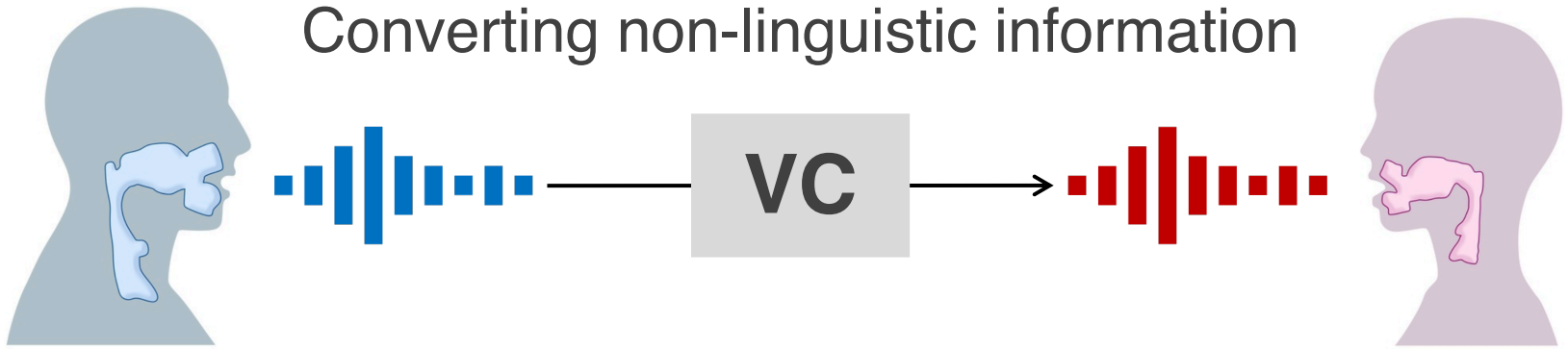
Master's thesis examination 21/01/2020

Real-Time, Full-Band, High-Quality Neural Voice Conversion with Sub-Band Modeling and Data-Driven Phase Estimation

48-196610 Takaaki Saeki

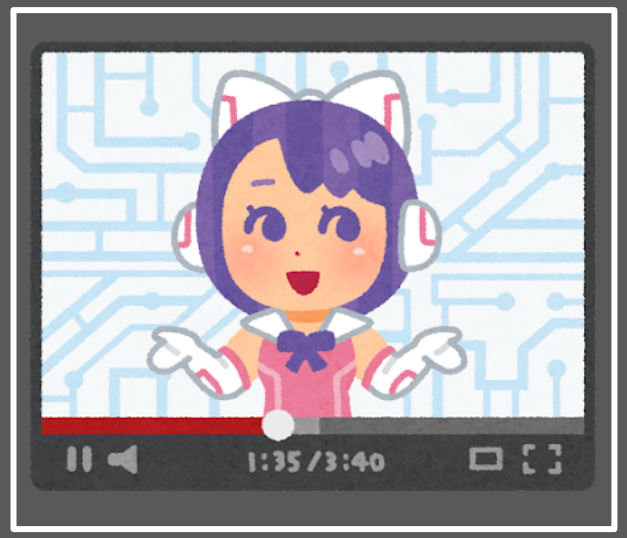
Supervisor: Prof. Hiroshi Saruwatari

Research Field: Voice Conversion (VC) [Stylianou+88]



VC removes physical constraints of human vocal organs

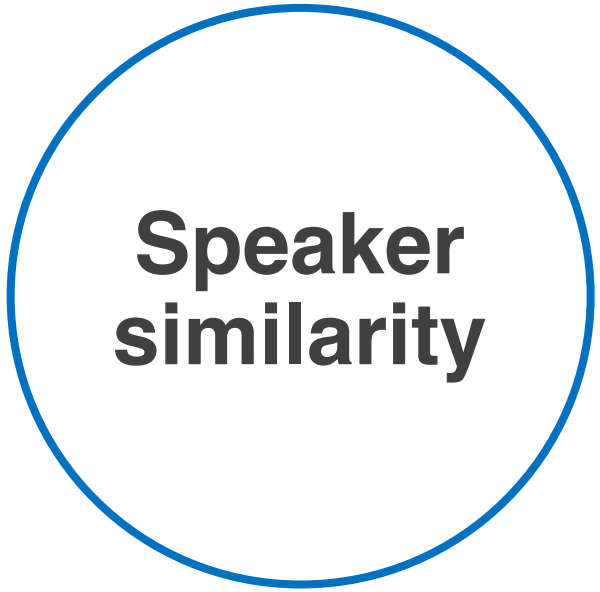
Live streaming with avatar



Assisting the speech-impaired

Enhancing musical expression

Requirements of VC



Deep neural network (DNN)-based VC

Full-band (0-24 kHz) VC

Real-time operation

VC needs all of these properties for **augmenting human communication**

Necessity for high-quality real-time VC

Overview of This Thesis

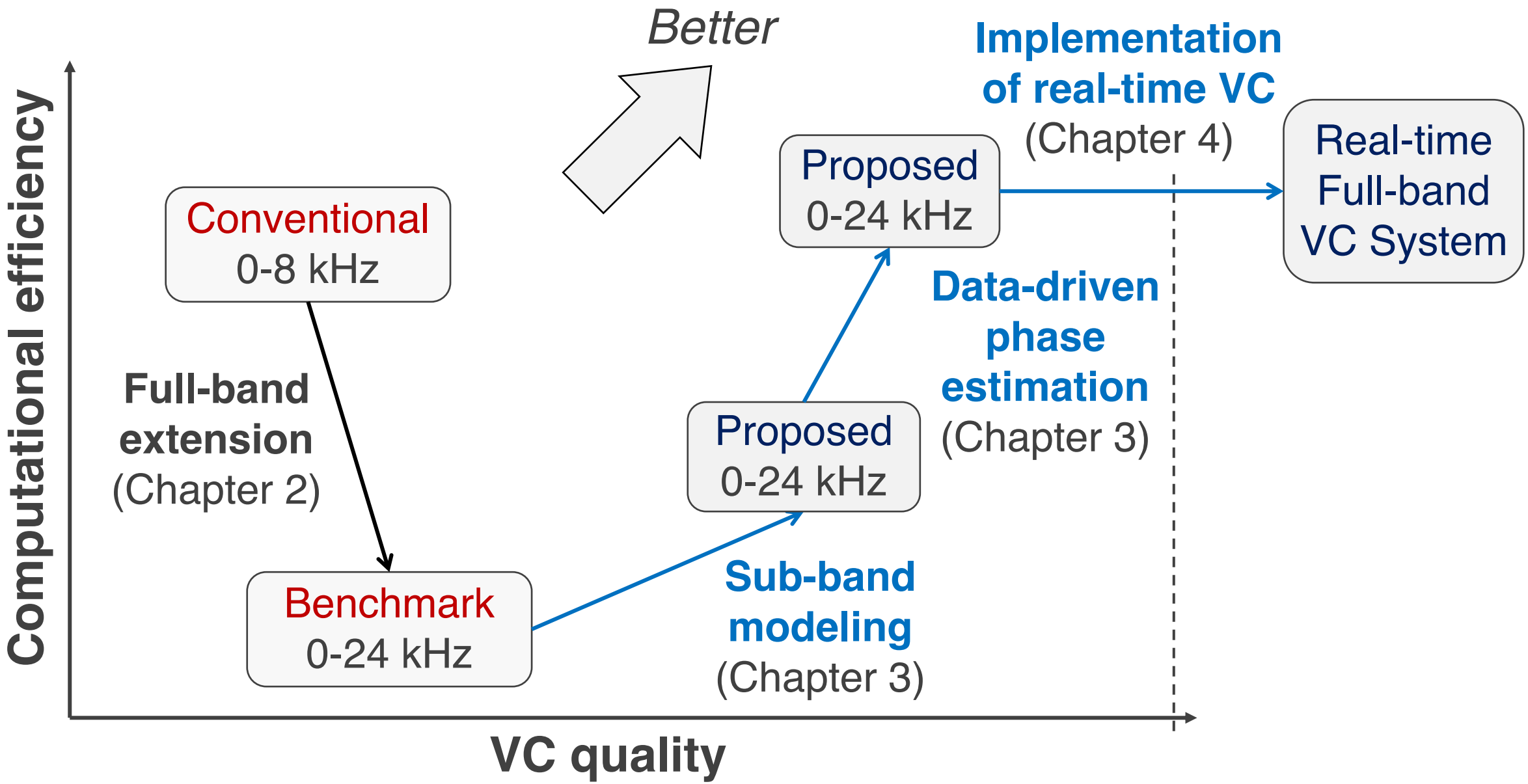
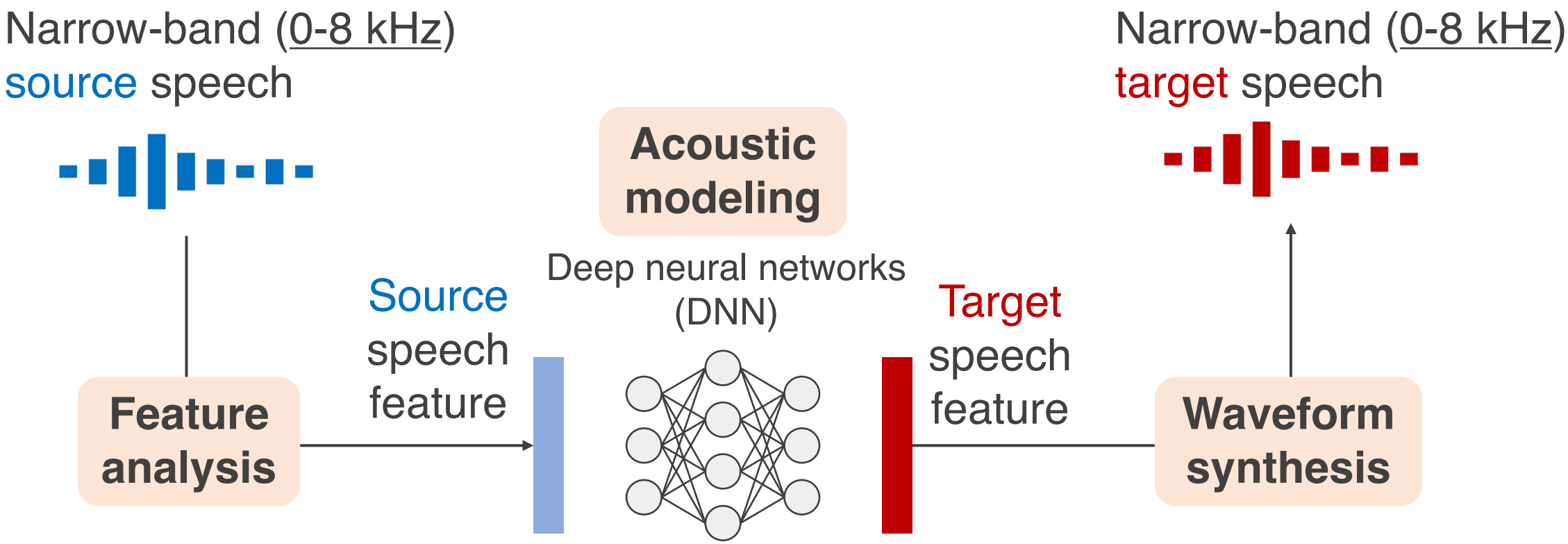


Table of Contents

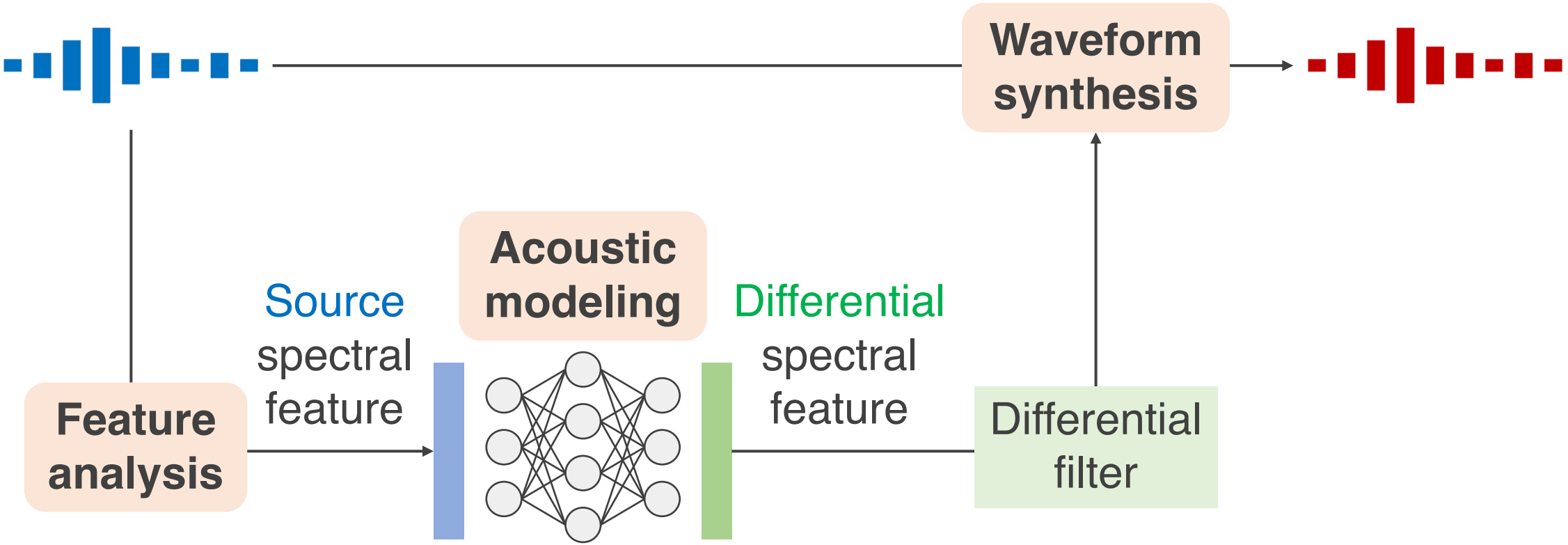
DNN-Based Real-Time VC [Arakawa+19]



Low speech quality due to narrow-band (0-8 kHz) speech

Quality degradation in waveform synthesis module

Spectral-Differential VC [Kobayashi+18]



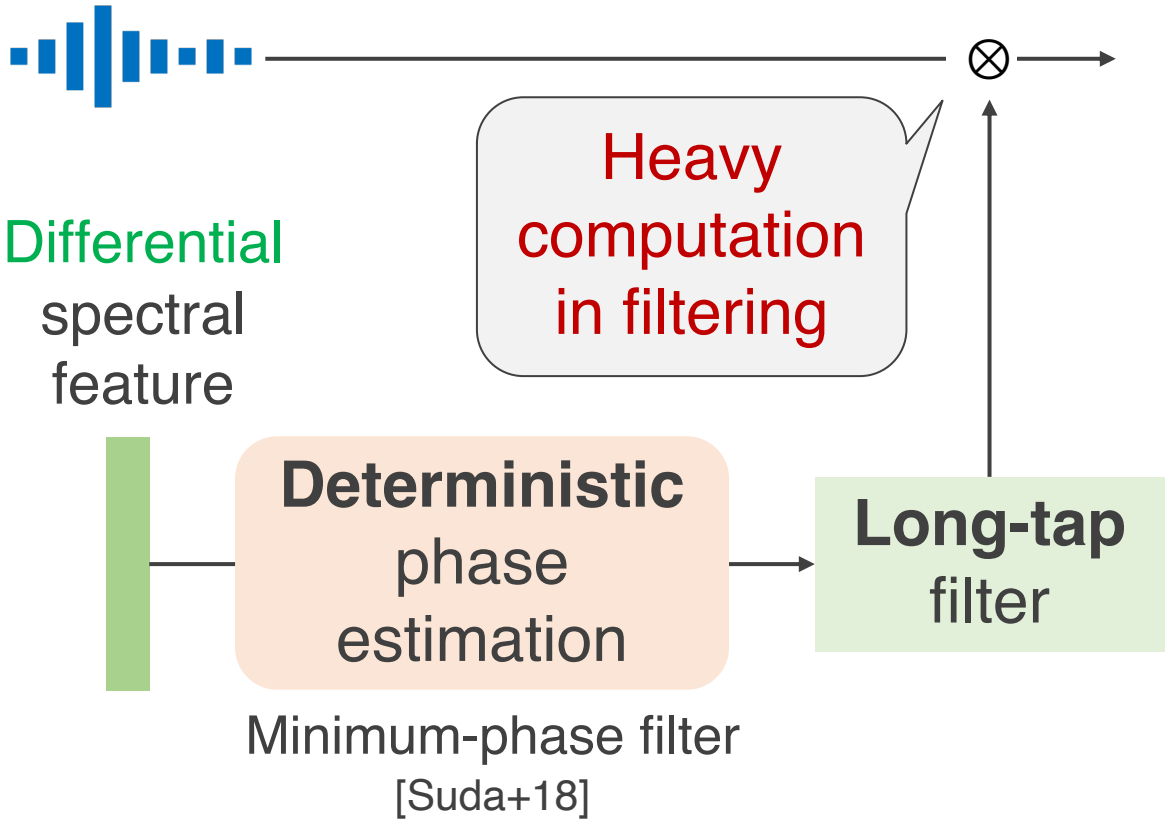
Spectral-differential VC uses filtering-based waveform synthesis

- ➢ High-quality synthesis with simple structure

Goal: real-time **full-band (0-24 kHz)** VC based on spectral-differential VC

Problems for Full-band Real-Time VC

High computational cost in waveform synthesis



Quality degradation of full-band converted speech

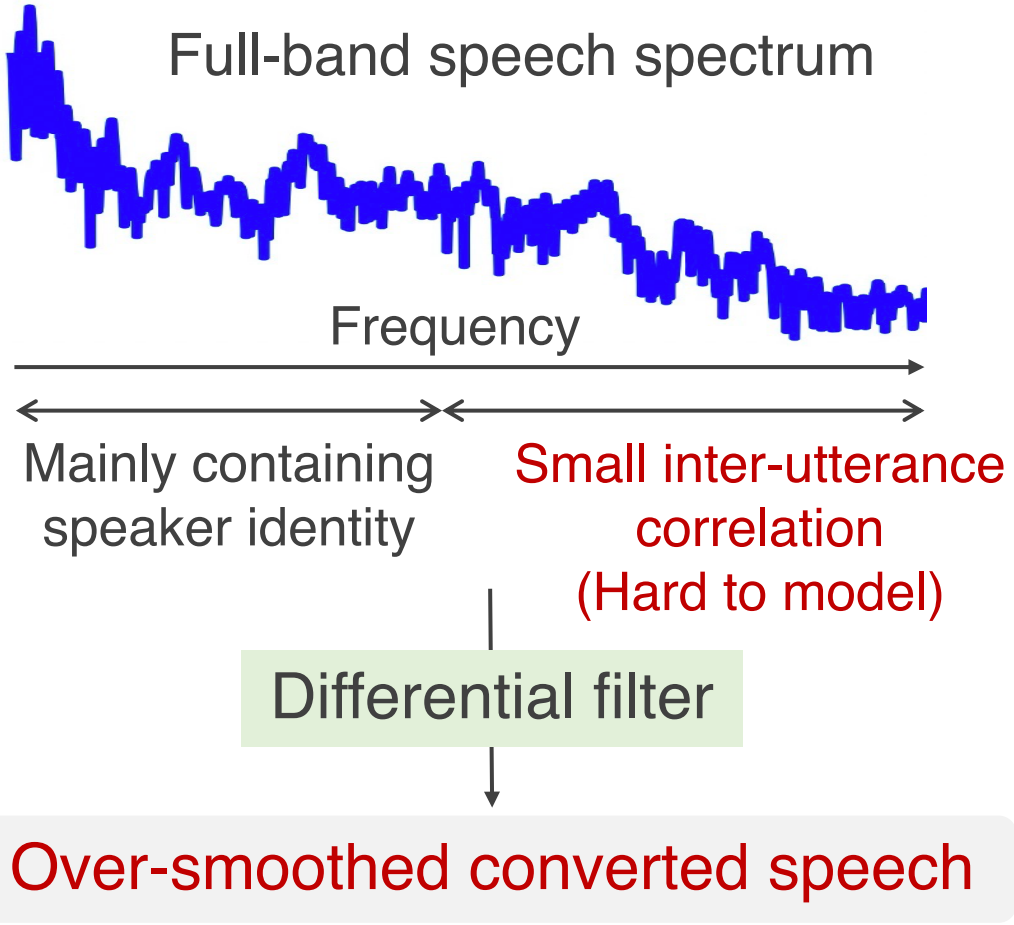
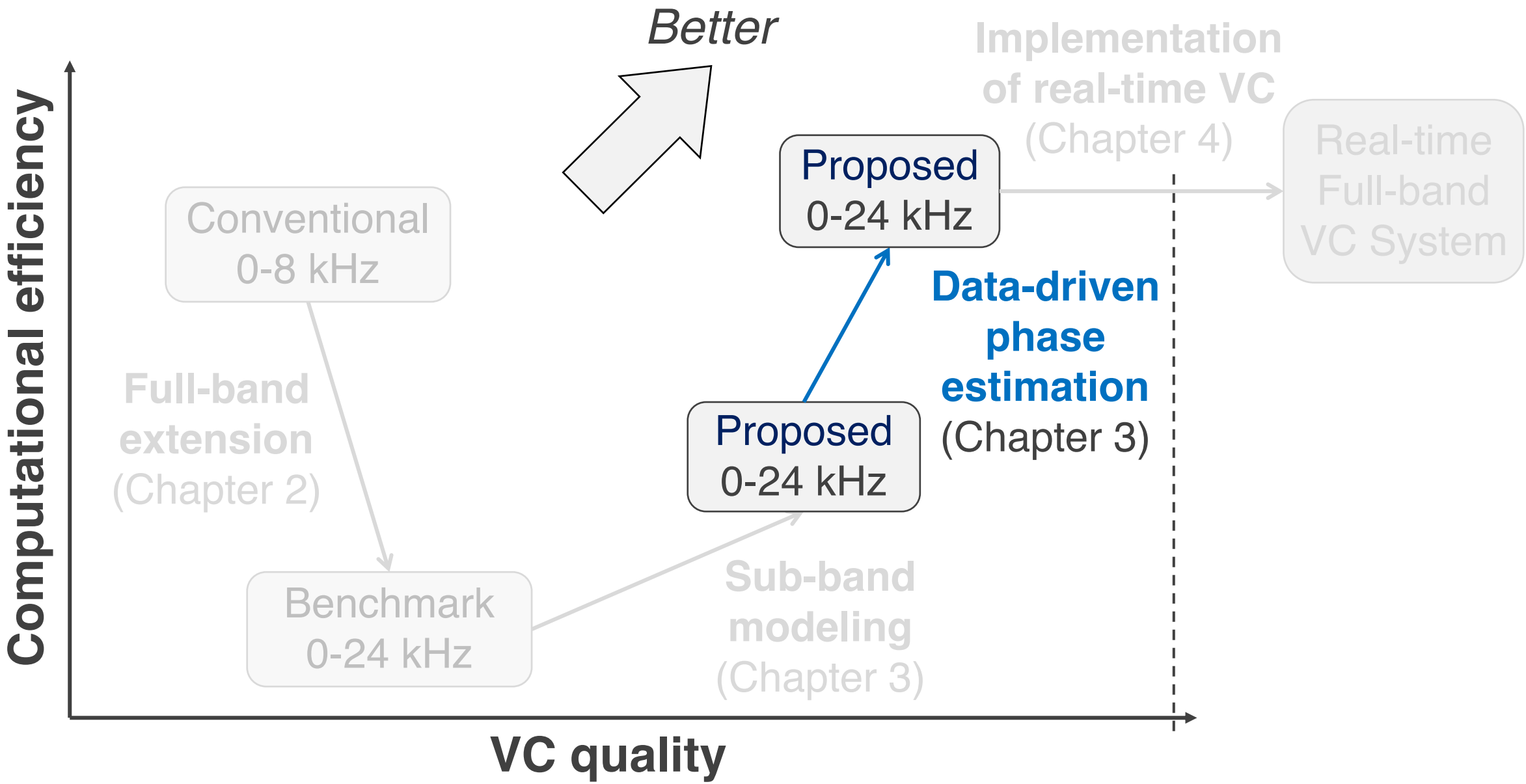


Table of Contents

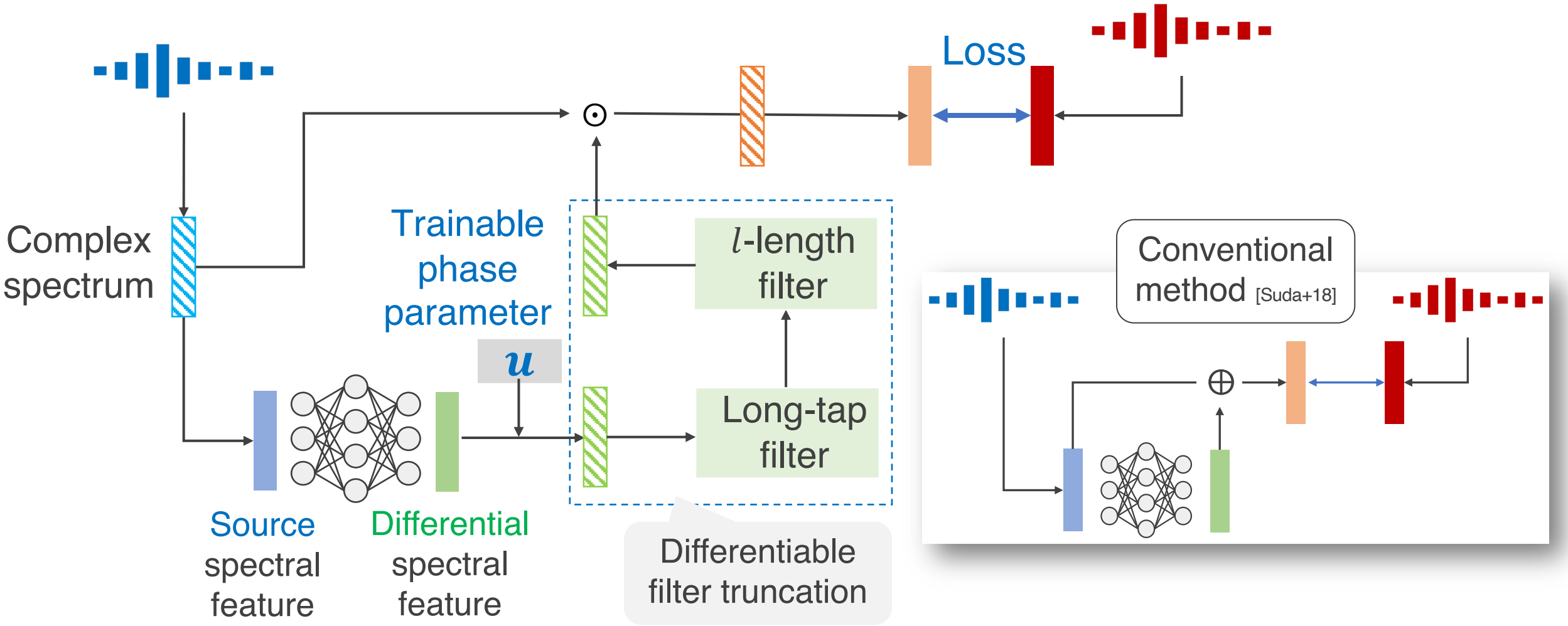
Overview of This Thesis



Training Process with Data-Driven Phase Estimation

Training phase parameter $\boldsymbol{u}$ and model parameter considering filter truncation

- Constructing short-tap filter without degrading quality



Evaluation of Data-Driven Phase Estimation

Subjective evaluations with different filter length for conventional and proposed methods (0-8 kHz)

Speaker similarity

Proposed	Score		Conventional
6 % length	**0.642**	0.358	6 % length
6 % length	**0.543**	0.457	100 % length

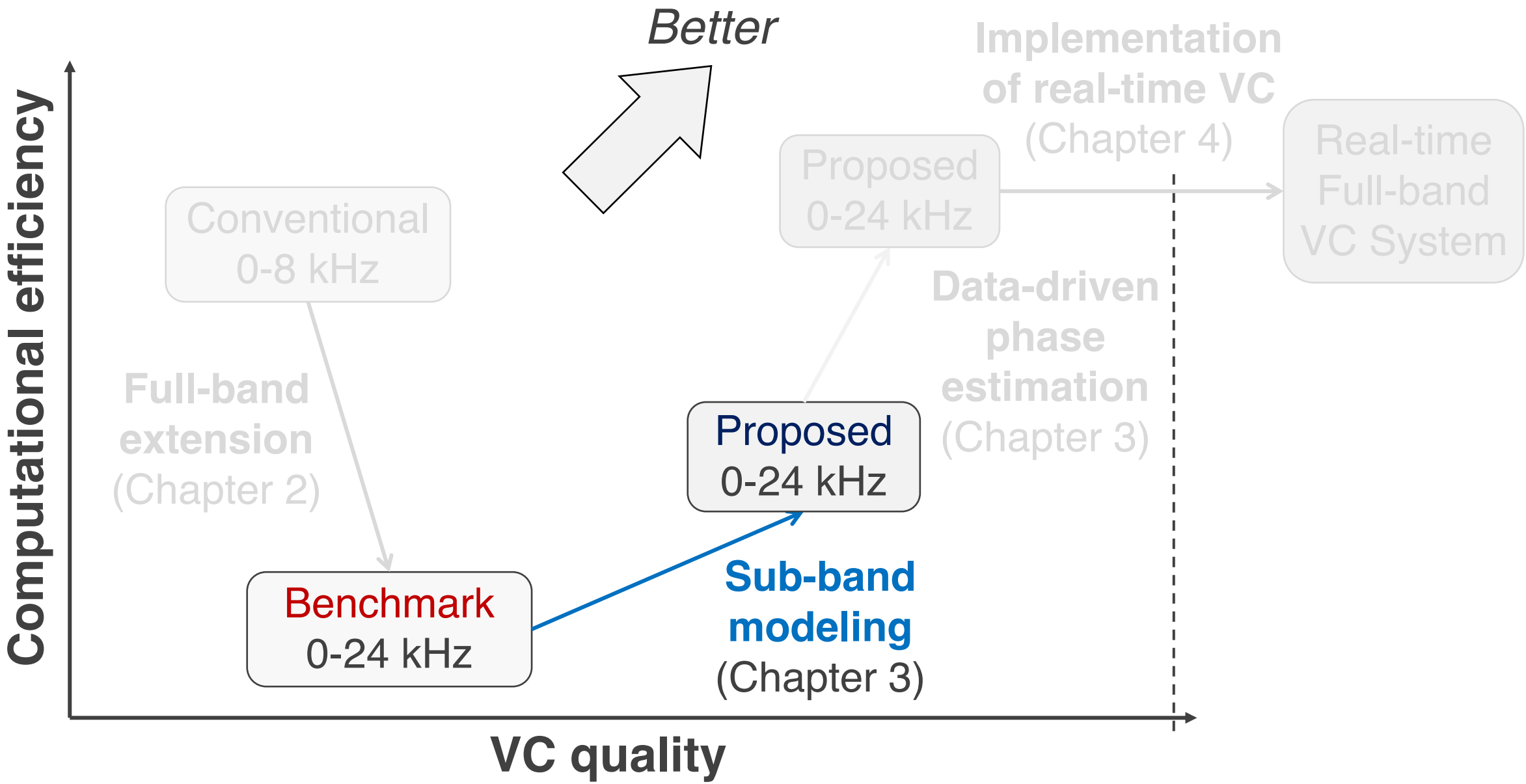
Speech quality

Proposed	Score		Conventional
6 % length	**0.807**	0.193	6 % length
6 % length	0.742	0.258	100 % length

- ❑ Short-tap proposed > Short-tap conventional
- ❑ Short-tap proposed >= Full-tap conventional

Proposed method can reduce computational cost by **94 %** while maintaining quality!

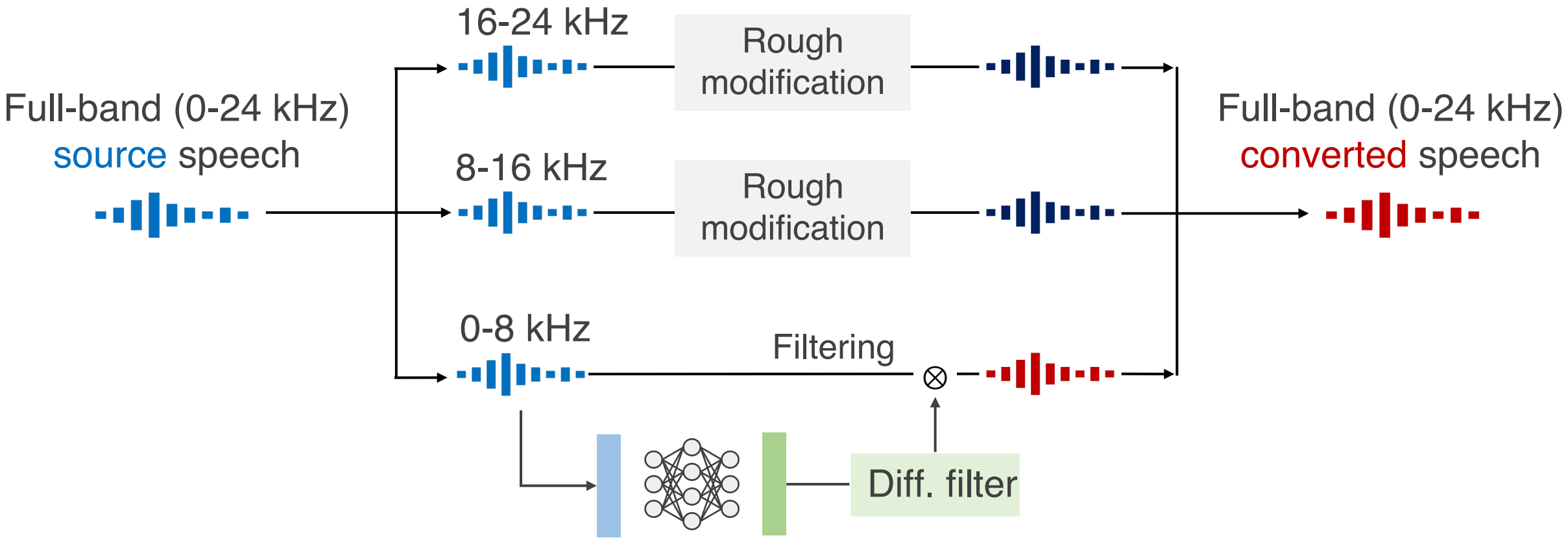
Overview of This Thesis



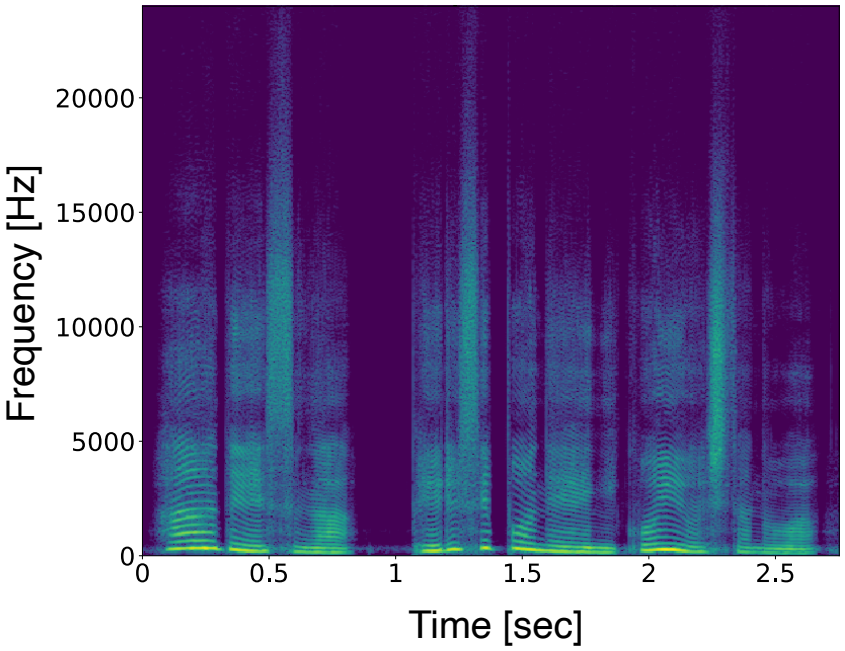
Sub-band Modeling Method

Separately model each frequency band with sub-band multirate processing [Crochiere+83]

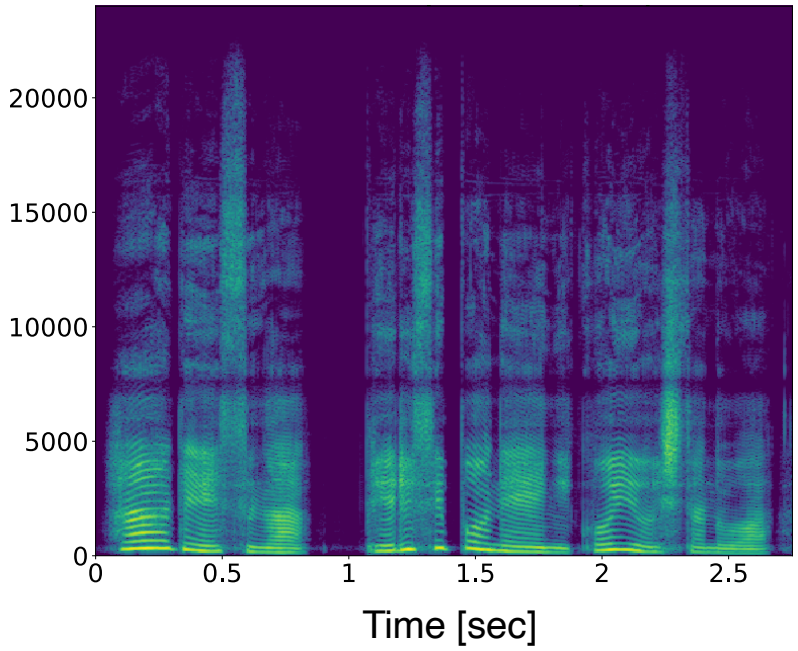
- Only converting lowest-frequency band (0-8 kHz) with DNN
- Roughly modifying or passing through higher-frequency band (8-24 kHz)



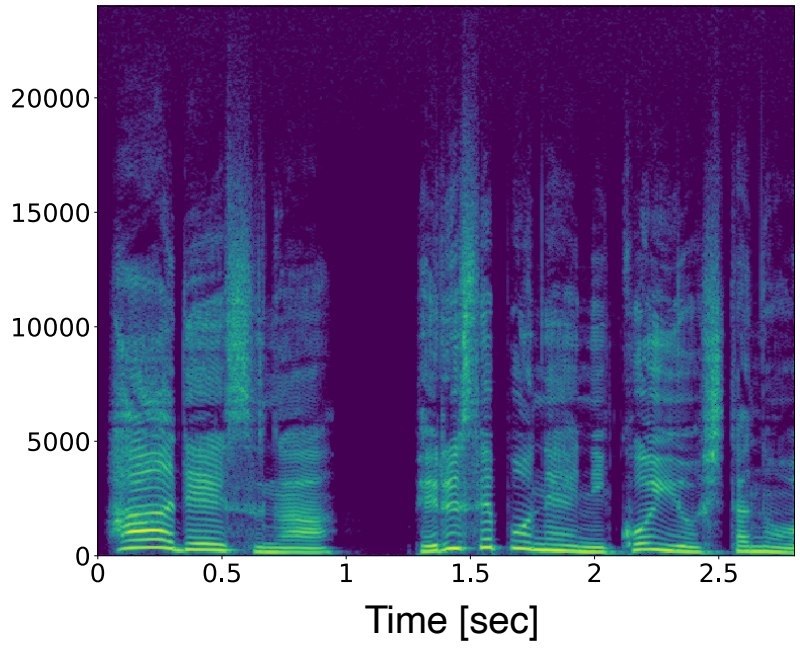
Analysis of Sub-Band Modeling Method



Without sub-band modeling



With sub-band modeling



Target speech

Sub-band modeling method **avoids over-smoothed spectrum** and **leads to fine structures** in whole frequency bands

Evaluation of Proposed Methods

Subjective evaluations for Benchmark* and combination of proposed methods (0-24 kHz)

* Benchmark: Full-band (0-24 kHz) spectral-differential VC without proposed methods

Speaker similarity

Proposed	Score		Benchmark
6 % length	0.516	0.484	100 % length

Short-tap proposed methods ≅ Full-tap benchmark

Speech quality

Proposed	Score		Conventional
6 % length	**0.840**	0.260	100 % length

Short-tap proposed methods >> Full-tap benchmark

Proposed methods can significantly improve speech quality while reducing computational cost in filtering!

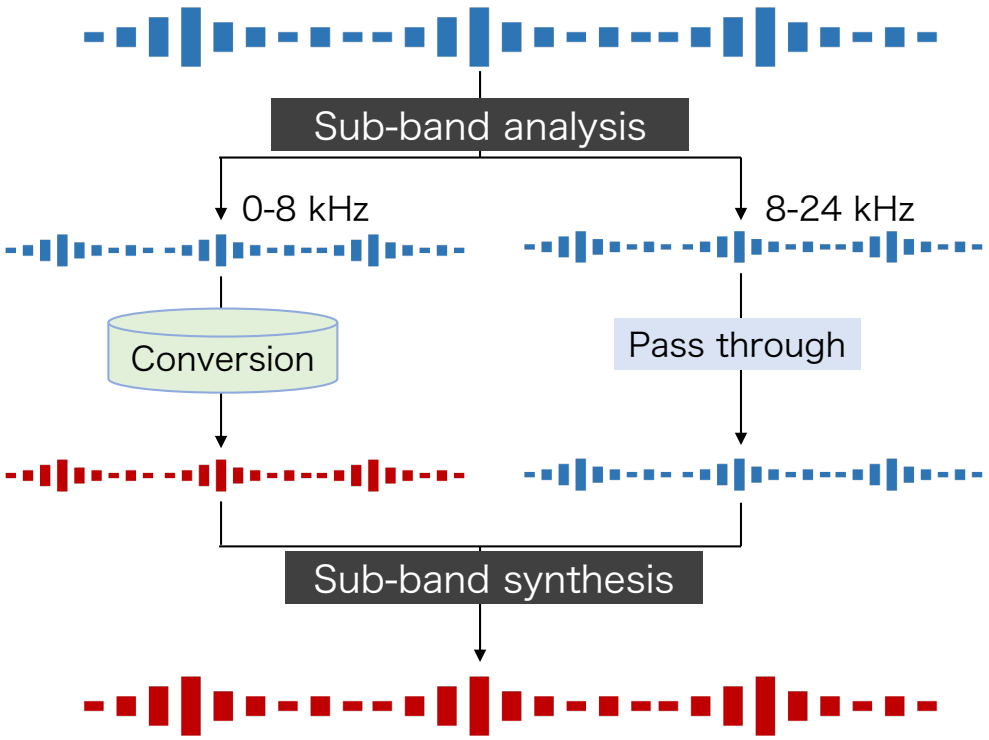
Table of Contents

Implementation of Real-Time VC System

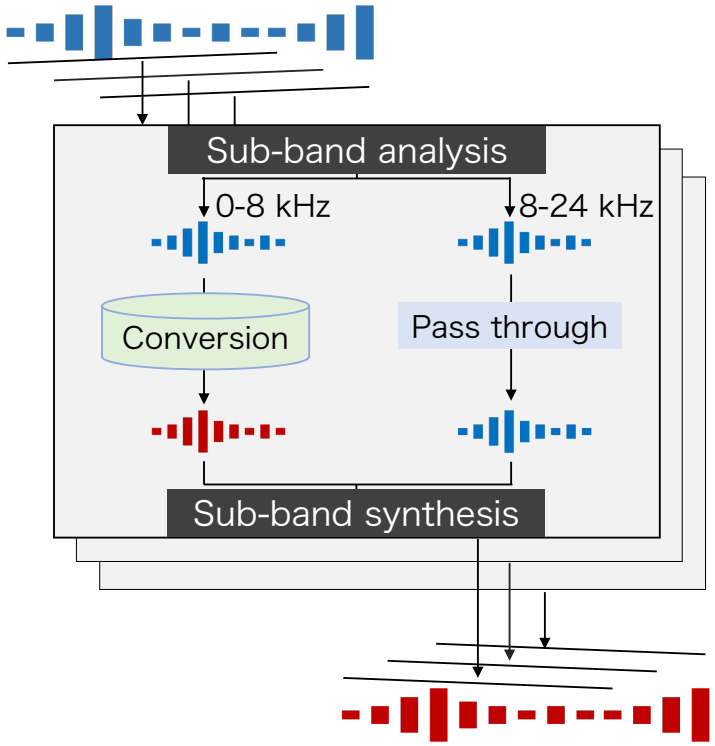
Requirements of real-time VC system for practical applications

- **Streaming** conversion
- **Pitch transformation** for cross-gender VC

Utterance-level VC (Chapter 3)

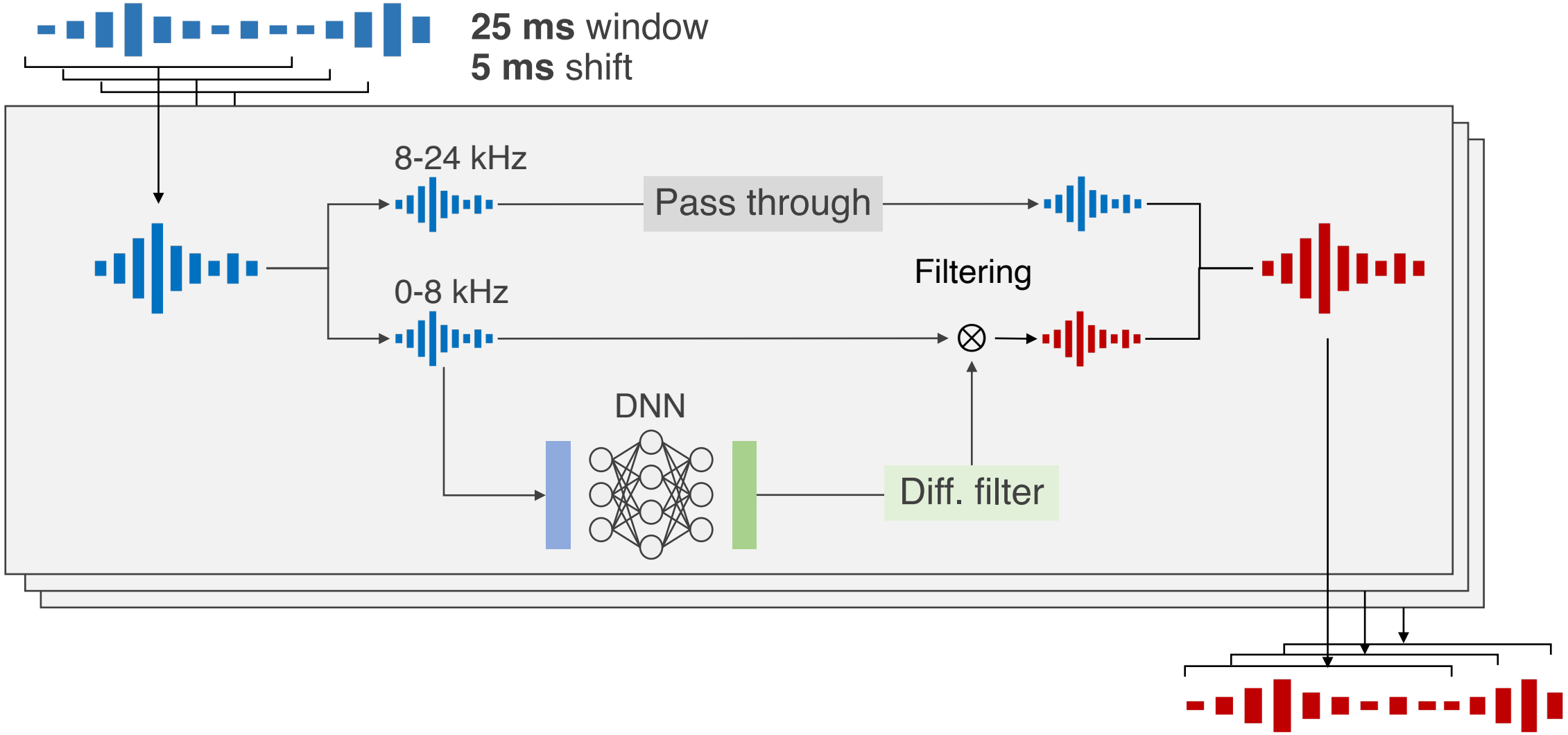


Streaming VC (Chapter 4)



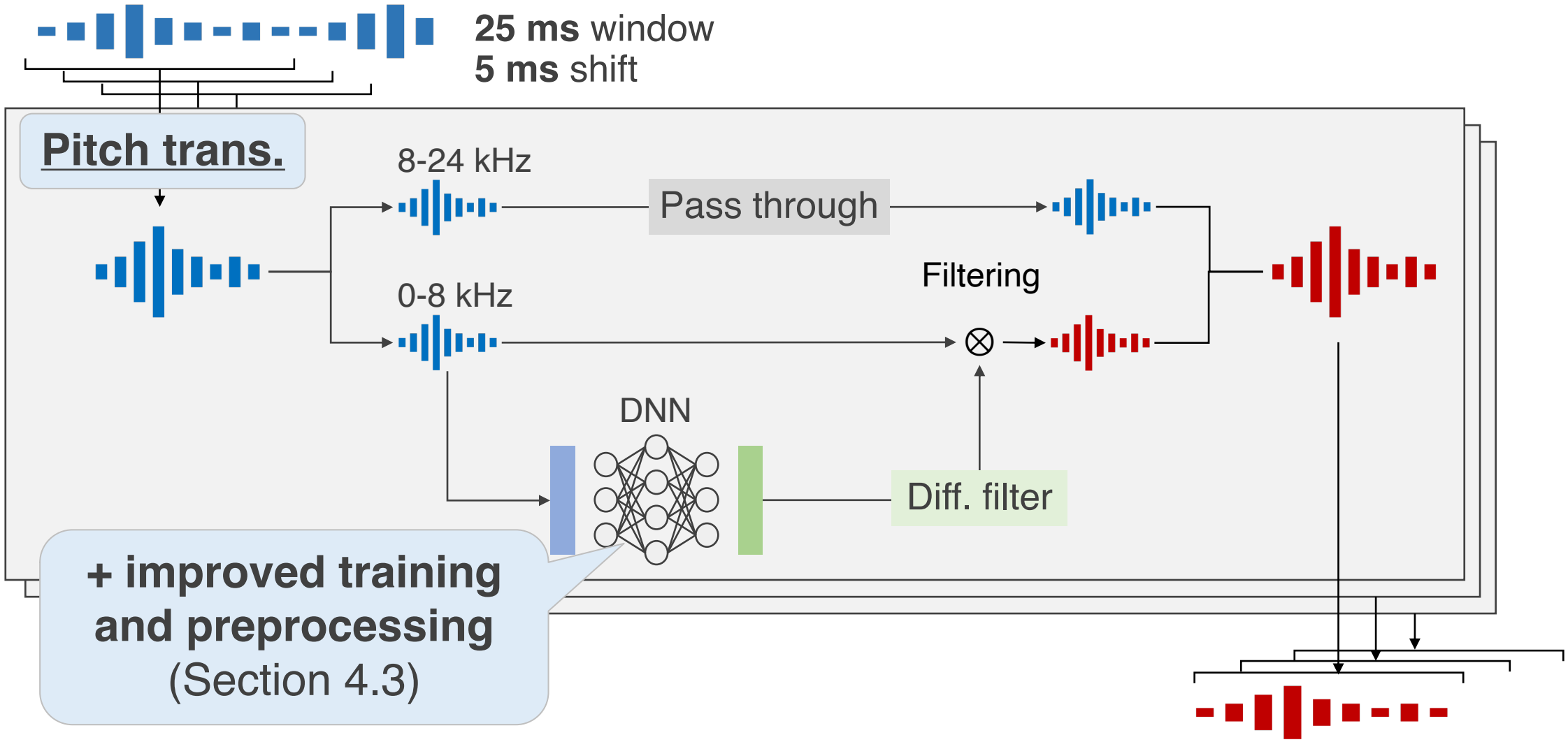
System Overview

Real-time system receives 25 ms frame and processes each frame within 5 ms

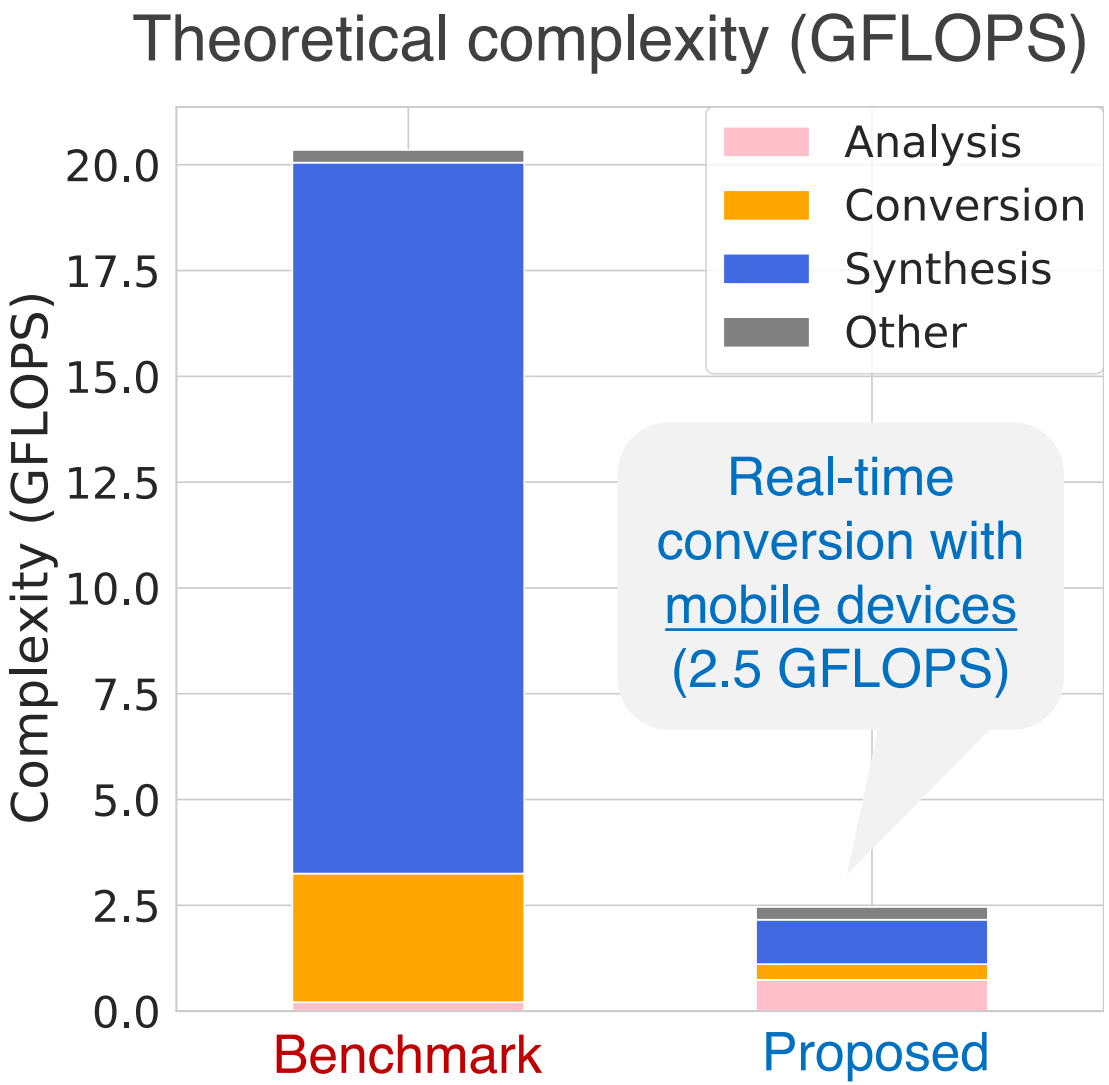


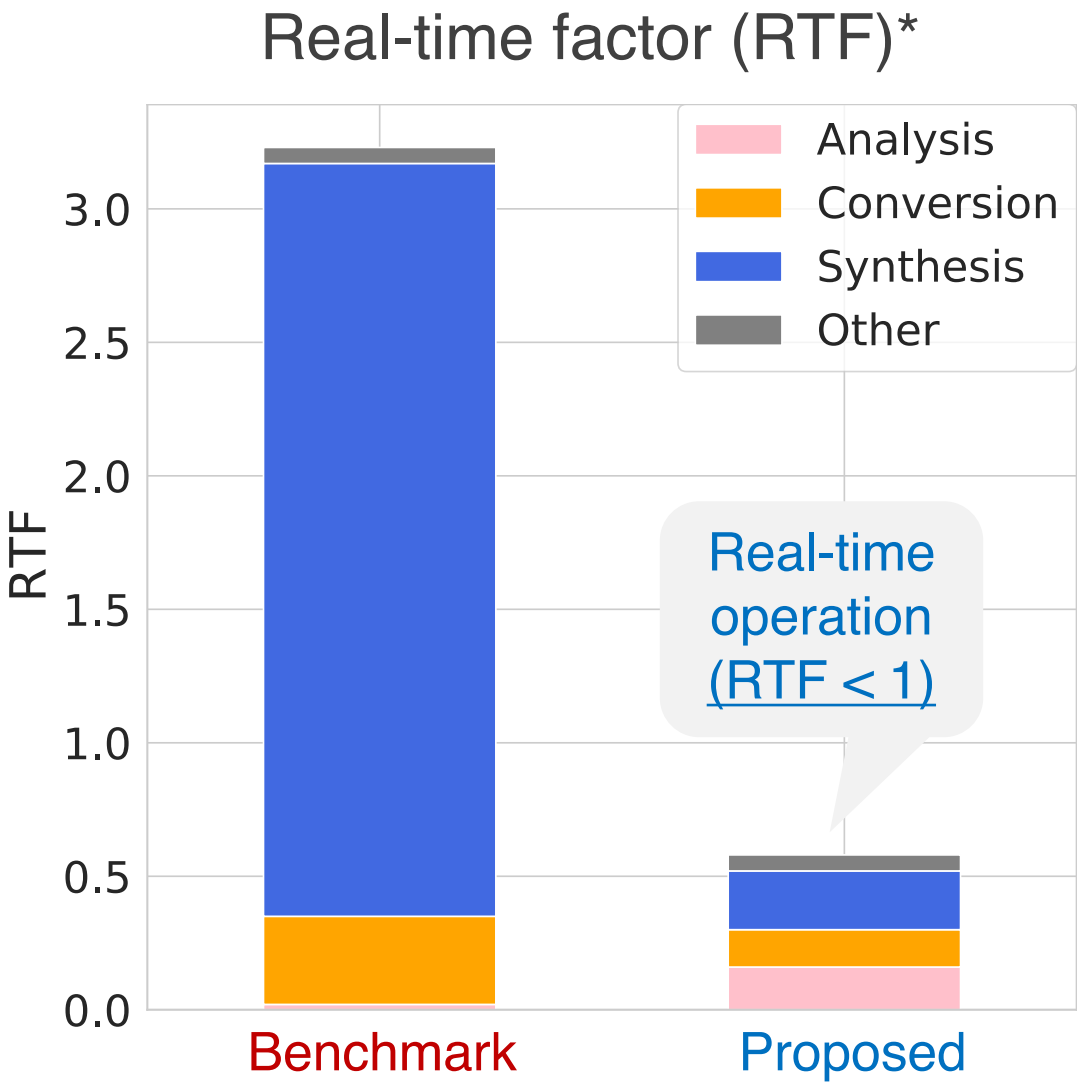
System Overview

System includes **pitch transformation** and improved training and preprocessing



Evaluations on Computational Efficiency

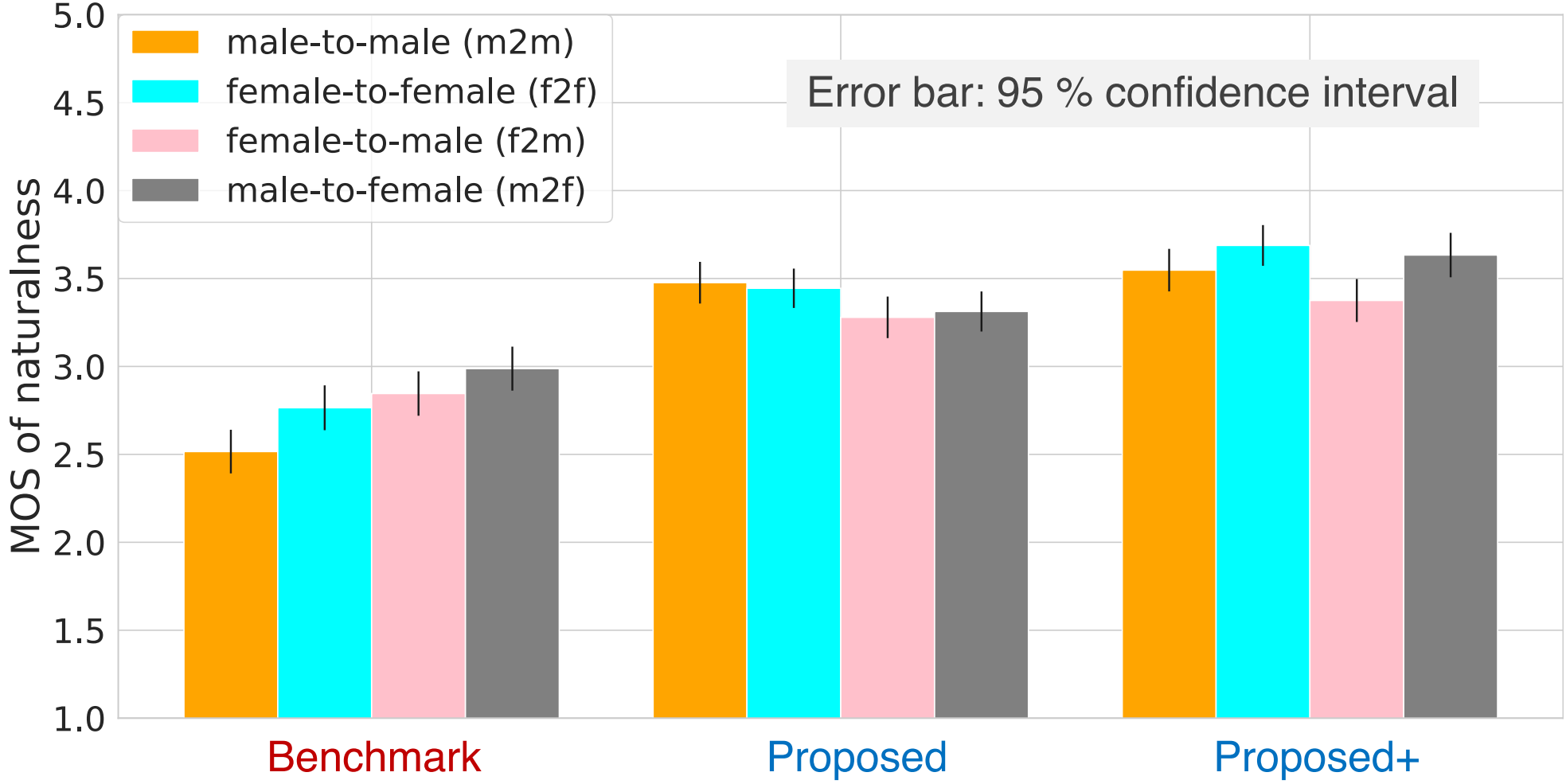




*Processing time (ms) / input waveform length (ms)

Experimental Evaluation on VC Quality

Evaluated naturalness of converted speech with mean opinion score (MOS)



Proposed system achieves significantly higher-quality than benchmark

Comparison with Other DNN-Based Real-Time VC

Subjective evaluations for Proposed+ and other DNN-based VC system [Arakawa+19]

Speaker similarity

Proposed+	Score		(Arakawa+19)
m2m	**0.727**	0.273	m2m
f2f	**0.907**	0.093	f2f
f2m	**0.777**	0.223	f2m
m2f	**0.880**	0.120	m2f

Speech quality

Proposed+	Score		(Arakawa+19)
m2m	**0.977**	0.023	m2m
f2f	**0.967**	0.033	f2f
f2m	**0.960**	0.040	f2m
m2f	**0.967**	0.033	m2f

Spkr	Source	Target	(Arakawa+19)	Proposed+
m2f				

Proposed real-time VC system achieves significantly higher speaker similarity and speech quality than other DNN-based real-time VC system

Publications and Research Activities

Original Journal Papers

[Saeki+, IEEE SPL (submitted)]

[Saeki+, IEICE Trans. (submitted)]

International conferences

[Saeki+, INTERSPEECH20]

[Kimura+, INTERSPEECH20]

[Saeki+, ICASSP20]

Domestic conferences

6 publications

Patents

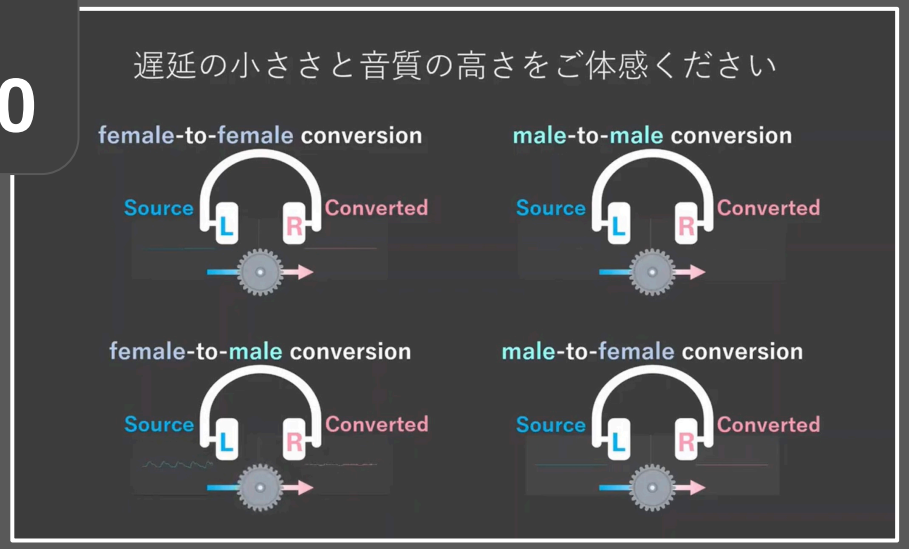
3 submissions

Exhibitions

Sainokuni Buisiness Arena 2021 ONLINE

CEATEC 2020 ONLINE

Pictures of CEATEC 2020



Summary and Conclusion

Purpose

Developing real-time, full-band, high-quality VC system

Contributions

Data-driven phase estimation method for reducing computational cost in filtering
Sub-band modeling method for high-quality and efficient full-band VC
Implementation of real-time full-band VC system

Results

Attaining 2.5 GFLOPS and RTF < 1 for converting full-band speech
Achieving high-quality output speech with 3.6 / 5.0 MOS of naturalness

Future works

Improving feature analysis for better speaker similarity