

# Davide Gabrielli

Un Approccio alla Voice Conversion a Spettro Ridotto  
attraverso la Sine-Wave Speech

Relatore  
Prof. Danilo Avola

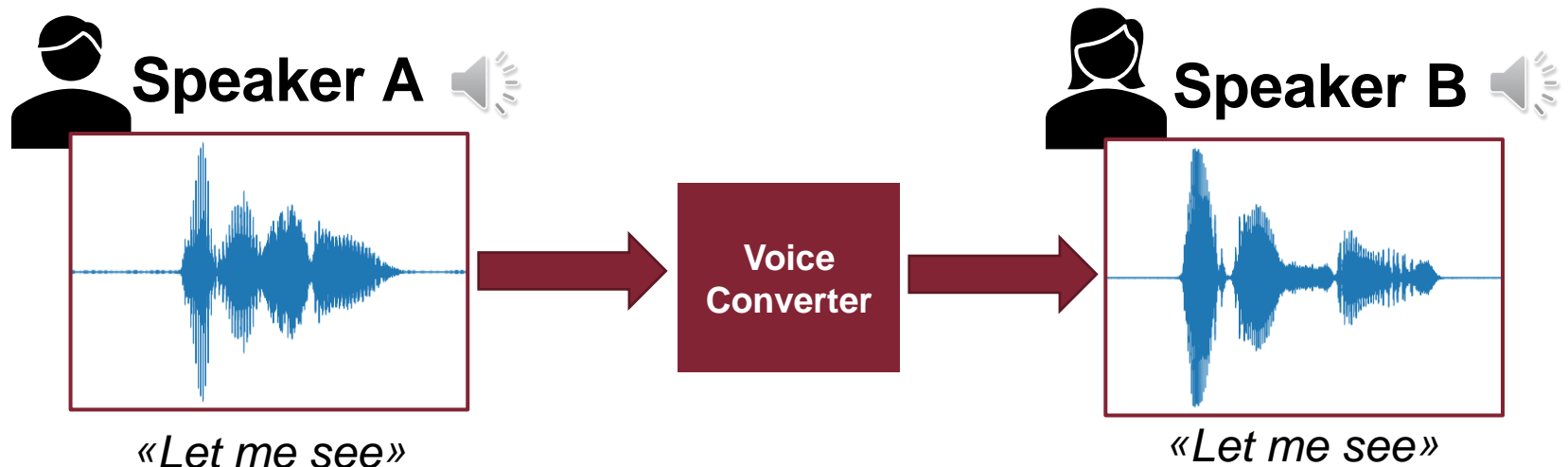
Correlatore  
Prof. Luigi Cinque  
Dr. Daniele Pannone



SAPIENZA  
UNIVERSITÀ DI ROMA

# Scopo del lavoro

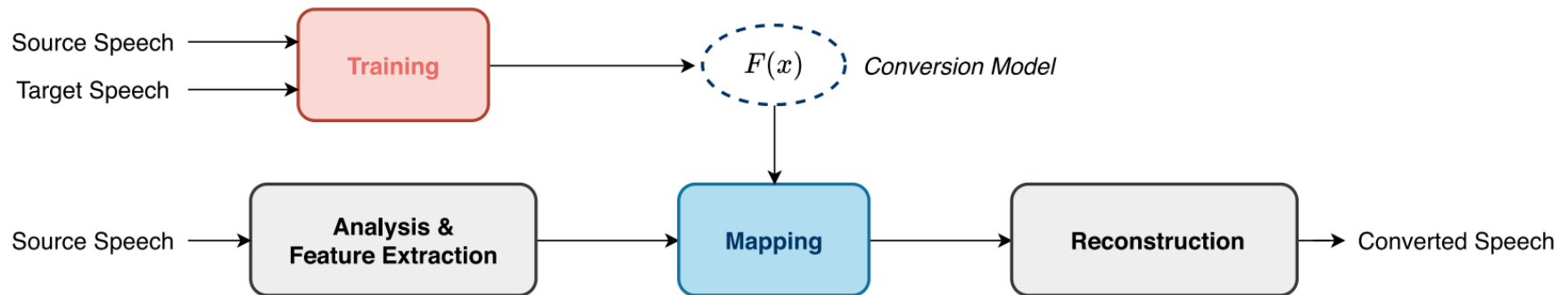
Realizzare un sistema di voice conversion che sfrutti rappresentazioni a spettro ridotto.



**Figura 1.** Esempio di una conversione di un audio da una voce maschile di uno speaker A ad una voce femminile di uno speaker B.

# Voice Conversion

Il processo di voice conversion si può suddividere in tre fasi principali:



**Figura 2.** Pipeline tipica della voice conversion.

# Il suono

Il suono è un segnale acustico prodotto dalle vibrazioni di un corpo e dalla trasmissione di queste attraverso un mezzo.

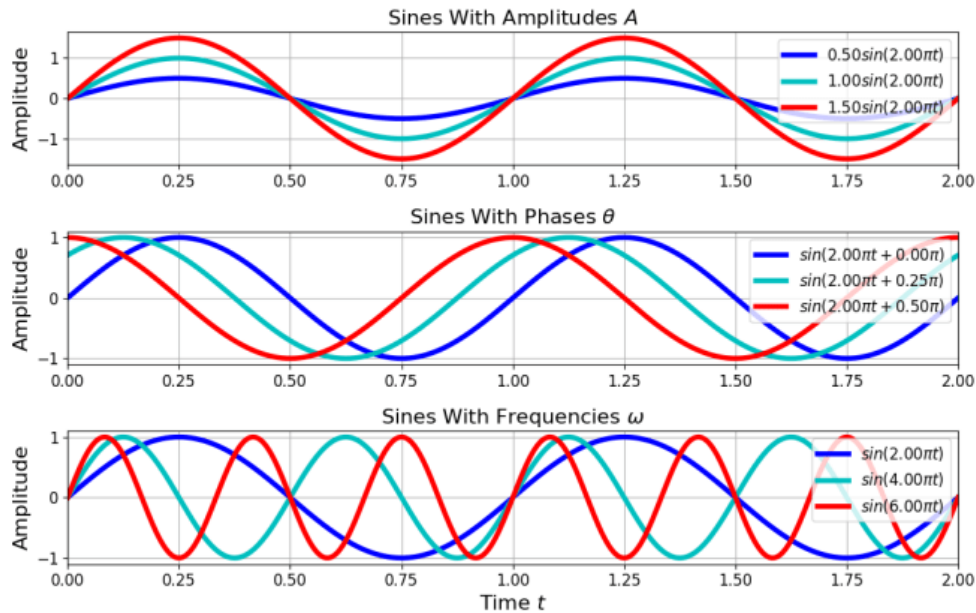


Figura 4. Onda sinusoidale.

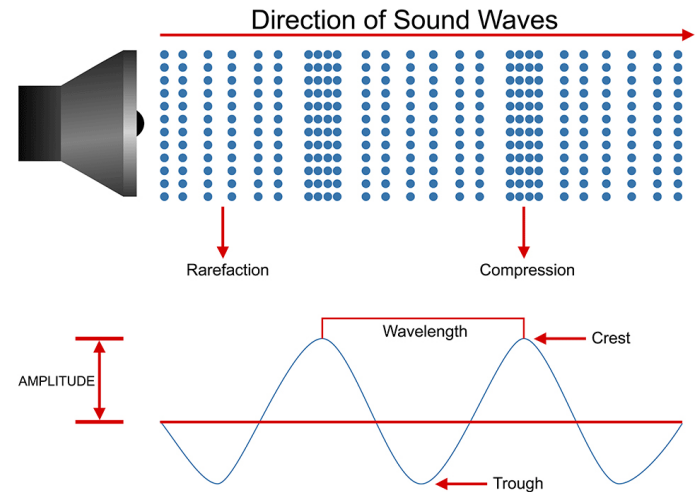


Figura 3. Il suono emesso da un oggetto.

# Rappresentazioni dell'audio (1)

Il suono, essendo un segnale continuo, per poter essere rappresentato digitalmente deve essere discretizzato:

- Campionamento

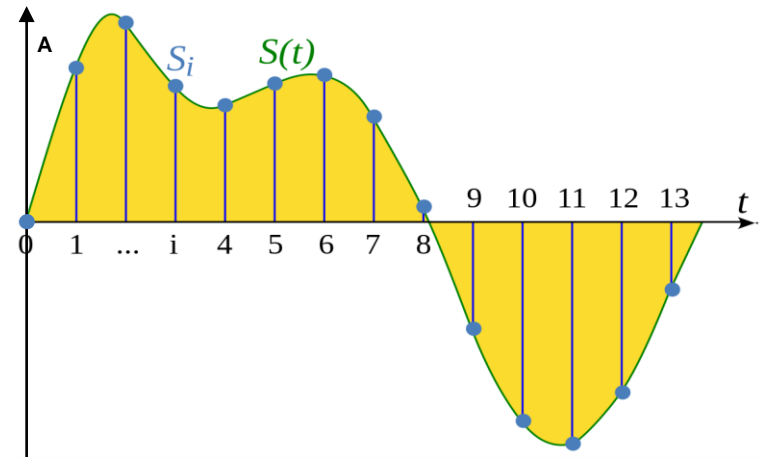


Figura 5. Esempio di un segnale campionato.

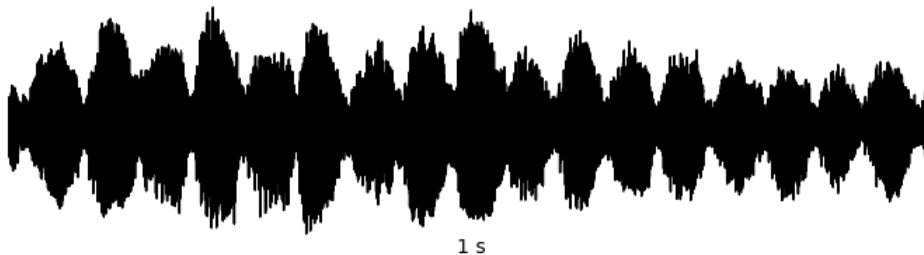
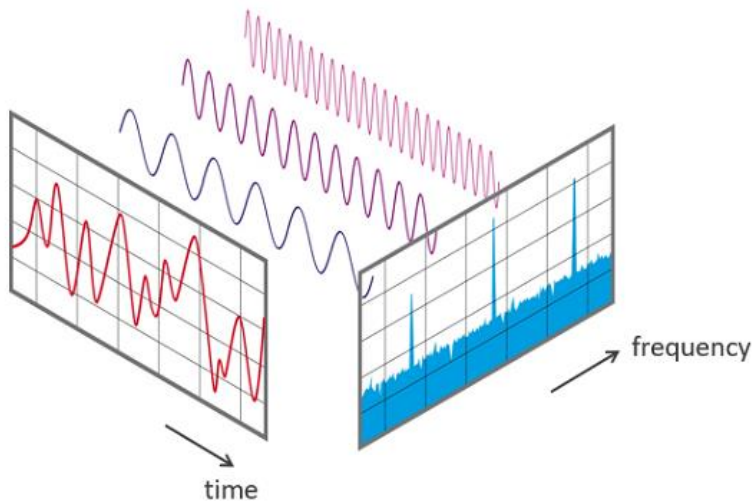


Figura 6. Un segnale complesso campionato.

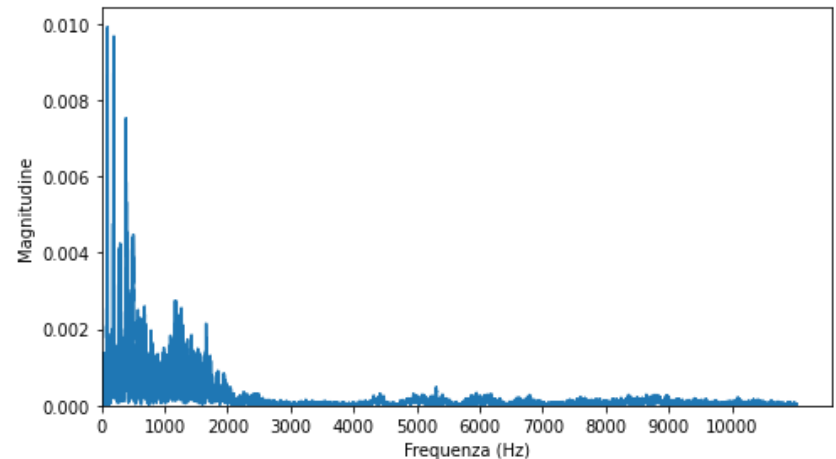
## Rappresentazioni dell'audio (2)

Tuttavia è possibile ottenere delle rappresentazioni sul dominio della frequenza:

- Spettro di potenza



**Figura 7.** Trasformata di Fourier.



**Figura 8.** Spettro di potenza.

## Rappresentazioni dell'audio (3)

Possiamo calcolare la trasformata di Fourier su piccole finestre di audio temporalmente consecutive:

- Short-time Fourier transform
- Spettrogrammi

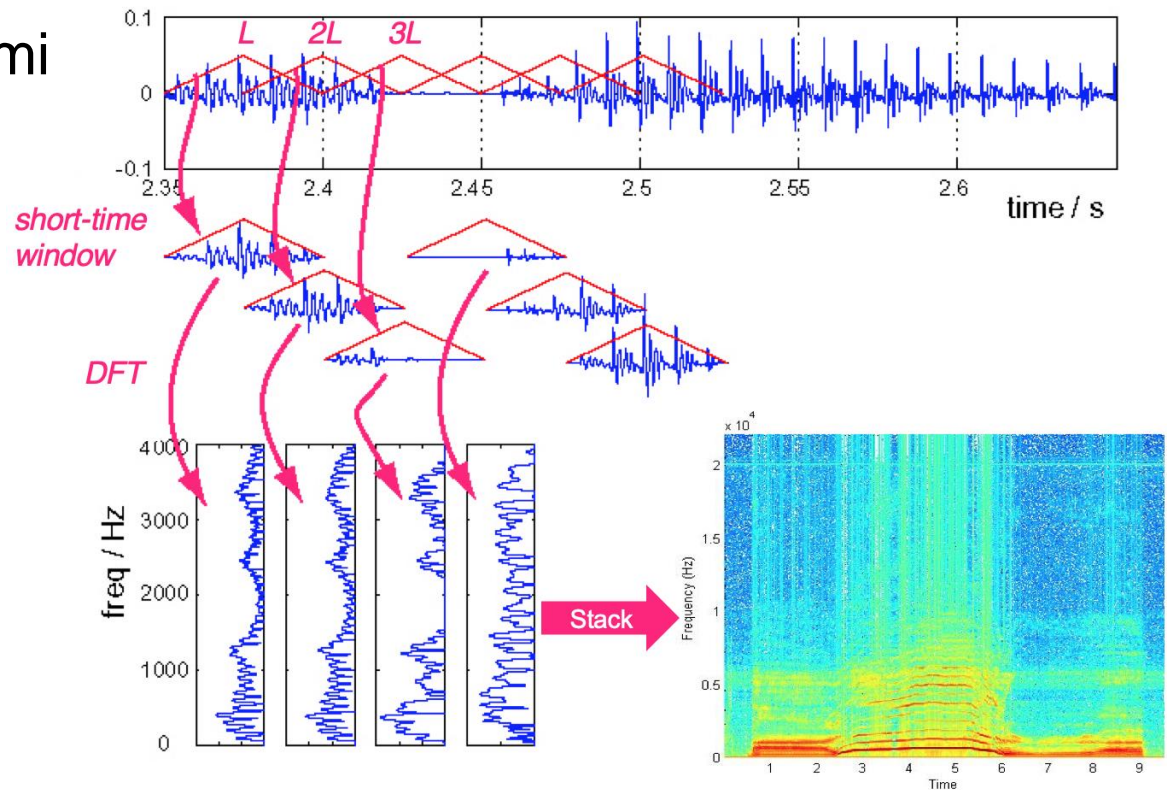


Figura 9. Short-time Fourier transform.

## Rappresentazioni dell'audio (4)

Tuttavia l'essere umano non ha una percezione lineare dell'altezza del suono (pitch) ma bensì logaritmica:

- Scala mel
- Spettrogrammi mel

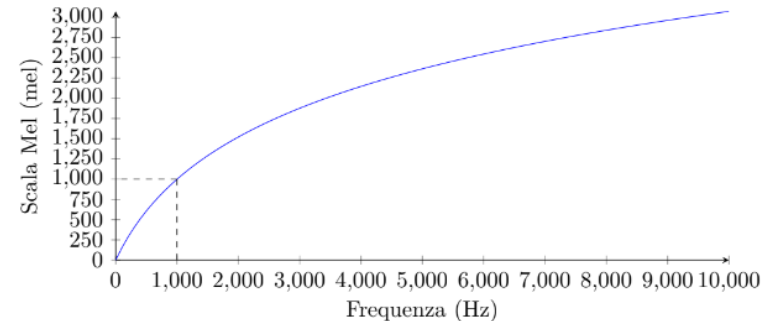


Figura 10. Scala mel.

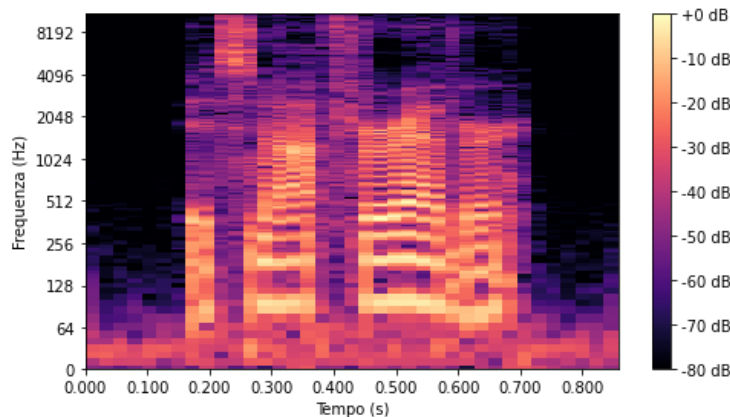


Figura 11. Spettrogramma.

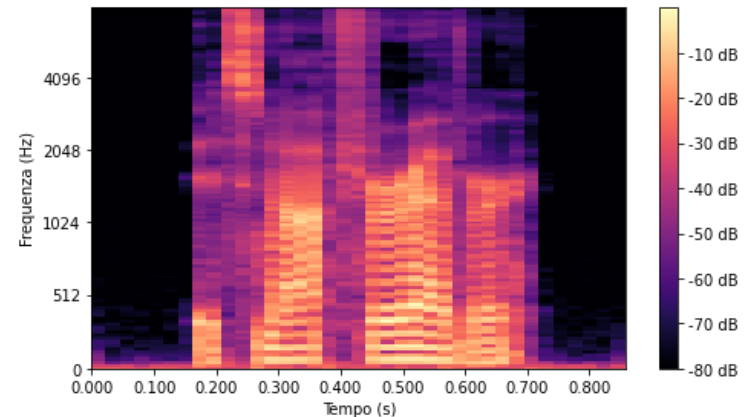
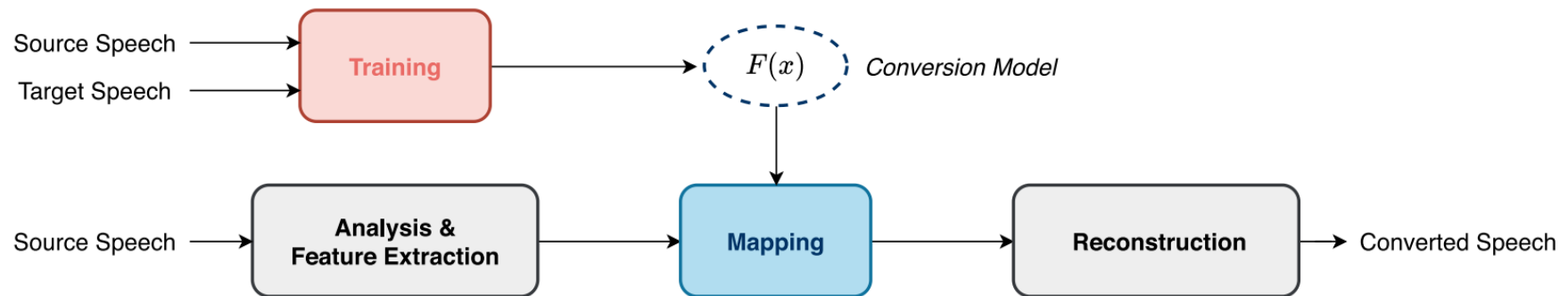


Figura 12. Spettrogramma mel.



## Voice Conversion: Mapping

Ora che abbiamo visto come rappresentare il suono, possiamo concentrarci sulla fase di mapping.



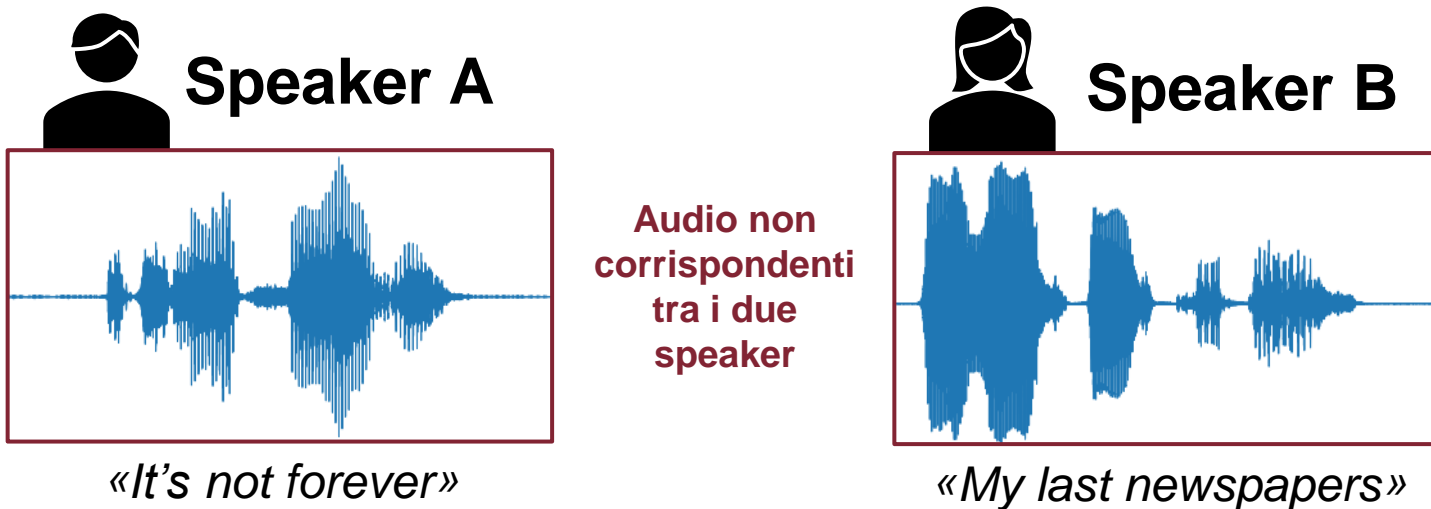
**Figura 13.** Pipeline tipica della voice conversion.

## Voice Conversion: Dati non paralleli

Addestrare un modello di conversioni tra due voci senza corrispondenza di audio tra essi.

**Pro:** facilità nel collezionare dati

**Contro:** risultati più difficili da ottenere.



**Figura 14.** Esempio di dataset formato da dati non paralleli.

# MaskCycleGAN-VC

La MaskCycleGAN-VC è lo stato dell'arte per quanto riguarda la voice conversion di una coppia di speaker su dati non paralleli.

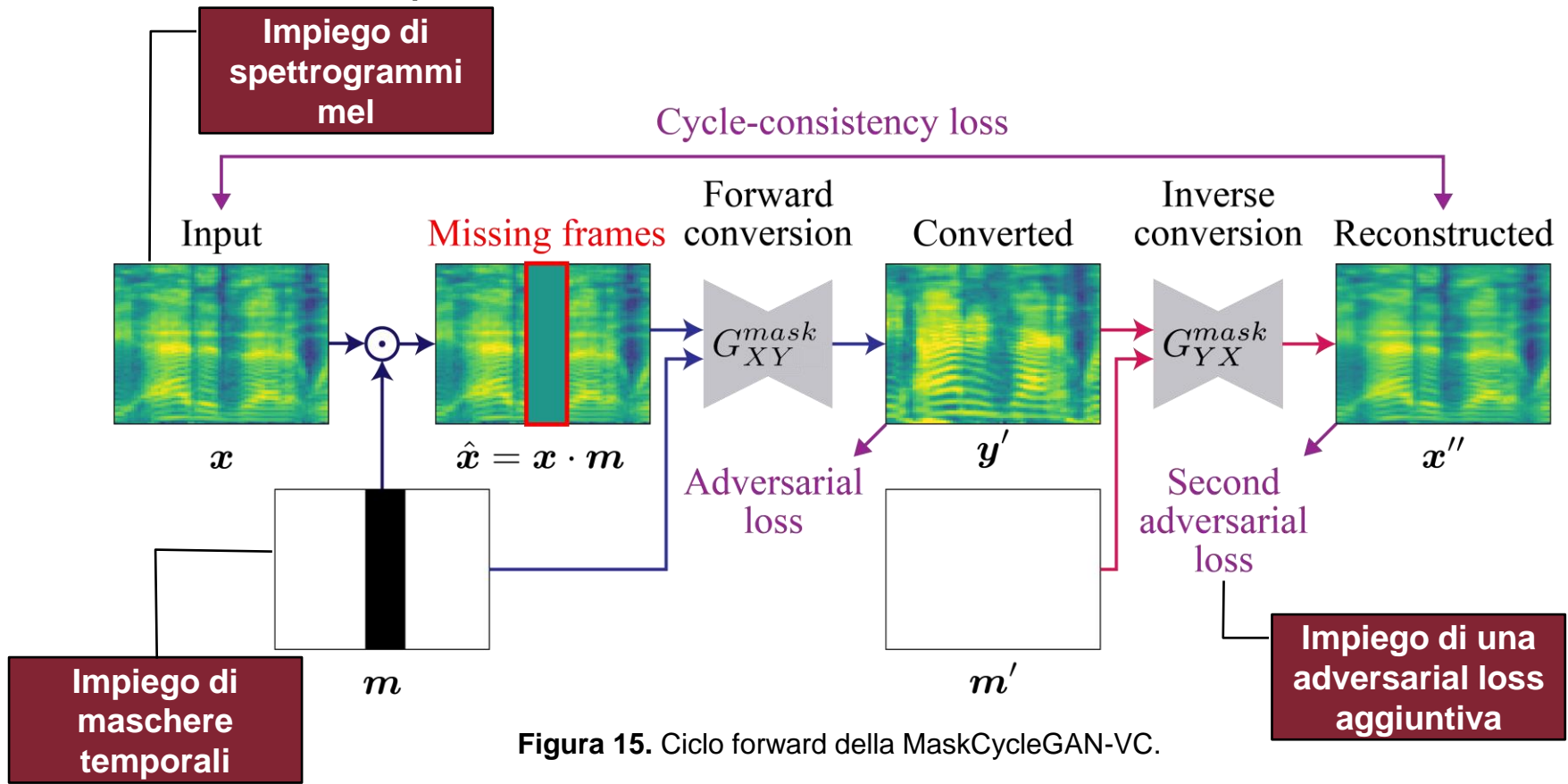
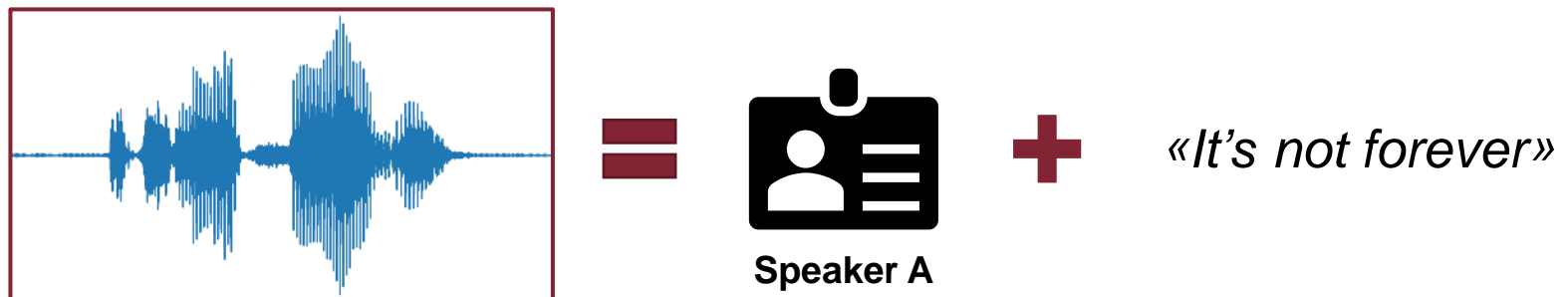


Figura 15. Ciclo forward della MaskCycleGAN-VC.

## Voce: contenuto o identità?

L'idea per cui sia possibile disaccoppiare il contenuto linguistico dall'identità vocale è ricorrente all'interno di varie aree dello speech processing.

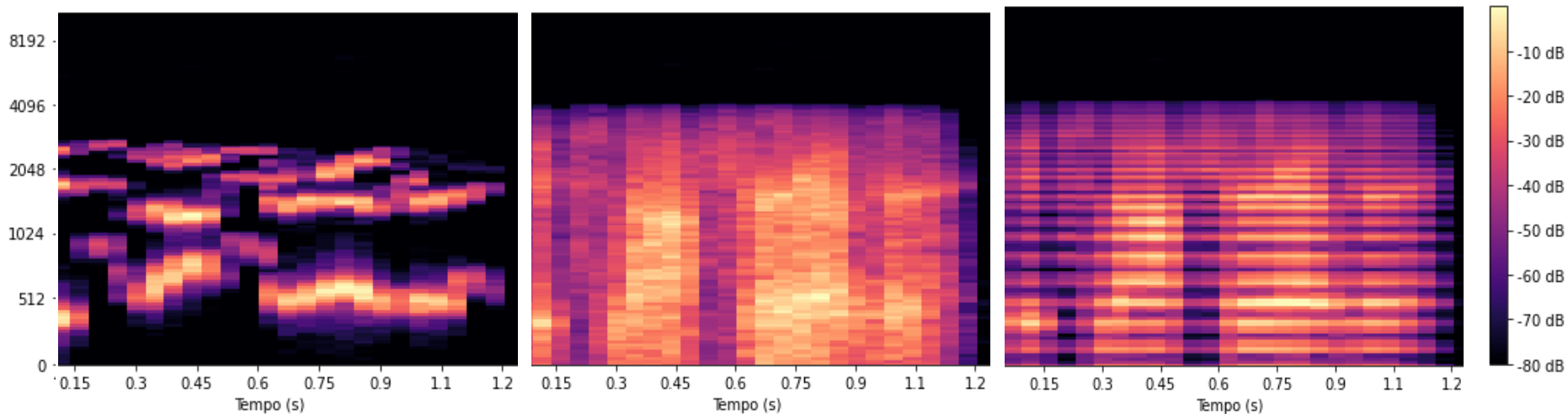


**Figura 16.** Da una voce possiamo ottenere due componenti: una linguistica e una di identità vocale.

## Rappresentazione a spettro ridotto

In questo lavoro vengono impiegate tre forme di audio a spettro ridotto:

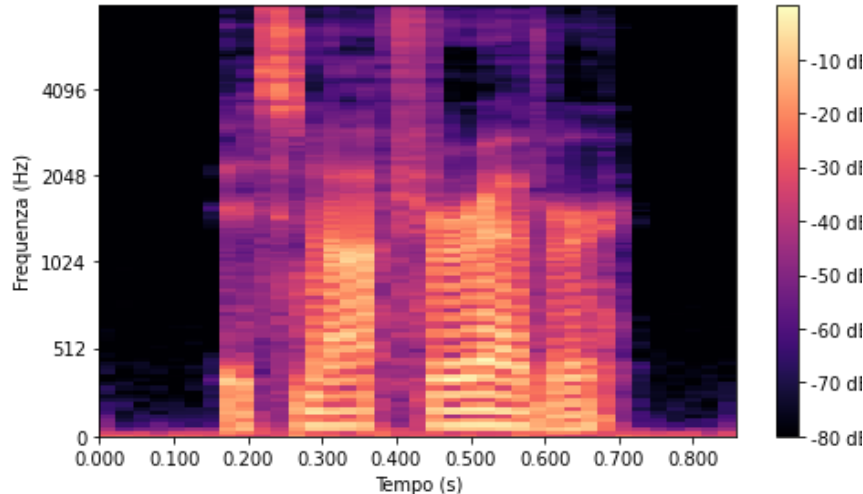
- Sine-Wave Speech
- Noise Vcoded Speech
- Buzz Vcoded Speech



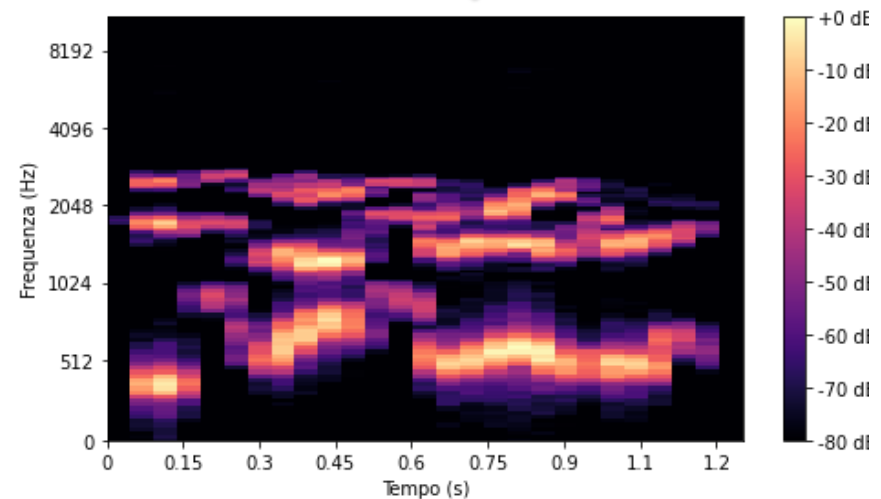
**Figura 17.** Spettrogramma mel delle tre forme a spettro ridotto impiegate in questo lavoro.

# Sine-Wave Speech

La sine-wave speech è una forma di audio del parlato umano a spettro ridotto, in cui vi sono presenti in genere tre o quattro componenti sinusoidali mobili.



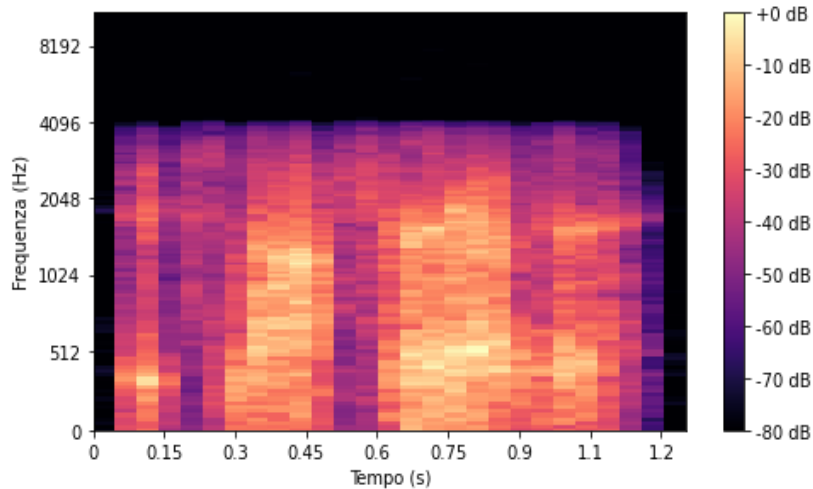
**Figura 18.** Spettrogramma mel originale.



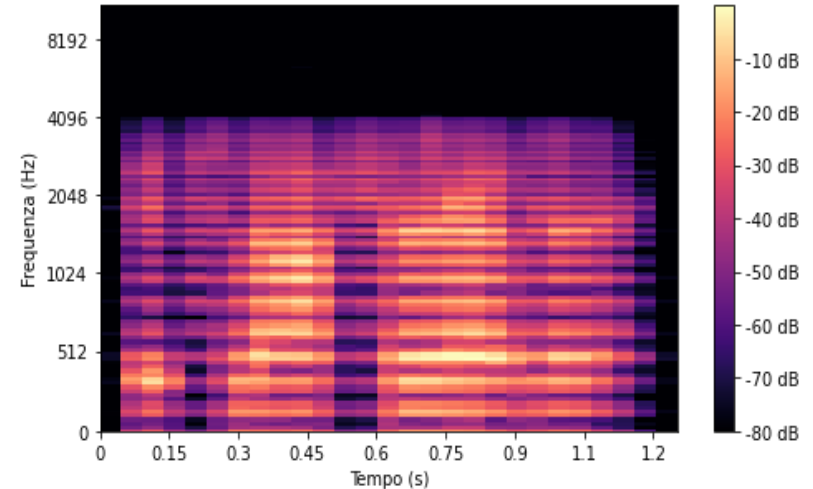
**Figura 19.** Spettrogramma mel di sine-wave speech.

# Vocoded Speech

Il vocoder è una tecnica di elaborazione dell'audio che richiede due sorgenti: un carrier, che accoglierà il suono, e un modulatore, che darà forma al suono del carrier.



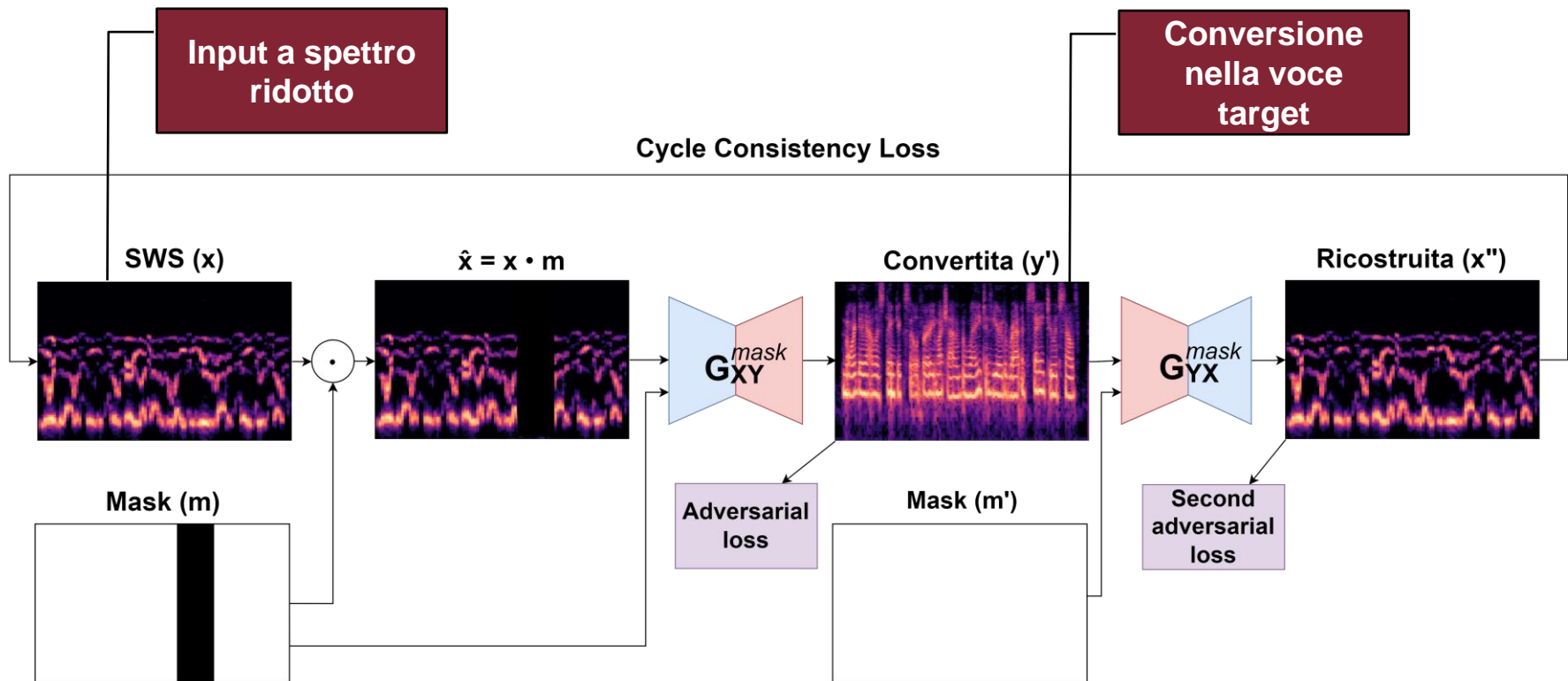
**Figura 20.** Spettrogramma mel di noise vocoded speech.



**Figura 21.** Spettrogramma mel di buzz vocoded speech.

## Architettura proposta

L'architettura proposta si ispira alla MaskCycleGAN-VC, utilizzando però audio a spettro ridotto come input.



**Figura 22.** Ciclo forward dell'architettura proposta impiegando il modulo di riduzione a SWS.



## Dataset

Il dataset impiegato è della Voice Conversion Challenge 2018, strutturato come segue:



### Training

- 4 source speaker (Task “*Hub*”)
- 4 source speaker (Task “*Spoke*”)
- 4 target speaker
- 81 audio ( $\approx 5$  min) per speaker



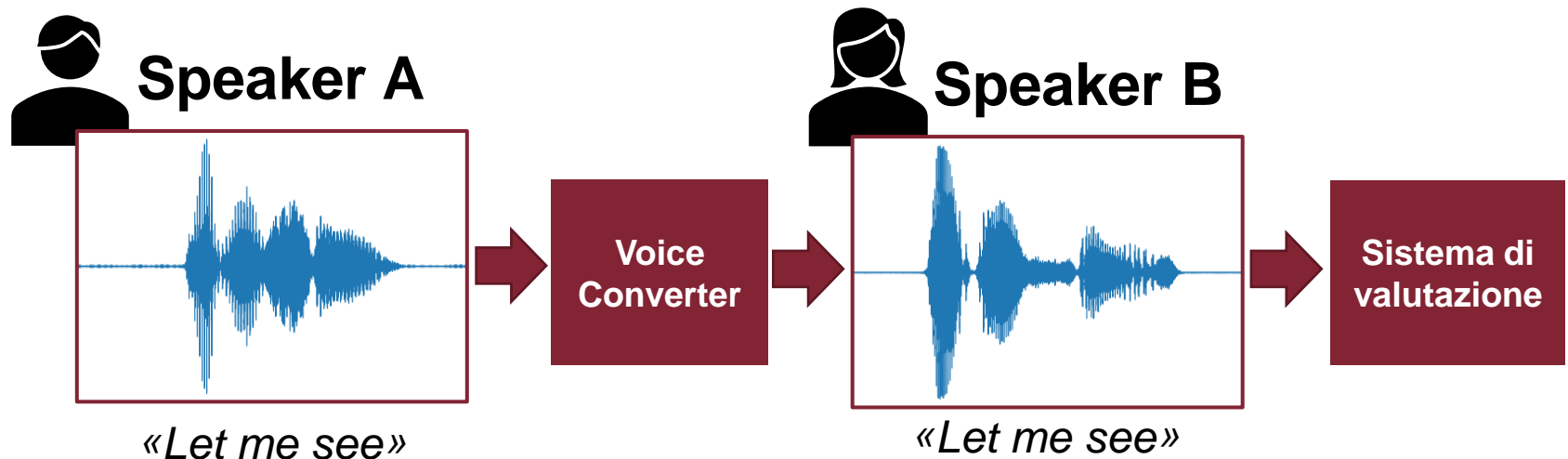
### Evaluation

- 4 source speaker (Task “*Hub*”)
- 4 source speaker (Task “*Spoke*”)
- 4 target speaker
- 35 audio ( $\approx 2$  min) per speaker

## Metriche utilizzate

Le metriche impiegate per la valutazione di audio sono le seguenti:

- Mel-cepstral Distortion
- Kernel DeepSpeech Distance
- MOSNet



**Figura 23.** Esempio di una conversione di voce che viene passata ad un sistema per la valutazione .

# Risultati

Riduzione spettro <sup>a</sup>	Valutazione	F→F	F→M	M→F	M→M
Nessuna riduzione <sup>b</sup>	MCD [dB]	6.61	6.57	6.98	6.89
	KDSD [ $\times 10^5$ ]	<b>2074</b>	<b>1755</b>	<b>2770</b>	<b>1583</b>
	MOSNet	3.84	4.46	3.92	4.58
Noise vocoded <sup>c</sup>	MCD [dB]	6.53	<b>6.47</b>	6.75	6.73
	KDSD [ $\times 10^5$ ]	3269	2247	3446	2032
	MOSNet	3.90	4.46	3.89	4.49
Buzz vocoded <sup>c</sup>	MCD [dB]	<b>6.49</b>	6.49	<b>6.70</b>	<b>6.71</b>
	KDSD [ $\times 10^5$ ]	3063	2155	3169	1823
	MOSNet	3.80	4.47	<b>3.94</b>	4.53
Sine-wave speech <sup>c</sup>	MCD [dB]	6.55	6.55	6.98	6.78
	KDSD [ $\times 10^5$ ]	3513	2621	4802	2492
	MOSNet	<b>3.91</b>	<b>4.48</b>	3.86	<b>4.61</b>

















<sup>a</sup> Modulo di riduzione dello spettro applicato sui dati di input.

<sup>b</sup> Nessuna riduzione spettrale applicata, modello trainato come proposta da Kaneko et al. in [14].

<sup>c</sup> Metodi di riduzione spettrale proposti, come descritti nella sezione 4.2.

# Risultati

Alcuni dei risultati delle conversioni con il metodo proposto:

	Source	Target	Converted	Paper
<b>F→F</b>				
<b>F→M</b>				
<b>M→F</b>				
<b>M→M</b>				

## Conclusioni e sviluppi futuri

I risultati ottenuti mostrano come sia possibile sfruttare rappresentazioni con riduzione di spettro nella voice conversion.

Per sviluppi futuri si ritiene interessante approfondire:

- Impiego per data augmentation
- Applicazione in altri campi (e.g. speech recognition)

**Grazie per l'attenzione**