Abstract

Strategie basate su Deep Learning
per sistemi di Riconoscimento Vocale di parlato disartrico

di Rosanna Turrisi

L'obiettivo di questa tesi è quello di sviluppare tecniche di *deep learning* per sistemi di riconoscimento vocale (RV) per persone affette da disartria. La disartria è un disordine motorio che comporta una compromissione della comunicazione verbale e difficoltà nel controllo motorio. Perciò, dispositivi basati sul riconoscimento vocale potrebbero rappresentare l'unica possibilità per soggetti disartrici di interagire con il mondo esterno. Tuttavia, le tecnologie oggi disponibili risultano inadeguate. Ad esempio, testando Google Speech API e IBM sul dataset TORGO, abbiamo ottenuto un Word Error Rate (WER) superiore all'80%. Poiché le caratteristiche del parlato dipendono dalla gravità e dal tipo di disartria, un modello di RV ideale dovrebbe essere addestrato su grandi dataset di parlato disartrico. Sfortunatamente, questi dati sono difficili da acquisire e sono oggi disponibili solo pochi e limitati dataset. Al fine di risolvere questo problema, presentiamo qui tre possibili strategie.

La prima strategia si basa sull'integrazione di informazioni sulla dinamica del tratto vocale nei sistemi di RV. Poiché la misurazione articolatoria è piuttosto invasiva e costosa, ci siamo prima soffermati su metodi per la sintesi di *feature* articolatorie (FA), a partire da *feature* fonetiche. Più precisamente, proponiamo l'uso delle *feature* fonetiche in aggiunta o in sostituzione a quelle acustiche nell'Acoustic Inversion (AI) map, con lo scopo di migliorare la generalizzazione a nuovi dataset. Successivamente, abbiamo introdotto metodi non supervisionati che combinano *feature* fonetiche, contenenti informazioni articolatorie grezze, e *feature* acustiche, contenenti informazioni sulla co-articolazione, per sintetizzare FA.
Infine, abbiamo esplorato due strategie per integrare le FA sintetizzate nei sistemi di RV. Dopo aver ottenuto incoraggianti risultati sul dataset TIMIT sulla classificazione di fonemi, abbiamo testato le performance del sistema di RV sul dataset CHiME-4. Le due strategie hanno portato a una riduzione relativa del WER di 1.9% e 5.4%, rispettivamente.

La seconda strategia si basa sul *multi-source domain adaptation* (MSDA), in cui vengono sfruttati dataset sorgente per apprendere un classificatore per un dataset target. L'algoritmo proposto, MSDA Weighted Joint Distribution Optimal Transport (MSDA-WJDOT), è ottimizzato per trovare il miglior allineamento, basato sul Trasporto Ottimo, tra la distribuzione di probabilità del target

e una combinazione convessa di quelle sorgente. Tale combinazione è pesata da un coefficiente che viene appreso in base alla distanza tra le distribuzioni sorgente e target. Abbiamo utilizzato poi questo algoritmo in due applicazioni. Nel primo caso, abbiamo adottato MSDA-WJDOT per imparare una funzione in grado di diagnosticare la disartria. In questo caso, sia i dataset sorgente che quello target sono dataset di parlato disartrico e normale, contenente ciascuno un diverso tipo di rumore. La seconda applicazione riguarda invece l'adattamento di un sistema di riconoscimento di comandi vocali a uno speaker disartrico. In entrambi i casi, MSDA-WJDOT ha ottenuto performance migliori sia della baseline che di altri metodi per il MSDA.

In ultimo, ci siamo focalizzati su contesti in cui è sufficiente il riconoscimento di un vocabolario limitato. Questo permette di semplificare il problema riducendolo al riconoscimento di comandi vocali. Lavorando in questa direzione, abbiamo acquisito un dataset di parlato disartrico contenente comandi relativi all'uso di un'applicazione Rubrica per smartphone. Queste registrazioni costituiscono il più esteso dataset di parlato disartrico in italiano. Infine, introduciamo il progetto AllSpeak in cui è stata sviluppata un'applicazione Android basata sul riconoscimento di comandi vocali. Questa permette a soggetti affetti da Sclerosi Laterale Amiotrofica (SLA) di comunicare i loro bisogni primari (e.g., "Ho sete") anche quando il loro parlato è a stento intelligibile.