato patologico

inspire

Il segnale audio

Non invasivo

Un microfono registra il respiro o la voce di pazienti con la minima invasività

Facile da raccogliere

Anche fuori dall'ospedale, per molte ore

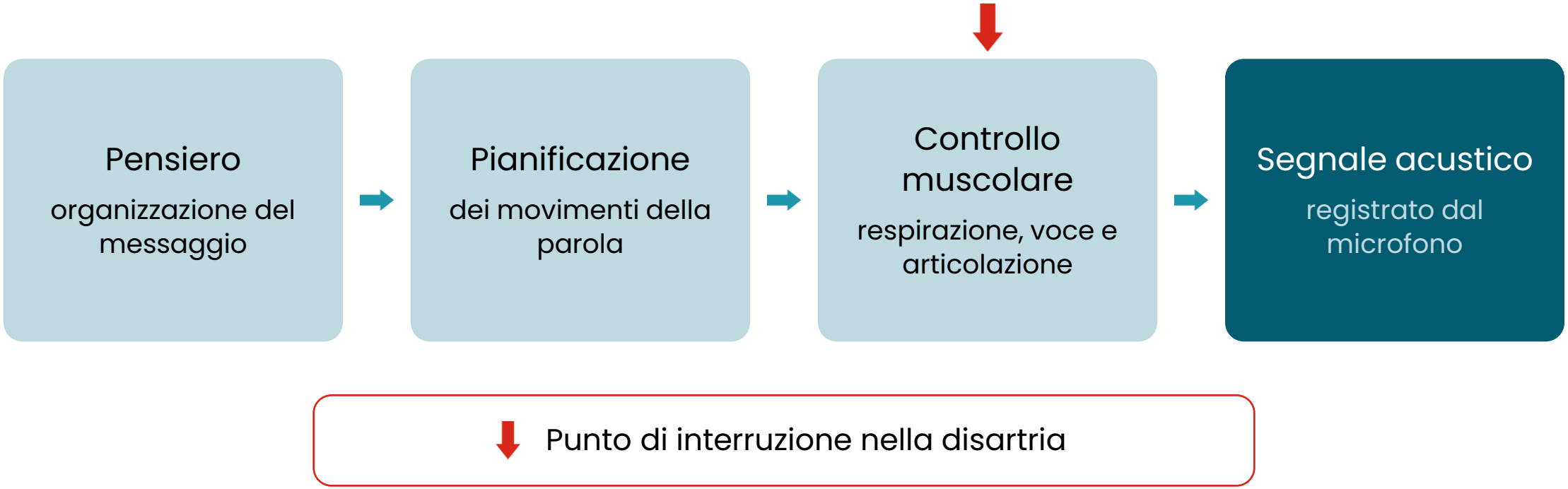
Ricco di informazioni

Le alterazioni del parlato possono essere indicatori precoci di malattie neurologiche

L'IA impara a riconoscere queste tracce.

inspire

Dal pensiero al suono



inspire

Disartria

Disturbo motorio della parola, di origine neurologica

Parlato lento, monotono, poco articolato

Compromette chiarezza e comprensibilità

FUNZIONI COINVOLTE

- Respirazione
- Fonazione
- Risonanza
- Articolazione
- Prosodia

inspire

Cause e Conseguenze cliniche

ALCUNE CAUSE

- SLA
- Malattia di Parkinson
- Paralisi cerebrale

CONSEGUENZE

- Riduce la comprensibilità e l'efficacia della comunicazione
- Incide sulle relazioni sociali e professionali
- Riduce l'autonomia e la qualità di vita del paziente
- **Può essere un indicatore precoce di malattia neurologica**

Lillini et al., MBEC (accettato)

inspire

Pratica clinica attuale

Il paziente parla
parole, frasi e compiti vocali

↓

Il logopedista ascolta
valutazione percettiva

↓

Scale standardizzate
RDP; FDA: 28 dimensioni in 8 categorie

↓

Punteggio
gravità e intelligibilità

BIOMARCATORI VALUTATI (RDP)

Respirazione	Fonazione
Muscoli oro-facciali	Diadococinesia
Riflessi	Articolazione
Prosodia	**Intelligibilità**

Biomarcatore chiave: l'intelligibilità.

Rudzicz et al., 2012 · Lillini et al., IJCNN 2025 · Lillini et al., Odyssey 2026

inspire

Intelligibilità: definizione e misura un esempio

DEFINIZIONE

$$SI = R / A$$

R: parole riconosciute dagli ascoltatori

A: parole totali presentate

CLASSI DI INTELLIGIBILITÀ

Molto bassa	0–25%
Bassa	26–50%
Media	51–75%
Alta	76–100%

Cinque ascoltatori trascrivono ogni parola: si media la percentuale corretta.

Feng e Chen, 2022 · Kim et al., 2008 · Woisard et al., 2020

inspire

Limiti della valutazione percettiva

SOGGETTIVA

Giudizio dell'esaminatore

Il clinico conosce il paziente: effetto di familiarità

Scarsa riproducibilità tra valutatori

POCO PRATICABILE

Lenta e costosa

Richiede ascoltatori e tempo per ogni paziente

Liste di parole apprese a memoria: intelligibilità sovrastimata

Serve una misura oggettiva e riproducibile.

Woisard et al., 2020 · Feng e Chen, 2022

inspire

Difficoltà nella raccolta dati

RACCOLTA

Affaticamento

I pazienti si stancano facilmente durante le registrazioni

Le sessioni cliniche sono brevi e il tempo è limitato

CONDIVISIONE

Dati sensibili

I dati clinici sono protetti da vincoli di privacy

Pochi corpora, spesso accessibili solo su richiesta

Pochi dati, poco vari e difficili da condividere.

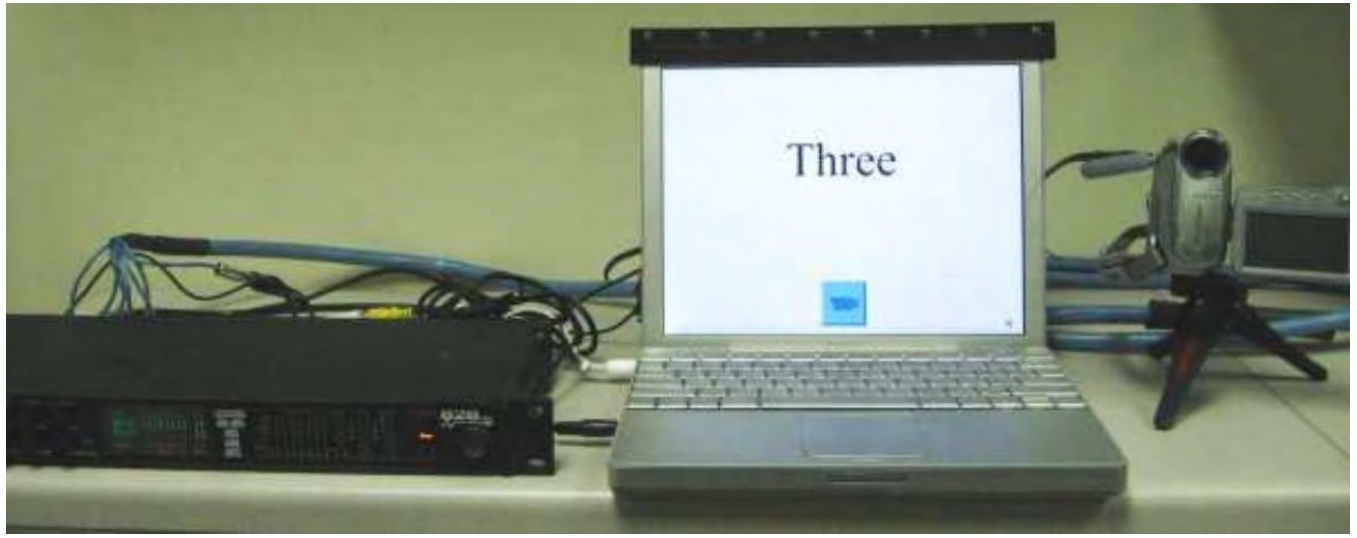
Marini et al., 2021 · Rudzicz et al., 2012 · Kim et al., 2008

inspire

Acquisizione dei dataset



Cabina anecoica, microfono a condensatore



Array di 8 microfoni, parola mostrata a schermo, videocamera

Woisard et al., 2020 · Kim et al., 2008 · Marini et al., 2021

inspire

Dataset disponibili

UA-Speech (inglese)	15 parlanti, paralisi cerebrale · 765 parole ciascuno
TORGO (inglese)	8 parlanti con Paralisi cerebrale o SLA · circa 15h di audio
HeyJay! (inglese)	36 parlanti, 6 eziologie neuromotorie · 8.669 enunciati
SAP (inglese)	524 parlanti · 5 eziologie · circa 415 ore di parlato
IDEA (italiano)	45 parlanti, 8 patologie · 16.794 registrazioni
EasyCall (italiano)	31 parlanti · 21.386 comandi
NeuroVoz (spagnolo)	54 parlanti con Parkinson · 2.977 registrazioni

I corpus accademici restano piccoli e poco vari.

Kim et al., 2008 · Rudzicz et al., 2012 · Marini et al., 2021 · Mendes-Laureano et al., 2024 · SAP Challenge 2025

inspire

Problemi aperti

Standardizzare la valutazione

dell'intelligibilità tramite algoritmi di IA

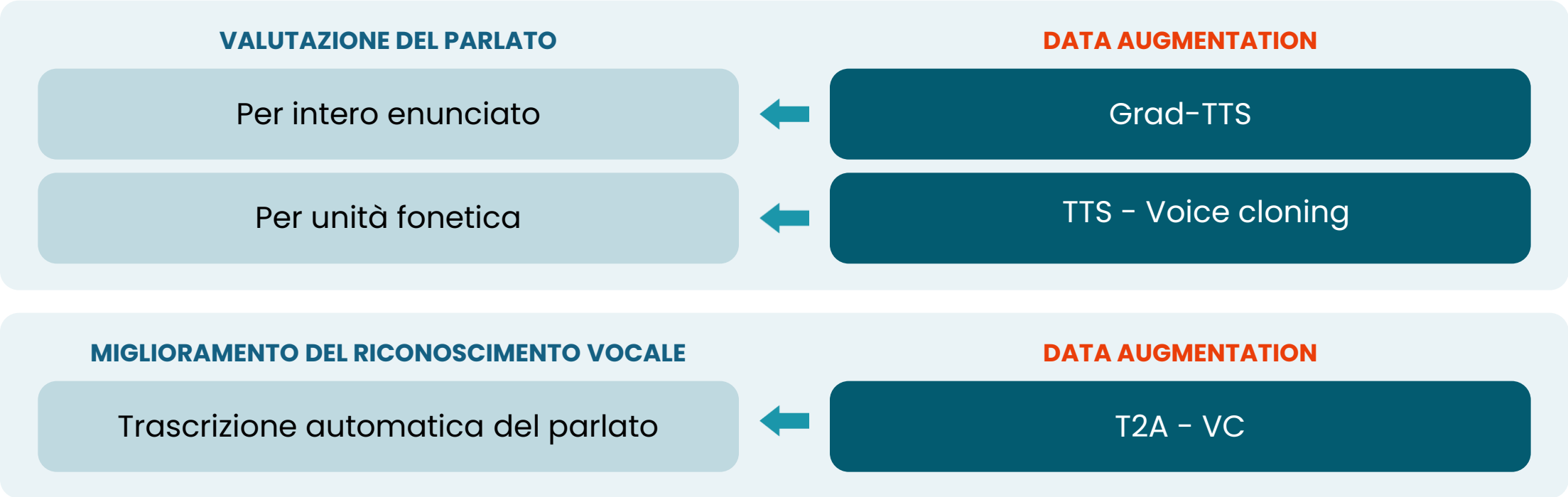
Generare parlato sintetico

patologico, per mitigare la scarsità di dati (data scarcity)

Lillini et al., IJCNN 2025 · Lillini et al., MBEC · Lillini et al., Odyssey 2026

inspire

Contributi sulla disartria



Lillini et al., IJCNN 2025 · MBEC · AII 2026

inspire

Sovrapposizione tra train e test

SPLIT PER ENUNCIATO

98,03–99,87%

accuratezza sugli stessi parlanti

PARLANTI SEPARATI

46,32–56,21%

accuratezza su parlanti mai visti

Stessi modelli: MobileNetV3, ResNet-50 e ResNet-152

Quasi perfetto solo su voci già viste: i modelli sfruttano i tratti del parlante, non marcatori robusti di disartria.

Per questo è importante partizionare i dati con pazienti disgiunti tra addestramento e test.

Lillini et al., IJCNN 2025

inspire

Grad-TTS: introduzione

PROBLEMA

Pochi dati di parlato disartrico etichettati: i classificatori di intelligibilità generalizzano male su parlanti mai visti.

APPROCCIO

Grad-TTS, modello a diffusione addestrato sulle voci disartriche di TORGO, sintetizza il vocabolario di UA-Speech.

CLASSIFICAZIONE

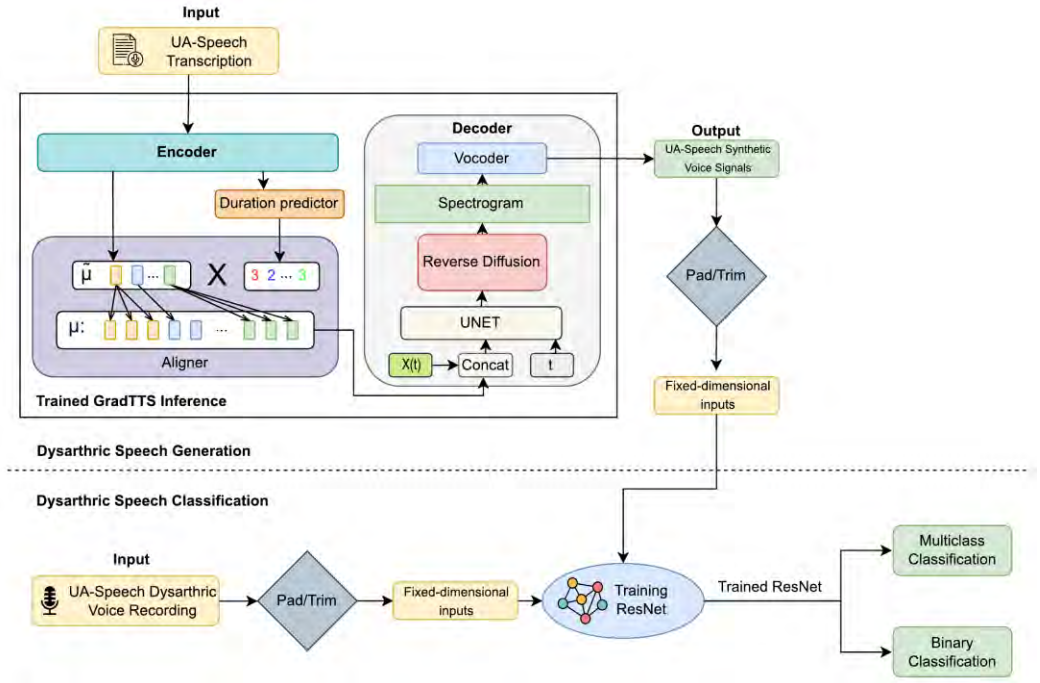
Si usa l'intero enunciato: viene trasformato in spettrogramma e usato come input per il modello di classificazione.

Il parlato sintetico migliora la classificazione dell'intelligibilità su parlanti mai visti?

Lillini et al., MBEC · Popov et al., 2021

inspire

Grad-TTS: pipeline



Voci TORGO, lessico UA-Speech
le etichette ereditano la scala FDA

Intero enunciato
trasformato in spettrogramma per la classificazione

Parlanti mai visti in test
3 fold, 4 parlanti di test per fold

Lillini et al., MBEC · Popov et al., 2021

inspire

Grad-TTS: risultati

INTELLIGIBILITÀ BINARIA

+24,3%

F1: 0,733 → 0,911

soli dati reali → con augmentation

INTELLIGIBILITÀ A 4 CLASSI

+48,0%

F1: 0,319 → 0,472

soli dati reali → con augmentation

UA-Speech · ResNet-50 · parlanti mai visti · media su 3 fold

Miglioramento relativo del F1 rispetto allo stesso modello con soli dati reali.

Lillini et al., MBEC

inspire

StyleTTS 2: introduzione

PROBLEMA

Quali unità del parlato segnalano meglio l'intelligibilità? Il giudizio sull'intero enunciato non lo dice.

APPROCCIO

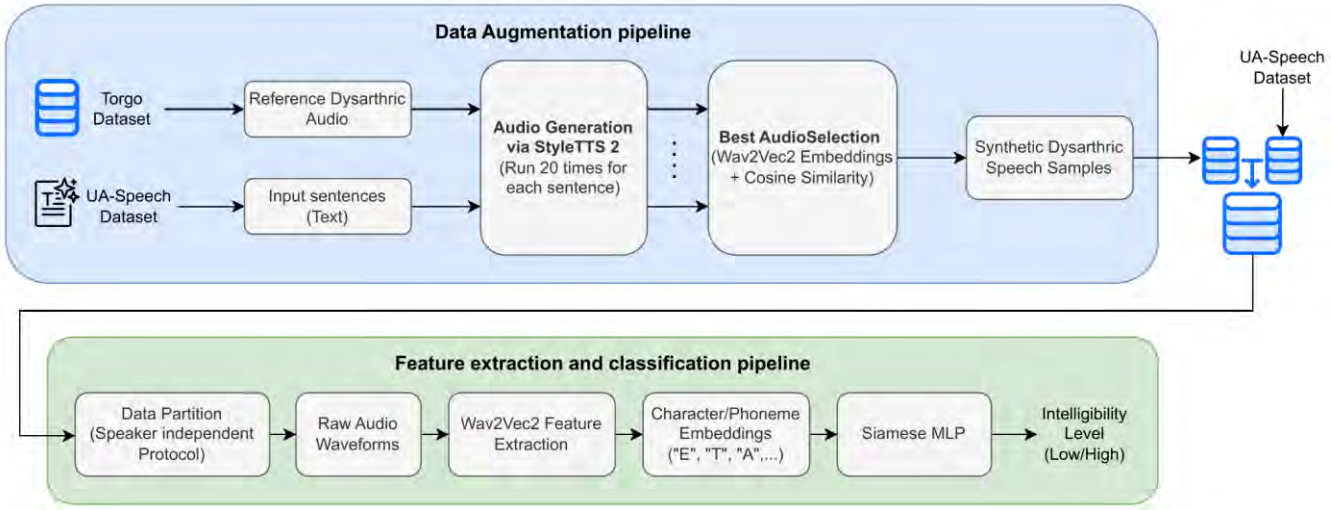
StyleTTS 2 con audio di riferimento TORGO genera frasi UA-Speech con voce patologica, selezionate per similarità coseno con Wav2Vec2.

ANALISI

18 unità (13 lettere e 5 fonemi) assegnate ai frame con CTC arg-max; un Classificatore per unità distingue bassa e alta intelligibilità.

Il parlato sintetico migliora la valutazione per unità fonetica?

StyleTTS 2: pipeline



Quali unità rivelano meglio la disartria?

StyleTTS 2

frasi sintetiche con voce di riferimento TORGO

13 lettere e 5 fonemi analizzati

assegnati ai frame con CTC arg-max

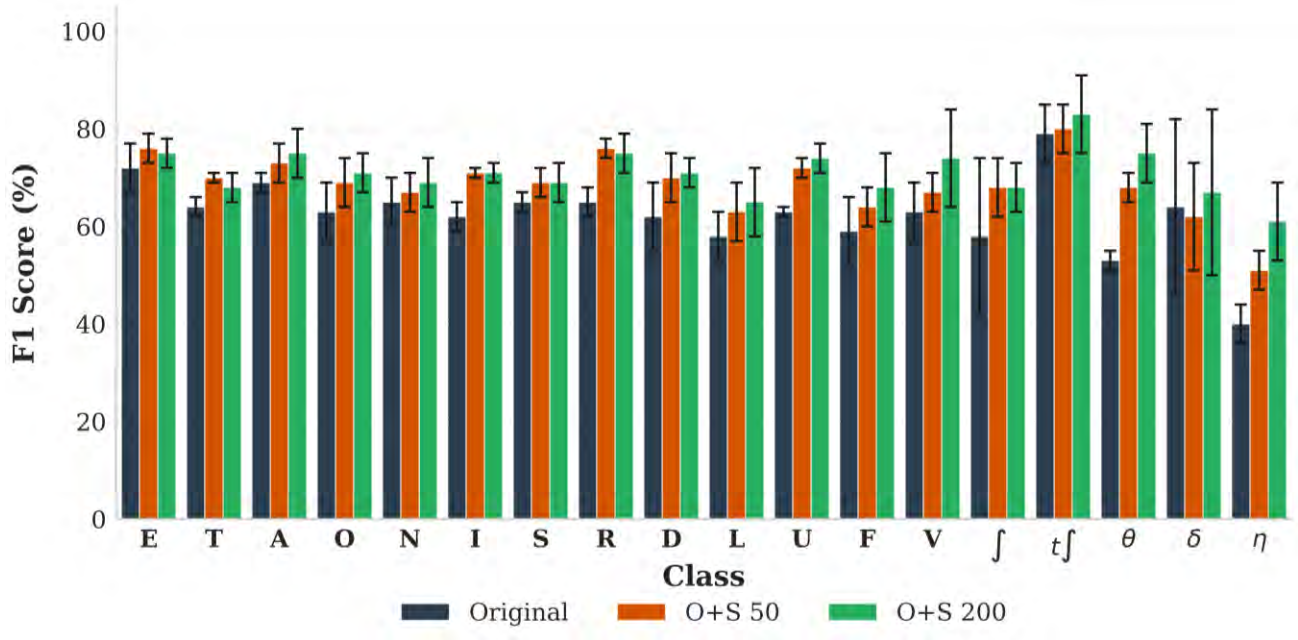
3 fold

parlanti separati; test solo su parlato reale

Lillini et al., AII 2026

inspire

StyleTTS 2: risultati



F1 per unità (soli originali, originali + sintetici); media su 3 fold

VARIAZIONE RELATIVA DEL F1

tʃ	79,3% → 82,7% **+4,3%**
E	71,7% → 75,3% **+5,0%**
R	65,3% → 75,0% **+14,9%**
A	69,0% → 74,7% **+8,3%**
θ	52,7% → 74,7% **+41,7%**

Media su 18 unità: 62,5% → 71,0% (+13,6%)

Lillini et al., AII 2026

inspire

T2A - VC: introduzione

PROBLEMA

I sistemi ASR sbagliano di più sul parlato disartrico, ma i dati per adattarli sono pochi, soprattutto fuori dall'inglese.

APPROCCIO

Una ricetta clinica (46 dimensioni percettive, eziologia, gravità) condiziona un modello a diffusione addestrato su SAP.

DATI SINTETICI

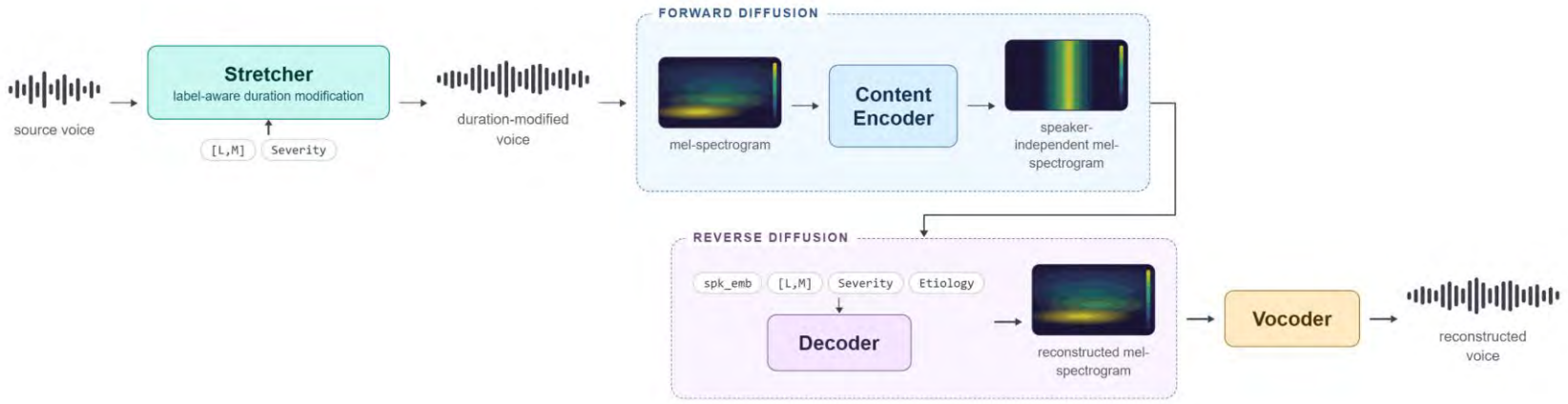
Pazienti sintetici in inglese, italiano e spagnolo per addestrare l'ASR; test su cinque corpora di parlato reale.

Il parlato sintetico multilingua riduce l'errore di riconoscimento sul parlato disartrico reale?

Lillini et al., Computer Speech & Language (in preparazione)

inspire

T2A - VC: pipeline



Set A: replica

100 profili di pazienti SAP (inglese)

Set B: nuovi pazienti

identità sana; inglese, italiano, spagnolo

Lillini et al., Computer Speech & Language (in preparazione)

inspire

T2A - VC: risultati

−21,5%	−17,3%	−10,6%	−6,3%	−37,0%
HeyJay (en)	**IDEA (it)**	**UA-Speech (en)**	**TORGO (en)**	**NeuroVoz (es)**

Variazione relativa del CER senza → con dati sintetici, migliore configurazione per corpus.

Meno errori su tutti e cinque i corpora reali.

Lillini et al., Computer Speech & Language (in preparazione) · dati preliminari

inspire

Riassunto dei risultati

CLASSIFICAZIONE DELL'INTELLIGIBILITÀ

Intelligibilità binaria Grad-TTS · ResNet-50	**0,733** F1 · SOLO REALI	**0,911** F1 · AUGMENTATION	**+24,3%** VARIAZIONE (+0,178)
Intelligibilità 4 classi Grad-TTS · ResNet-50	**0,319** F1 · SOLO REALI	**0,472** F1 · AUGMENTATION	**+48,0%** VARIAZIONE (+0,153)
Unità fonetiche StyleTTS 2 · 18 unità	**62,5%** F1 · SOLO ORIGINALI	**71,0%** F1 · +200 SINTETICHE	**+13,6%** VARIAZIONE (+8,5 pt)

RICONOSCIMENTO AUTOMATICO DEL PARLATO (ASR)

HeyJay (en) Parlato multilingua			**-21,5%** CER · VARIAZIONE
IDEA (it) Parlato multilingua			**-17,3%** CER · VARIAZIONE
NeuroVoz (es) Parlato multilingua			**-37,0%** CER · VARIAZIONE

Lillini et al., MBEC · AII 2026 · Computer Speech & Language (in preparazione)

inspire

Conclusioni e take-away

TAKE-AWAY

- Respiro e voce: il solo audio offre misure oggettive per la valutazione della disartria
- Tre approcci generativi (Grad-TTS, StyleTTS 2, multilingua) attenuano la scarsità di dati
- Migliorano classificazione (F1 da +13,6% a +48,0%) e riconoscimento (CER da -6,3% a -37,0%)

PROSPETTIVE

- Valutazione degli audio sintetizzati da logopedisti esperti
- Riduzione degli errori di classificazione tra classi di intelligibilità adiacenti
- Verifica sperimentale degli algoritmi nella pratica clinica reale

L'IA non sostituisce il clinico: gli offre misure oggettive e strumenti di supporto.

inspire

Grazie per l’attenzione

Domande e discussione benvenute

Davide Lillini · d.lillini@pm.univpm.it

Dipartimento di Ingegneria dell’Informazione, Università Politecnica delle Marche

