**SLIDE 1**

Partendo dalla definizione del problema, Il problema dell’audio super resolution consiste nell’aumentare la frequenza di campionamento di un segnale, ovvero stimare un segnale in alta risoluzione partendo da una sequenza in bassa risoluzione contenente solo una piccola frazione dei campioni originali. Questo concetto lo possiamo anche visualizzare con lo spettrogramma riportato nella slide, in cui vediamo come l’obiettivo sia proprio quello di recuperare le alte frequenze del segnale originale partendo dalle basse frequenze. In questo lavoro la frequenza di campionamento target è di 16 kHz, mentre quella dei segnali in input è di 4 kHz quindi abbiamo un rapporto pari a 4.

Formalmente, questo si traduce in un problema supervisionato, quindi un dataset composto da coppie di segnali a bassa e alta risoluzione, in cui cerchiamo di ricostruire il segnale con un approccio regressivo, utilizzando un modello di DL. Abbiamo dunque un problema dei minimi quadrati, e una naturale conseguenza di questa formulazione è l’utilizzo dell’errore quadratico medio come funzione di loss.

**SLIDE 2**

Questo lavoro si basa principalmente sullo studio di due articoli, che hanno portato importarti contributi in letteratura. Nel primo articolo gli autori implementano un modello denominato time-frequency network (TFNet) che consiste in una rete neurale con una architettura ramificata. Sostanzialmente in questo modello ci sono 2 rami per processare il segnale: uno per processarlo nel dominio del tempo ed uno per processarlo nel dominio delle frequenze. Gli autori di questo paper hanno dimostrato che questo approccio migliora le performance rispetto ad una più tradizionale stima che elabora il segnale solo nel dominio del tempo. Questa rete si occupa di ottimizzare i due rami congiuntamente al fine di fornire un'unica previsione.

Il secondo paper sviluppa un modello denominato TFiLM Net. Il maggior contributo di questo modello è lo sviluppo di layer che sfruttano le reti ricorrenti per modificare le attivazioni dei layer convoluzionali. Questo viene fatto perché i layer convoluzionali, che sono ampiamente utilizzati in letteratura per processare segnali digitali, presentano un limite per processare i dati sequenziali, come per l’appunto i segnali audio. Infatti le reti convoluzionali hanno un campo recettivo limitato, che comporta un’insufficiente capacità di catturare dipendenze a lungo raggio tra i dati. Per risolvere questo problema TFiLM Net utilizza delle reti ricorrenti per modificare le attivazioni dei layer convoluzionali. E questo permette di espandere il campo ricettivo delle reti convoluzionali.

Quindi visto che il modello TFNet utilizza layer convoluzionali per processare i dati, l’idea di questo lavoro è quella di combinare questi due metodi per proporre una nuova architettura, erenditando le migliori rispettive caratteristiche dei due paper.

**SLIDE 3**

In questo schema possiamo appunto visualizzare i principali componenti che caratterizzano l’approccio proposto.

Andando con ordine, la prima operazione che viene effettuata in questo sistema è quella di upsampling, ovvero il segnale a bassa risoluzoine viene interpolato con le spline in modo tale che input e output abbiano da subito la stessa dimensione.

Dopodichè, possiamo osservare la presenza di una architettura ramificata: come nel modello TFNet c’è un ramo che permette di processare il segnale nel dominio del tempo (che è quello in basso) e un ramo un po’ più complesso, che permette di processare il segnale nel dominio delle frequenze.

Partendo dal dominio del tempo, vediamo che il processing del sistema avviene attraverso la rete neurale citata in precedenza, ovvero TFiLM Net. Guardandola più nel dettaglio, essa si compone di un’architettura a collo di bottiglia, ispirata agli autoencoder con l’ausilio di connessioni residue. Questo viene fatto per facilitare un apprendimento gerarchico delle feature in fase di training. Il cuore di questa rete neurale è composto dai layer TFiLM che sono rappresentati a destra. L’input di questi layer è sostanzilamente un tensore ottenuto come output di un layer convoluzionale, con funzione di attivazione reLU. L’idea è quella di pesare le arrivazioni dell’input in base all’output ottenuto da una rete ricorrente. In particolare, il tensore di input viene suddiviso in diversi blocchi, ciascuno dei quali viene ridotto di dimensionalità con un’operazione di pooling, e successivamente passa in input ad una rete LSTM unidirezionale al contrario di quello che è scritto nello schema. Infine, l’output della LSTM viene utilizzato per pesare i valori in input così che il layer restituisca un tensore di attivazioni pesate che espandono il campo ricettivo di una semplice convoluzione.

Ritornando allo schema di partenza, andiamo adesso a vedere il processing nel ramo delle frequenze. Nel ramo in alto il segnale viene per prima cosa passato nel dominio delle frequenze attraverso le trasformate di Fourier. L’obiettivo di questo ramo è stimare esclusivamente lo spettro di ampiezza del segnale in alta risoluzione, escludendo quindi la stima della fase, che è riservata al ramo temporale. Come abbiamo detto in precedenza l’ipotesi alla base di un modello di super resolution è che le alte frequenze di un segnale possano essere stimate partendo dalle basse frequenze. Per esplicitare questa dipendenza viene utilizzato appunto lo Spectral Replicator layer. Sostanzialmente questa componente del sistema non fa altro che replicare le basse frequenze dello spettro di ampiezza in input nelle corrispettive alte frequenze. In questo modo viene esplicitata la dipendenza tra basse e alte frequenze e la rete è facilitata nella stima.

A questo punto vediamo che l’output dello spectral replicator layer viene processato dalla rete TFiLM Net che abbiamo presentato in precedenza, e in questo modo abbiamo due stime che provengono dai due rami, denominate qui come m e z. Queste due stime vengono combinate grazie allo spectral fusion layer. Sostanzialmente si compone di 2 equazioni. La prima permette di ottenere la stima finale dello spettro di ampiezza. Essa come vediamo consiste nella stima pesata delle rispettive stime dei due rami. Il peso di ciascun ramo in questa stima è determinato con un approccio data driven. Infatti il peso w che vediamo nell’equazione è una variabile stimata dalla rete, la cui distribuzione è dunque determinata nella fase di training. Una volta ottenuto il valore di M, si passa alla seconda equazione in cui si ottiene l’effettivo segnale stimato in alta risoluzione. Come possiamo vedere, mentre lo spettro di ampiezza è stimato da entrambi i rami della rete, la fase è stimata solo nel dominio del tempo.

**SLIDE 4**

Venendo agli ambienti di valutazione, visto che per valutare il modello proposto lo confronteremo nella sezione dei risultati con i due paper di riferimento, è importante sottolineare quali sono i dati utilizzati e l’ambiente di valutazione dei paper originali. Sia in TFiLM che in TFNet il dataset utilizzato è il VTCK, che possiamo considerare un benchmark per il problema di audio sr. Il dominio di riferimento per questi dati è quello delle voci, infatti questo dataset contiene 44 h di registrazioni in inglese effettuate da 109 speaker. Perciò la dimensione dei dati è piuttosto elevata. Anche i modelli sono piuttosto grandi, contano infatti all’incirca 35 M di parametri e l’insieme delle risorse computazionali richieste supera i limiti di calcolo a disposizione per questo progetto. Di conseguenza sono state adottate due strategie: un campionamento stratificato per genere dei dati ed una riduzione dei modelli fino a 22 M di parametri. Perciò sono stati riallenati in queste condizioni sia i modelli di riferimento, TFiLM e TFNet, sia il modello proposto.

**SLIDE 5**

Per quanto riguarda la preparazione dei dati, brevemente, sono stati effettuati i seguenti step. Per prima cosa sono stati convertiti i dati originali attraverso un downsampling così da ottenere coppie di registrazioni ad alta e bassa risoluzione, necessarie per l’allenamento del modello. Successivamente sono state filtrate le sequenze di dati con una bassa energia, ovvero le sequenze che potremmo definire di “silenzio” all’interno di una clip audio. Questa tecnica si è dimostrata in letteratura essere efficace per stabilizzare il gradiente e favorire la convergenza del modello. Dopodichè, sono state estratte dalle registraizoni originali patch di lunghezza pari a 8192 campioni. Questa è appunto la dimensione dell’input, che viene ottenuta attraverso un upsampling con interpolazione spline, come avevamo già visto in precedenza.

**SLIDE 6**

Venendo alle metriche utilizzate per la valutazione del modello, abbiamo in primo luogo il signal to noise ratio, che mette in relazione la potenza del segnale originale del rumore, che potremmo definire banalmente come la differenza tra la ricostruzione del modello e la ground truth.

Mentre l’SNR è un segnale che misura la qualità della ricostruzione nel dominio del tempo, abbiamo la LSD che la misura nel dominio delle frequenze. Queste due metriche sono piuttosto standard nei problemi di audio SR.

E’ stato inoltre ideato un approccio meno tradizionale per la valutazione dei modelli che coinvolge un sistema di speech to text. Sostanzialmente mi sono chiesto se un modello di SR può essere utile per aiutare un sistema di STT nel trascrivere una registrazione pre-processando l’input. Per far ciò è stato utilizzato un modello open source di STT sviluppato da Baidu che è Deep Speech. E per valutare la qualità di ogni singola trascrizione è stata utilizzata una metrica standard per i problemi di STT che è il WER.

**SLIDE 7**

I 3 modelli sono dunque stati valutati con queste 3 metriche ciascuna delle quali è stata misurata in 2 diversi regimi. Ogni modello è infatti stato allenato sia su un task di Multi-Speaker, in cui appunto il training e il test set sono composti da diversi speaker e quindi la difficoltà è maggiore in quanto ci sono diverse distribuzioni di voci, di genere e di accenti. Una volta allenati i modelli sul multi-speaker, è stata effettuato un fine-tuning e ciascun modello è stato allenato nel task di single speaker, in cui appunto viene preso in considerazione un singolo speaker del dataset.

Come possiamo vedere già dal SNR i 3 modelli risultano decisamente migliori dell’interpolazione spline. In entrmbi i task il modello TFNet registra risultati leggermente migliori sia rispetto a TFiLM che al modello proposto.

Venendo alle altre metriche, abbiamo che per la LSD, il modello proposto registra il risultato migliore. Perciò la valutazione della ricostruzione nel dominio delle frequenze risulta essere maggiormente accurata nell’approccio proposto rispetto sia ai paper di riferimento sia all’interpolazione effettuata con le cubic spline in cui la differenza supera abbondantemente i 2 dB sia in single che multi-speaker.

Infine abbiamo i risultati del WER. Con le linee tratteggiate sono state evidenziate le metriche di errore registrate dal modello di DeepSpeech nei dati ad alta risoluzione, quindi sulla ground truth. Come possiamo vedere il modello di STT è molto accurato nella trascrizione e i risultati sembrano confermare la metrica precedente. Il modello proposto infatti risulta essere il migliore nel pre-processare l’input di Deep Speech superando TFNet e TFiLM net all’incirca di 0.5 punti percentuali.

**SLIDE 8**

Per concludere ho preparato una breve demo in cui possiamo ascoltare l’effetto del modello su una registrazione appartenente al test set del task multi-speaker. La prima che ascolteremo è la versione originale…. Questa è la versione in bassa risoluzione… Questa è la ricostruzione del modello.

Come possiamo vedere dagli spettrogrammi in basso, il modello riesce con successo a stimare correttamnete il contenuto delle alte frequenze del segnale originale soprattutto nel range di frequenze compreso tra 2 e 4 kHz.

Questo è confermato anche dai risultati nella trascrizione del modello di STT.

In alto infatti possimao vedere la trascrizione reale del file audio, e al centro possiamo vedere le varie trascrizioni del modello di Deep Speech nelle 3 diverse versioni della registrazione. Come possiamo vedere il modello quando il segnale è in bassa risoluzione non riesce praticamente a trascrivere neanche una parola, mentre con il pre-processing del modello di SR è in grado di ricostruire quasi tutta la frase correttamente.

**CONCLUSIONI**

Per concludere, il modello ha registrato risultati promettenti.

Da un lato, sicuramente si potrebbe pensare di inserire anche una metrica di valutazione qualitativa, che quindi si basi sulla percezione soggettiva di un campione di ascoltatori.

Infine, sarebbe interessante valutare il modello proposto senza le eccessive riduzioni di dimensinoalità dovute ai limiti computazionali a disposizione.