Classificazione automatica dell'eloquio disartrico: un'analisi comparativa dei metodi di supporto decisionale

D. Lillini, C. Aironi, L. Migliorelli, L. Gabrielli, S. Squartini
Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche
Via Brecce Bianche 12, 60131, Ancona, Italia, e-mail: s.squartini@univpm.it

La disartria compromette la chiarezza dell'eloquio e può rappresentare un indicatore precoce di patologie neurologiche, quali la sclerosi laterale amiotrofica (SLA) ad esordio bulbare e la paralisi cerebrale. Il monitoraggio dell'evoluzione della disartria riveste un ruolo fondamentale nella predizione dell'avanzamento della malattia, nella prescrizione di ausili comunicativi e nello sviluppo di nuove metriche di outcome clinico. Tuttavia, la valutazione dell'intelligibilità del parlato si basa ancora prevalentemente su strumenti percettivi, come il Robertson Dysarthria Profile, somministrati ai pazienti dagli Speech Language Pathologists (SLPs). Tali valutazioni sono qualitative e soggette a variabilità inter- e intra-valutatore, limitandone l'affidabilità. Per superare questi limiti, la letteratura recente ha esplorato metodi automatici per la stima dell'intelligibilità, distinguibili in approcci data-driven, che si avvalgono di tecniche di apprendimento automatico, e approcci non data-driven, basati su modelli ASR pre-addestrati o algoritmi di digital signal processing (DSP). In [1] gli autori utilizzano Convulutional Neural Network (CNN), e reti Long Short-Term Memory (LSTM) per la classificazione automatica del livello di disartria, mentre Tripathi et al. in [2] propone un approccio basato su confronto tra trascrizioni ASR e ground-truth tramite metriche di similarità come Levenshtein Distance e Sequence Matching.

Il dataset impiegato per l'analisi è l'UA-Speech [3], composto da registrazioni di 15 soggetti affetti da paralisi cerebrale con differenti livelli disartria, suddivisi in quattro classi: very low, low, mid e high. Dopo una fase di pre-processing dei dati, sono stati implementati e confrontati tre modelli data-driven, ovvero MobileNetV3, ResNet50 e ResNet152, utilizzando spettrogrammi log-Mel come input, accompagnati da due approcci non data-driven: il primo fondato sull'impiego di sistemi ASR pre-addestrati (Whisper e Wav2Vec), il secondo basato su algoritmi di elaborazione del segnale (DSP) finalizzati al calcolo delle metriche STOI ed ESTOI, attraverso il confronto tra segnali vocali sani e disartrici.

Per garantire una valutazione robusta, sono state adottate quattro strategie di divisione dei dati: il Preset A, che prevede la separazione dei soggetti tra training e test set; il Preset B, che utilizza una divisione casuale per utterance; il Preset C, che impiega un set di training ridotto ed il preset D dove tutti i dati sono stati inseriti nel test set per la valutazione dei soli approcci non-data-driven. I modelli sono stati valutati in termini di accuratezza e coefficiente di correlazione di Pearson (PR-C), rispetto agli score di intelligibilità percettiva forniti dagli SLPs insieme al dataset.

I modelli data-driven hanno ottenuto accuratezze superiori al 99% nei preset B e C, risultato attribuibile all'overlap dei pazienti tra training e test set. Tuttavia, imponendo la separazione dei soggetti come previsto dal preset A, l'accuratezza è diminuita drasticamente fino a valori compresi tra il 46% e il 56%, evidenziando problemi di generalizzazione [4]. Al contrario, l'approccio ASR con il modello Whisper ha raggiunto un coefficiente di correlazione PR-C pari a 0.93 e un'accuratezza dell'80%, dimostrandosi più robusto anche in scenari realistici. L'approccio DSP, basato su misure STOI, pur essendo meno performante (PR-C = 0.80; accuratezza = 60%), rappresenta una valida alternativa in contesti con risorse limitate.

Di seguito vengono riportati in tabella 1 e 2 i risultati relativi agli algoritmi data-driven e non-data-driven relativi ai preset utilizzati.

Tabella 1 - Confronto tra modelli non-data-driven per prest

Model	Preset	Model size	PR-C	Test acc [%]
Whisper	D	Base	**0.93**	**80.0**
		Large	**0.93**	**80.0**
Wav2Vec	D	Base	0.87	73.3
		Large	0.89	73.3
DSP approach	D	–	0.80	60.0

Tabella 2 - Confronto tra modelli data-driven per preset

Model	Preset	Learning rate	Epoche	Test acc [%]
MobileNetV3	A	0.01	100	48.09
ResNet152	A	0.01	100	**56.21**
ResNet50	A	0.01	100	46.32
MobileNetV3	B	0.01	100	99.40
ResNet152	B	0.01	100	**99.87**
ResNet50	B	0.01	100	99.75
MobileNetV3	C	0.01	100	**99.20**
ResNet152	C	0.01	100	98.03
ResNet50	C	0.01	100	98.23

L'analisi condotta mette in evidenza come la presenza di campioni dello stesso paziente nei set di training e test influisca pesantemente sulle prestazioni apparenti dei modelli data-driven, inducendo un overfitting non rappresentativo delle condizioni cliniche reali. Gli approcci non-data-driven, e in particolare i sistemi ASR pre-addestrati, mostrano una maggiore capacità di generalizzazione, rendendoli più adatti per applicazioni cliniche pratiche. Gli sviluppi futuri si concentreranno sull'esplorazione di approcci ibridi che integrino i punti di forza di entrambe le metodologie, affrontando al contempo le limitazioni dei dataset, come la mancanza di riferimenti a parlanti sani, al fine di migliorare ulteriormente l'affidabilità e l'utilità di questi sistemi.

Bibliografia

[1] Joshy, Amlu Anna, and Rajeev Rajan. "Automated dysarthria severity classification using deep learning frameworks." *2020 28th European signal processing conference (EUSIPCO)*. IEEE, 2021.

[2] Tripathi, Ayush, Swapnil Bhosale, and Sunil Kumar Kopparapu. "A novel approach for intelligibility assessment in dysarthric subjects." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[3] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. R. Gunderson, T. S. Huang, K. L. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in Interspeech, vol. 2008, 2008, pp. 1741 1744.

[4] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches," in ICASSP 2023-2023 IEEE