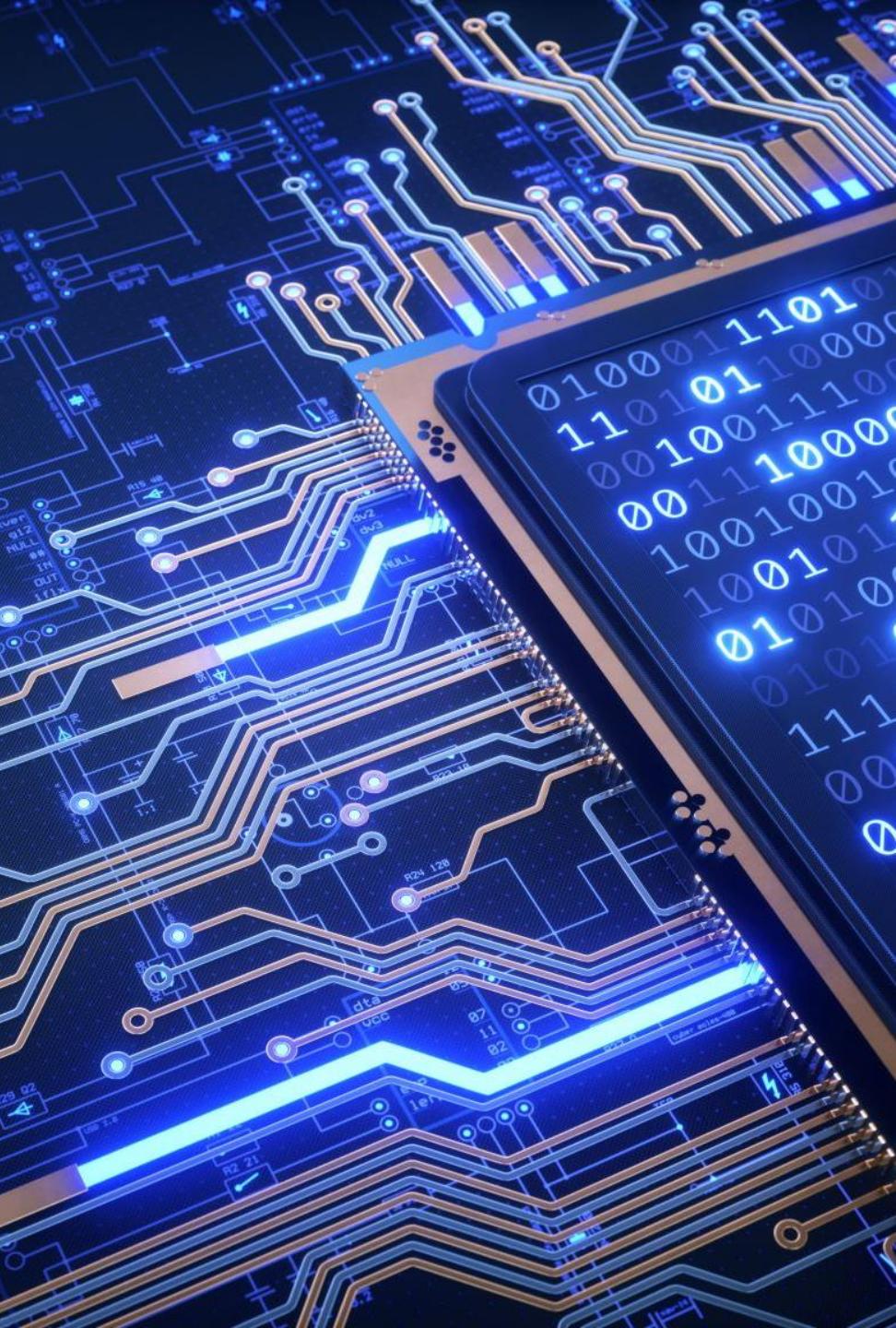


# ANALISI DELLE PRESTAZIONI DELLA RETE IBRIDA PARCNET PER IL PACKET LOSS CONCEALMENT SU SEGNALI VOCALI

ELABORATO SVOLTO DA MICHELE MARMORÈ



# OBIETTIVO

1. Studio della rete PARCnet;
2. Adattamento del dataset;
3. Training della rete con segnali speech;
4. Test di ricostruzione;
5. Valutazione dei risultati ottenuti.

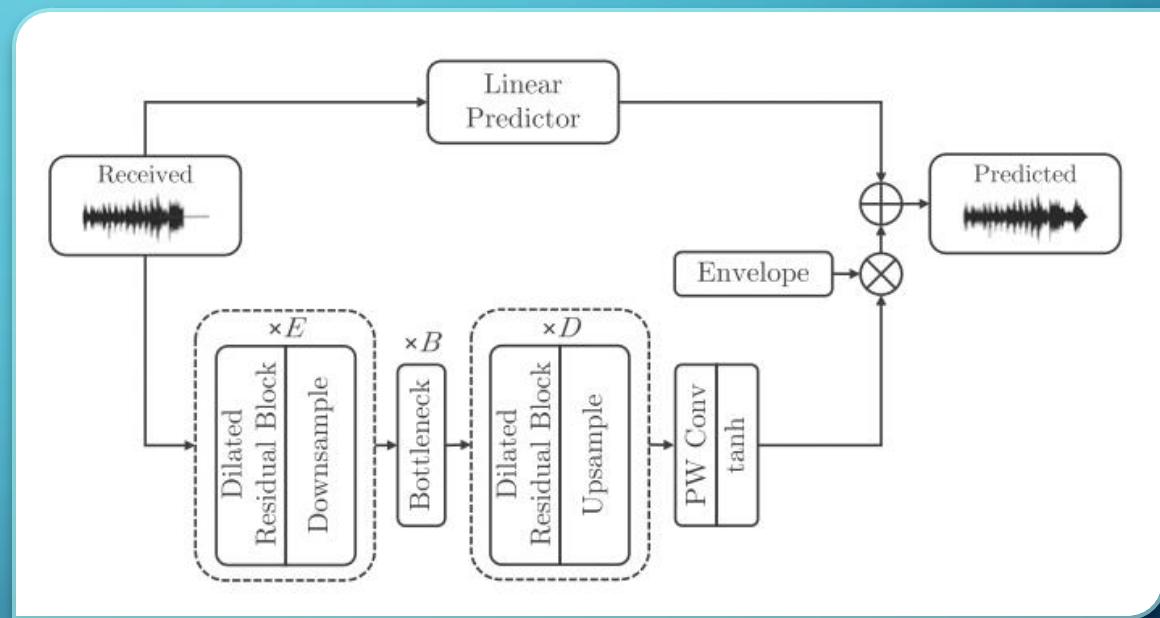
# PROBLEMA

Le comunicazioni in tempo reale su IP sono essenziali nell'uso quotidiano sia professionale che non. Un segnale per la comunicazione in tempo reale è trasmesso con reti di pacchetti. Tuttavia, queste si basano su protocolli "RTP/UDP", concepiti per garantire una massima velocità di trasmissione a discapito di una minore qualità. Si presenta quindi il problema della perdita dei pacchetti trasmessi. Questo rende necessario l'utilizzo di algoritmi in grado di compensare la perdita di frame al ricevitore e ripristinare una buona qualità di ascolto. In questo elaborato vengono valutate le prestazioni della rete PARCnet per il Packet Loss Concealment

# PARCnet

PARCnet è un metodo PLC ibrido che utilizza una rete neurale feed-forward per stimare il segnale residuo nel dominio del tempo di un modello autoregressivo lineare parallelo. La rete è formata da due rami paralleli:

1. Linear Predictor;
2. ramo a blocchi di convoluzione



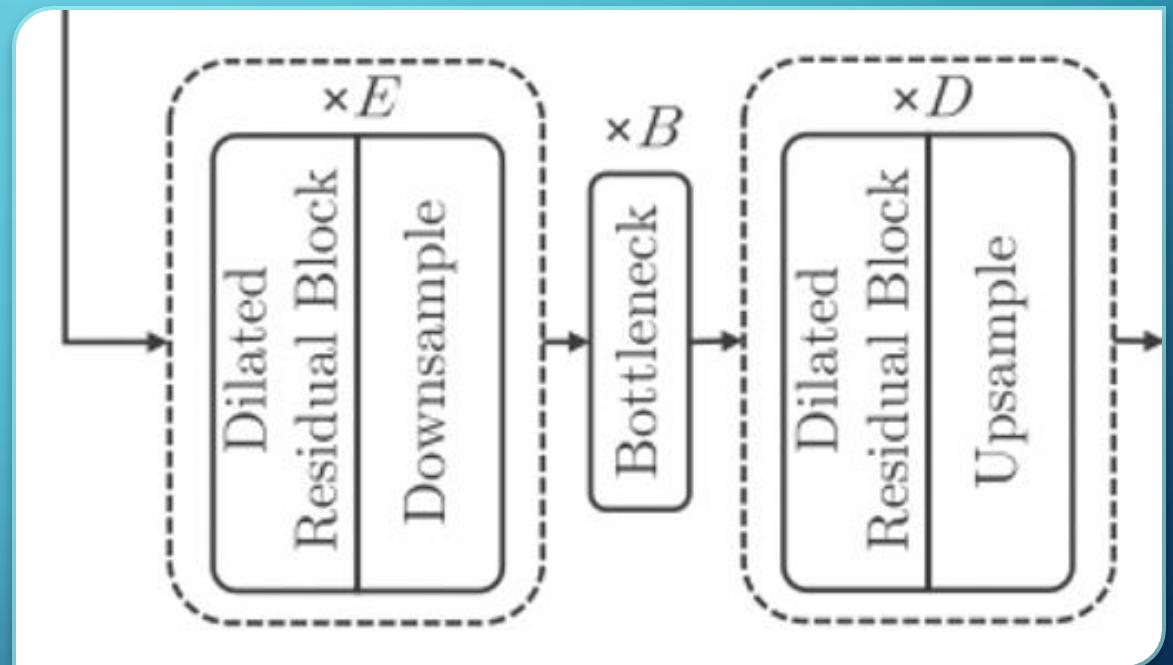
# LINEAR PREDICTOR

Il Linear Predictor è un modello che predice un valore futuro  $\hat{x}[n]$  in una sequenza di dati basandosi su una combinazione lineare di "k" valori precedenti. Questo coglie bene la dipendenza temporale del segnale a breve termine e permette quindi una previsione veloce ed accurata.

$$\hat{x}[n] = \sum_{i=1}^k a_i \cdot x[n - i]$$

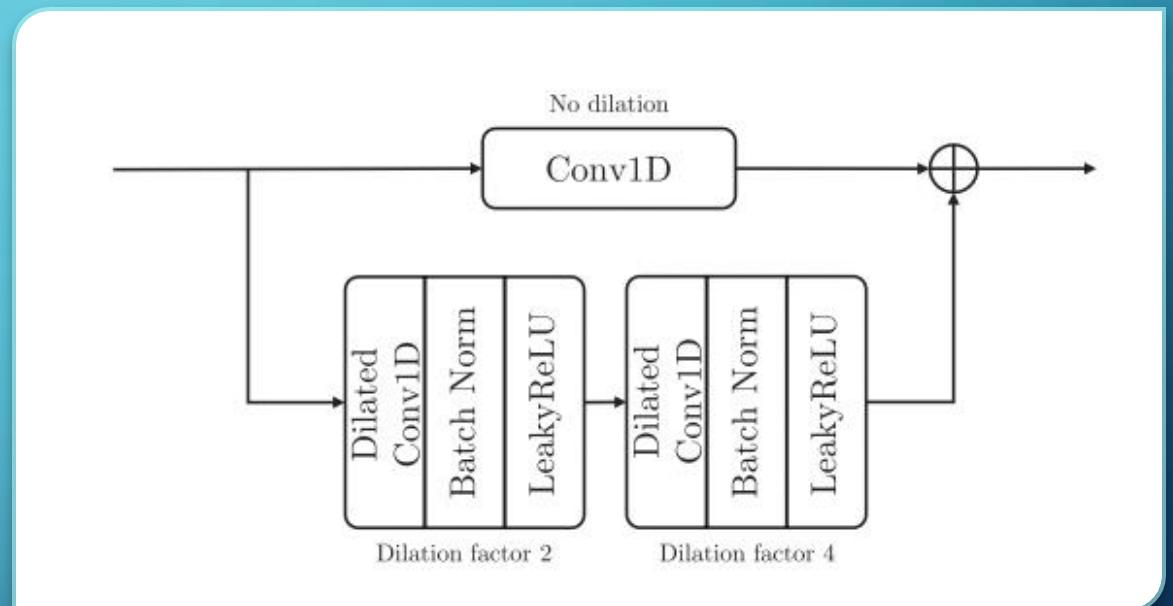
# RAMO A STRATI CONVOLUTIVI

Il ramo a strati convolutivi è a bottleneck.  
I blocchi "E" e "D" rappresentano  
rispettivamente il codificatore ed il  
decodificatore. Ognuno comprende  
blocchi di dilatazione residua, con in  
caduta rispettivamente un blocco  
downsample basato su max-pooling ed  
un blocco upsample, entrambi di ordine  
2.



# DILATED RESIDUAL BLOCK

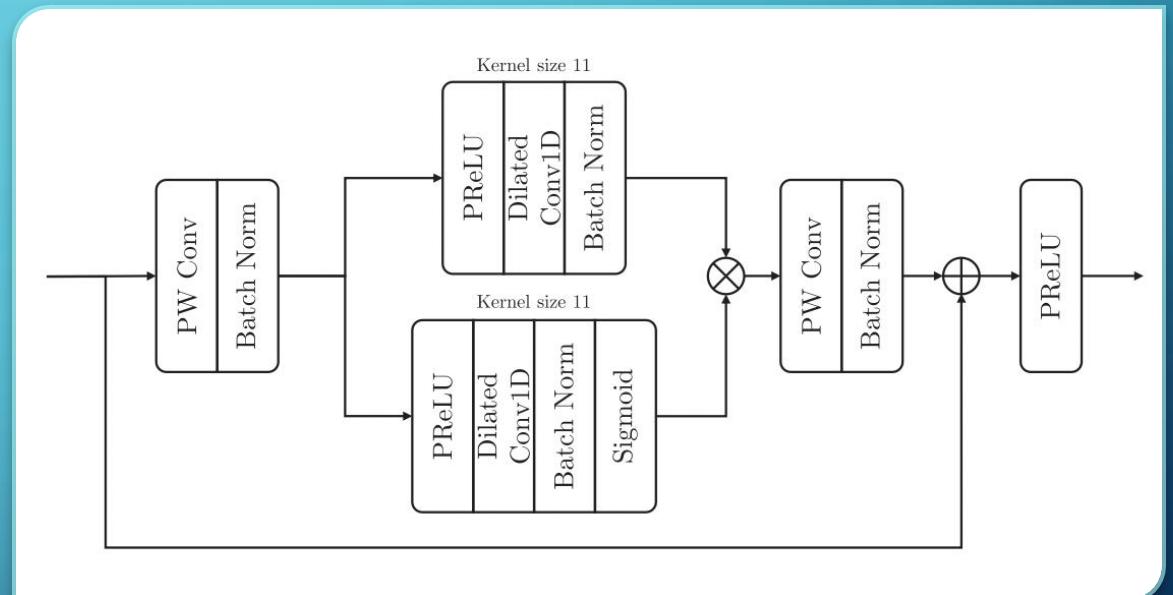
Ogni blocco comprende una connessione che passa attraverso uno strato convoluzionale senza dilatazione e due pile di convoluzioni con fattore di dilatazione rispettivamente di due e quattro, normalizzazione batch e LeakyReLU con pendenza  $\alpha= 0,2$ . Il numero di filtri cresce progressivamente nell'encoder (8, 16, 32, 64) e diminuisce simmetricamente nel decoder (64, 32, 16, 8). Gli strati convoluzionali hanno filtri di dimensione 11 nell'encoder e sette nel decoder.



# BOTTLENECK

Il bottleneck , invece, comprende  $B=6$  blocchi costituiti da uno stack di input, una gated linear unit (GLU) con dimensione kernel 11 e uno stack di output, nonché un percorso residuo che abbrevia l'input e l'output del blocco, seguito da PReLU. Gli stack di input e output presentano una convoluzione puntuale (PW) con fattore di dilatazione di 1 e 32 e 64 canali, rispettivamente, seguita da normalizzazione batch. Il tasso di dilatazione cresce esponenzialmente con ogni GLU al fine di catturare la correlazione tra campioni sempre più distanti. Vale a dire, lo strato  $j$ -esimo del collo di bottiglia ha una velocità di dilatazione di  $2^{(j-1)}$  ,  $j=1,\dots,B$ .

L'output del decoder viene quindi immesso in una convoluzione PW seguita da una funzione di attivazione della tangente iperbolica. Tutte le convoluzioni nel modello sono 1D.



# ADATTAMENTO DEL DATASET

Il Dataset utilizzato è quello fornito per la challenge "INTERSPEECH 2022".

Il dataset originale comprende tre gruppi:

1. Train-set: per la procedura di addestramento;
2. Validation-set: per eseguire la validation;
3. Blind-set: per la procedura di test.

Le porzioni train-set e validation-set contengono per ogni file audio incluso:

1. Clean\_signals;
2. Lossy\_signals;
3. File di metadati.

Il dataset è stato riadattato per l'utilizzo in questo elaborato: infatti per la fase di train sono stati utilizzati i 23184 segnali forniti nel "train-set", dei quali una piccola percentuale è stata assegnata alla validation. Per quanto riguarda invece il test-set è stata utilizzata la porzione "blind-set" scartando 2 segnali che sono risultati essere silenziosi. Quindi in tutto il test è stato eseguito su 964 segnali.

# TRAIN-SET

Per il train-set, i segnali da elaborare sono stati ricampionati ad una SR di 32kHz. Non viene effettuata alcuna normalizzazione in fase di caricamento in modo da rendere il sistema robusto in temini di silenzio e grandi variazioni di ampiezza del segnale. L'addestramento è stato completato con 500 epoche ognuna caratterizzata da 500 step, batch size di 128 e fade dimension di 64. Le funzioni di loss utilizzate nella fase di training sono le seguenti:

1. SpectralConvergenceLoss =  $\frac{\|Y - \tilde{Y}\|_F}{\|\tilde{Y}\|_F}$
2. LogMagnitudeSTFTLoss =  $\|\log(|Y| + \epsilon) - \log(|\hat{Y}| + \epsilon)\|_F$
3. SingleResolutionSTFTLoss =  $\frac{\|Y - \tilde{Y}\|_F}{\|\tilde{Y}\|_F} - \|\log(|Y| + \epsilon) - \log(|\hat{Y}| + \epsilon)\|_F$
4. MultiResolutionSTFTLoss =  $\frac{1}{R} \sum_{r=1}^R SR - STFTLoss_r$

# TEST-SET

Per il test, sono state simulate perdite uniformemente distanziate con un tasso di perdita del 10% e successivamente valutate le prestazioni PLC del metodo proposto rispetto alla tecnica zero-filling. Oltre ai file audio la procedura di test richiede in input altri elementi:

1. il checkpoint dell'addestramento;
2. un file .npy contenente le informazioni sui pacchetti contrassegnando con “0” un pacchetto trasmesso e ricevuto e con “1” un pacchetto perso.

Al termine del test-set viene generato un file excel che contiene tutti i risultati per le metriche calcolate per ogni segnale.

# METRICHE CALCOLATE

In fase di test-set sono state valutate diverse metriche:

1. NMSE
2. Mel-sc
3. Mel-cepstral
4. PLCMOS
5. DNSMOS
6. PESQ
7. STOI

# METRICHE CALCOLATE II

Nella tabella 1 sono inseriti i risultati delle metriche calcolate per 10 segnali di test selezionati in modo random. In questo caso i segnali sono stati ricostruiti con PARCnet

File name	nmse sig (dB)	nmse pckt (dB)	mel-sc sig	mel-sc pckt	PESQ	STOI	DNS mos	PLC mos	Mel cepstral
425422	-15,232	-6,685	0,104	0,283	2,596	0,924	3,264	2,818	-6,421
63022	-11,987	-2,672	0,158	0,524	3,152	0,960	2,570	2,365	-8,461
475285	-15,283	-6,288	0,098	0,268	3,096	0,969	3,923	3,249	-9,107
12865	-11,709	-2,418	0,149	0,404	2,388	0,973	3,489	2,875	-9,466
307528	-10,792	-1,626	0,156	0,418	2,398	0,949	3,385	2,673	-7,980
2307	-13,666	-4,048	0,117	0,298	2,814	0,977	3,259	3,064	-8,943
407950	-11,484	-2,607	0,161	0,422	2,040	0,967	3,677	3,190	-10,356
276335	-10,586	-1,367	0,167	0,493	2,863	0,964	3,637	2,874	-9,876
288077	-10,283	-1,028	0,188	0,593	1,949	0,956	3,482	1,874	-8,688
61483	-11,937	-3,027	0,137	0,371	3,072	0,965	3,819	3,241	-9,889

Tabella 1: Risultati metriche calcolate con segnali elaborati con PARCnet

# METRICHE CALCOLATE III

Nella tabella 2 sono inseriti i risultati delle metriche calcolate per 10 segnali di test selezionati in modo random. In questo caso i segnali sono stati ricostruiti con il metodo zero-filling.

File name	nmse sig (dB)	nmse pckt (dB)	mel-sc sig	mel-sc pckt	PESQ	STOI	DNS mos	PLC mos	Mel cepstral
425422	-10,128	0,000	0,220	1,000	1,409	0,907	2,261	2,767	-6,254
63022	-10,176	0,000	0,196	1,000	2,153	0,937	2,134	2,199	-8,403
475285	-9,853	0,000	0,230	1,000	1,487	0,917	2,536	3,038	-8,890
12865	-10,419	0,000	0,188	1,000	1,620	0,956	2,314	2,921	-9,387
307528	-9,921	0,000	0,218	1,000	1,973	0,947	2,591	2,805	-7,891
2307	-10,522	0,000	0,212	1,000	1,553	0,947	2,350	3,058	-8,806
407950	-9,861	0,000	0,217	1,000	1,576	0,946	2,027	3,288	-10,314
276335	-9,925	0,000	0,212	1,000	2,177	0,956	2,725	2,760	-9,958
288077	-10,031	0,000	0,198	1,000	1,788	0,954	2,697	1,890	-8,640
61483	-9,983	0,000	0,211	1,000	2,250	0,959	3,069	3,285	-9,839

Tabella 2: Risultati metriche calcolate con segnali elaborati con zero-filling

# RISULTATI A CONFRONTO

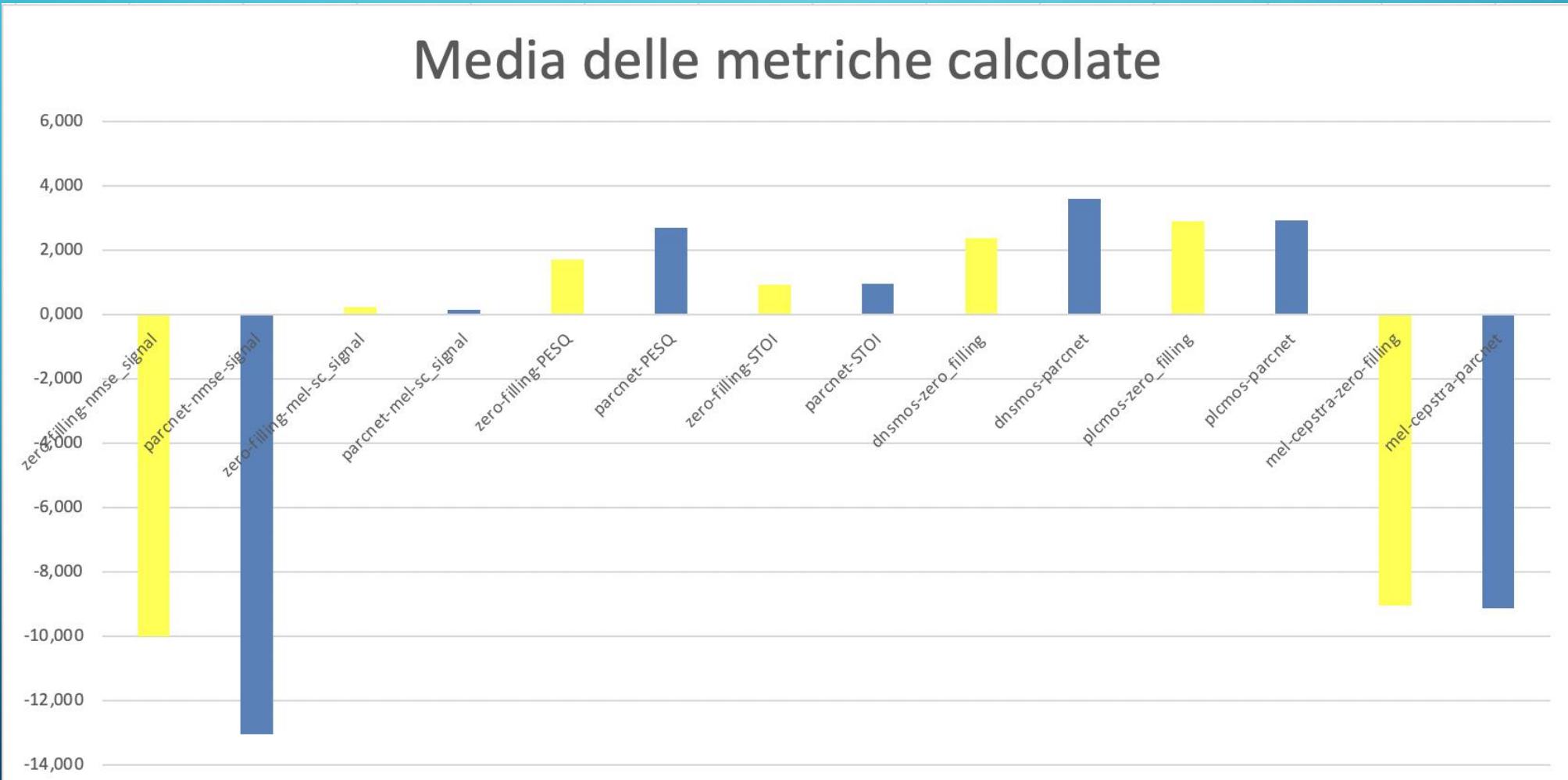
File name	nmse sig (dB)	nmse pckt (dB)	mel-sc sig	mel-sc pckt	PESQ	STOI	DNS mos	PLC mos	Mel cepstral
425422	-15,232	-6,685	0,104	0,283	2,596	0,924	3,264	2,818	-6,421
63022	-11,987	-2,672	0,158	0,524	3,152	0,960	2,570	2,365	-8,461
475285	-15,283	-6,288	0,098	0,268	3,096	0,969	3,923	3,249	-9,107
12865	-11,709	-2,418	0,149	0,404	2,388	0,973	3,489	2,875	-9,466
307528	-10,792	-1,626	0,156	0,418	2,398	0,949	3,385	2,673	-7,980
2307	-13,666	-4,048	0,117	0,298	2,814	0,977	3,259	3,064	-8,943
407950	-11,484	-2,607	0,161	0,422	2,040	0,967	3,677	3,190	-10,356
276335	-10,586	-1,367	0,167	0,493	2,863	0,964	3,637	2,874	-9,876
288077	-10,283	-1,028	0,188	0,593	1,949	0,956	3,482	1,874	-8,688
61483	-11,937	-3,027	0,137	0,371	3,072	0,965	3,819	3,241	-9,889

Tabella 1: Risultati metriche calcolate con segnali elaborati con PARCnet

File name	nmse sig (dB)	nmse pckt (dB)	mel-sc sig	mel-sc pckt	PESQ	STOI	DNS mos	PLC mos	Mel cepstral
425422	-10,128	0,000	0,220	1,000	1,409	0,907	2,261	2,767	-6,254
63022	-10,176	0,000	0,196	1,000	2,153	0,937	2,134	2,199	-8,403
475285	-9,853	0,000	0,230	1,000	1,487	0,917	2,536	3,038	-8,890
12865	-10,419	0,000	0,188	1,000	1,620	0,956	2,314	2,921	-9,387
307528	-9,921	0,000	0,218	1,000	1,973	0,947	2,591	2,805	-7,891
2307	-10,522	0,000	0,212	1,000	1,553	0,947	2,350	3,058	-8,806
407950	-9,861	0,000	0,217	1,000	1,576	0,946	2,027	3,288	-10,314
276335	-9,925	0,000	0,212	1,000	2,177	0,956	2,725	2,760	-9,958
288077	-10,031	0,000	0,198	1,000	1,788	0,954	2,697	1,890	-8,640
61483	-9,983	0,000	0,211	1,000	2,250	0,959	3,069	3,285	-9,839

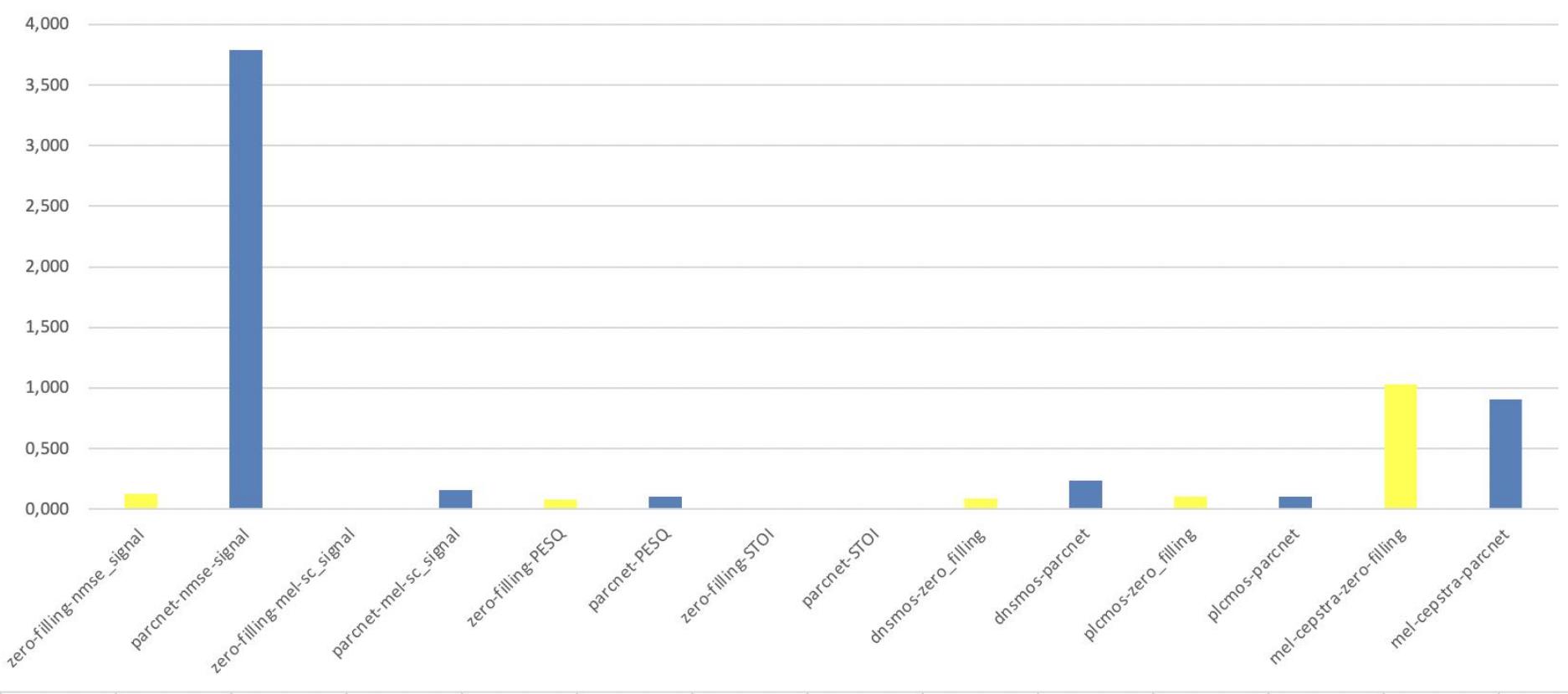
Tabella 2: Risultati metriche calcolate con segnali elaborati con zero-filling

# ISTOGRAMMI DI MEDIA E VARIANZA



# ISTOGRAMMI DI MEDIA E VARIANZA

Varianza delle metriche calcolate



# Word Error Rate (WER)

Un altro parametro calcolato a valle del test-set è il Word Error Rate. Questo è comunemente utilizzato per valutare la precisione dei sistemi di riconoscimento vocale, traduzione automatica e altri sistemi di elaborazione del linguaggio naturale. Misura l'accuratezza con cui un sistema riesce a trascrivere o tradurre il parlato o il testo, confrontando il testo prodotto dal sistema con il testo di riferimento corretto. Insieme al WER è stato valutato anche l'ACC (accuracy)

Entrambi sono valutati su:

1. Segnali clean (riferimento);
2. Segnali ricostruiti con zero-filling;
3. Segnali ricostruiti con PARCnet.

# RISULTATI WER

Dal calcolo del Word Error Rate e Accuracy risulta:

- WER-clean = 9,85% ACC= 90,15%;
- WER-zero-filling = 10,46% ACC = 89,54%;
- WER-PARCnet = 10,06% ACC = 89,94%;

# TEST DI ASCOLTO

Proponiamo 2 tracce di ascolto e per entrambe sono riportati: il file clean, il file lossy e il file PARCnet



1) CLEAN



LOSSY



PARCnet



2) CLEAN



LOSSY



PARCnet

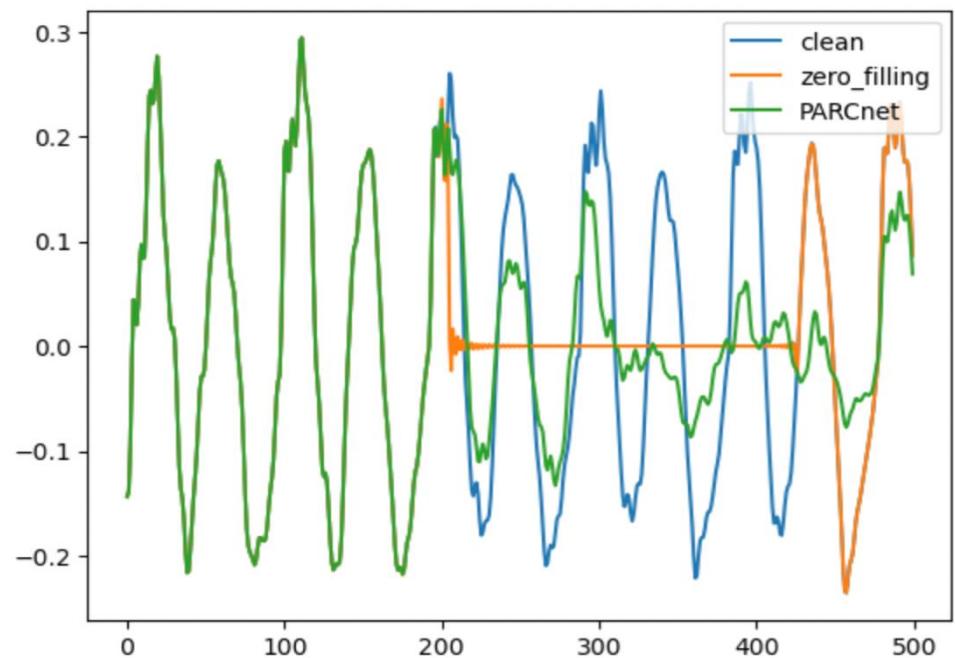
# GRAFICI DI RICOSTRUZIONE

Nella figura è illustrata la ricostruzione effettuata con i due metodi a confronto.

Nello specifico fatta sulla traccia 44660.wav. Nella figura è graficata solo quella presente nella porzione che va tra 2000 e 2500 campioni.

```
plt.plot(y_clean[2000:2500],label= 'clean')
plt.plot(y_zerofilling[2000:2500], label ='zero_filling')
plt.plot(y_parcnet[2000:2500], label= 'PARCnet')
plt.legend(loc='upper right')
plt.show()
```

✓ 0.1s

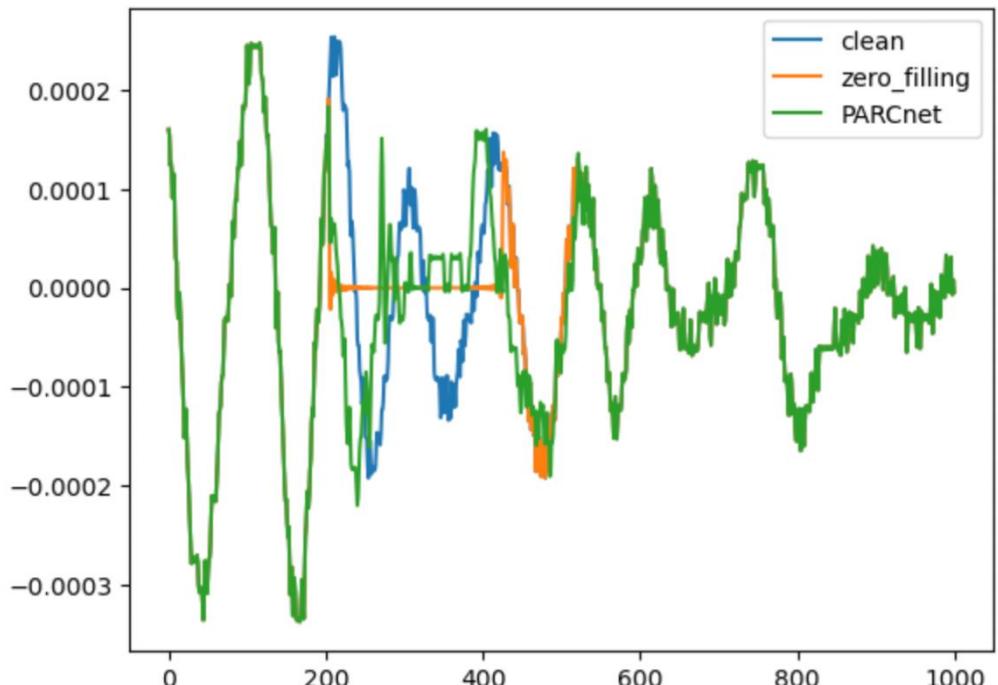


# GRAFICI DI RICOSTRUZIONE

Nella figura è illustrato un altro esempio di ricostruzione effettuata con i due metodi a confronto. Nello specifico fatta sulla traccia 54.wav. Nella figura è graficata solo quella presente nella porzione che va tra 2000 e 3000 campioni.

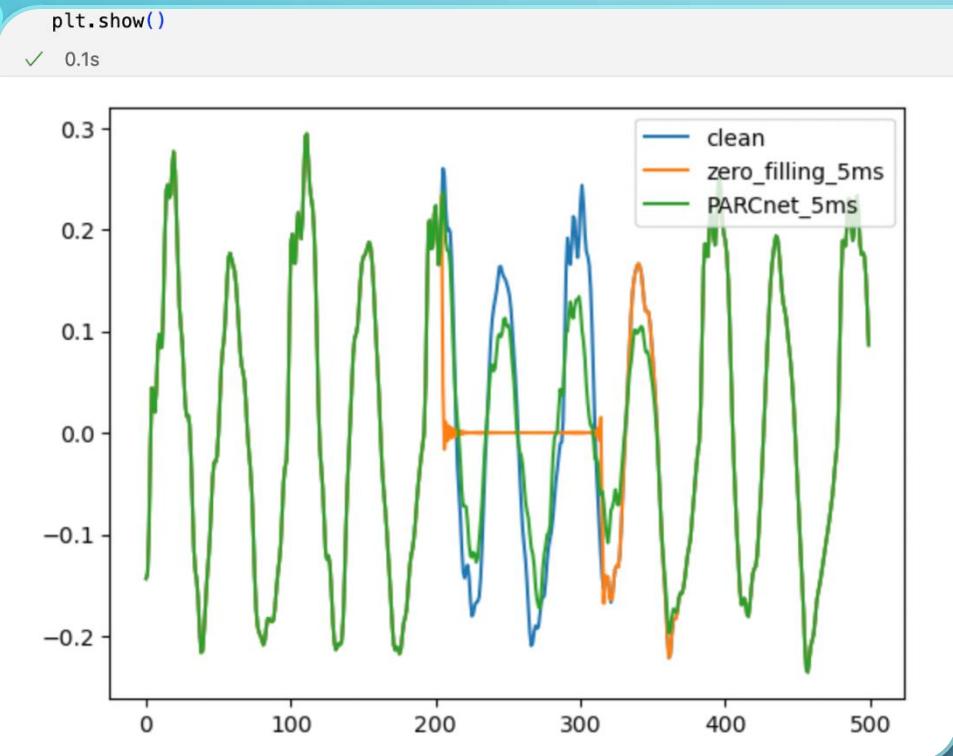
```
plt.plot(y_clean[2000:3000],label= 'clean')
plt.plot(y_zerofilling[2000:3000], label ='zero_filling')
plt.plot(y_parcnet[2000:3000], label= 'PARCnet')
plt.legend(loc='upper right')
plt.show()
```

✓ 0.1s

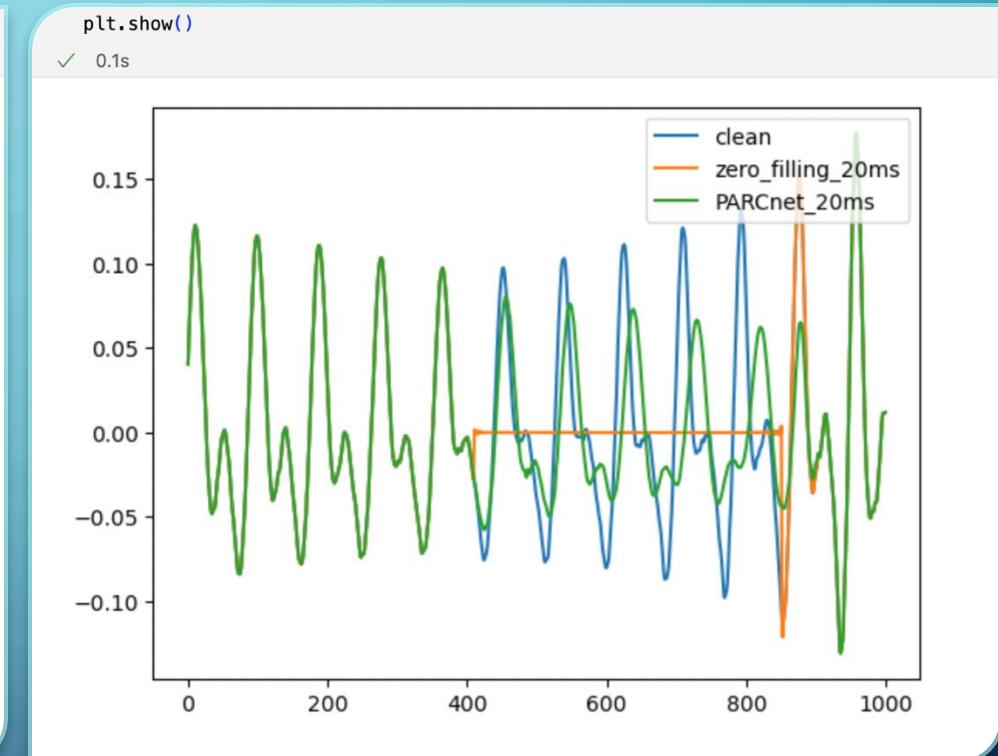


# VARIAZIONE DEL GAP

5ms



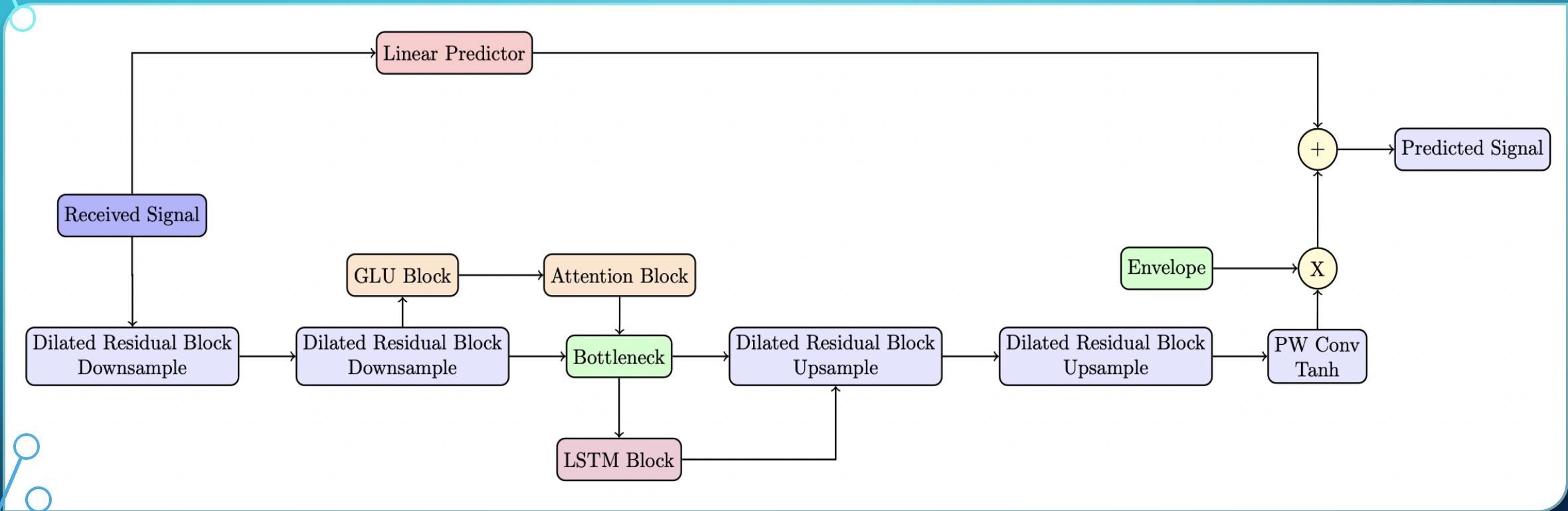
20ms



# CONCLUSIONI E CONSIDERAZIONI

1. Segnali elaborati con zero-filling
2. Segnali elaborati con PARCnet
3. Miglioramento e degrado delle prestazioni per entrambi al variare del gap
4. Modifiche effettuate al modello per migliorare il train-set

# MODELLO MODIFICATO





GRAZIE PER L'ATTENZIONE