AUDIO CODING BASED ON SPECTRAL RECOVERY BY CONVOLUTIONAL NEURAL NETWORK

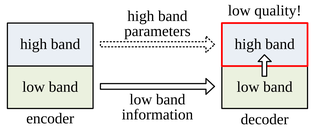
Seong-Hyeon Shin[1], *Seung Kwon Beack*[2], *Taejin Lee*[2] *and Hochong Park*[1]

[1]Kwangwoon University, Seoul, Korea

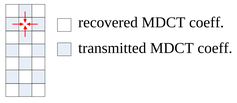
[2]Electronics and Telecommunications Research Institute, Daejeon, Korea

Introduction

- New audio coding method based on 2D spectral recovery by convolutional neural network
  - Objective : better coding performance than current state-of-the-art codecs
- Novelty of 2D spectral recovery
  - Conventional method : spectral recovery on a block basis
    - High-band recovery based on low band with optional high-band parameters
  - Proposed method : recovery of individual spectral coefficients based on neighboring information
    - Arrangement of transmitted data and recovered data in 2D check pattern



< Conventional spectral recovery : high-band coding >

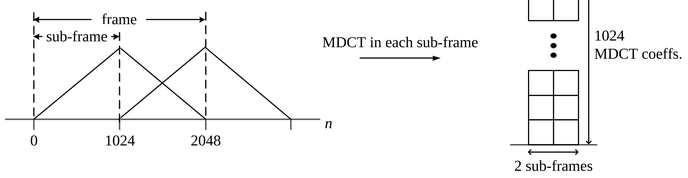


< Proposed method >

- The proposed method can be integrated with USAC (state-of-the-art audio codec).
- Performance
  - Significantly better than USAC frequency-domain mode at 39.4 kbps

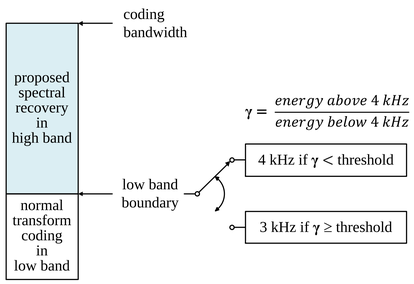
Proposed method : Settings

- Sub-frame-based transform
  - 1024 × 2 MDCT coefficients for each frame of 2048 samples



< Sub-frame-based transform >

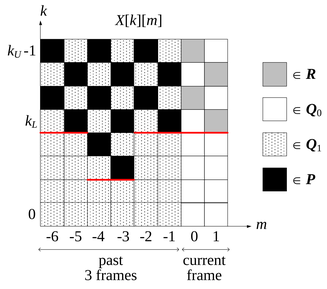
- Recovering only MDCT magnitudes
  - Ensuring high correlation among data
  - Signs are transmitted or randomly assigned.
    - Transmitted signs are selected based on their estimated importance.
- Recovering only high-band spectral data
  - Spectral recovery in low band yields unacceptable performance.



< Band boundary for recovery>

Proposed method : Structure of recovery

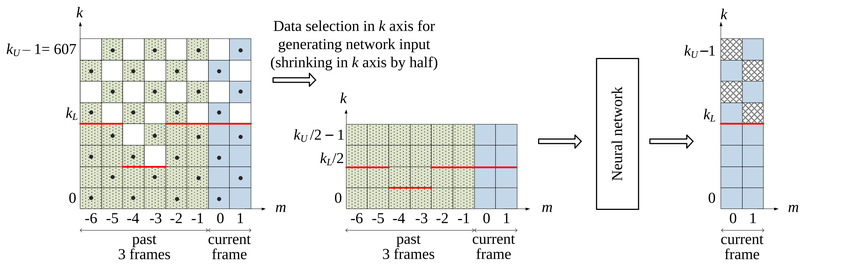
- Using previous frames to recover current frame
- Overall structure of 2D MDCT coefficients : k_U × 8 MDCT magnitudes
  - Current frame and past 3 frames



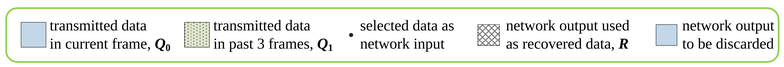
R : non-transmitted data in current frame
Q**0** : transmitted data in current frame
Q**1** : transmitted data in past frame
P : non-transmitted data in past frame
k_U : coding bandwidth
k_L : low-band boundary

< Overall structure of 2D MDCT coefficients for recovery >

- Reconstruction of MDCT magnitudes using CNN
  - k_U is set to 608 (0 ~ 14.25 kHz at 48 kHz sampling rate).



< Transmitted data > < Network input data > < Network output data >



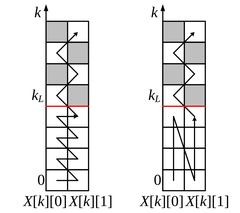
- Convolutional neural network (CNN)
  - Activation function
    - Hidden layer : ReLU
    - Output layer : tanh
  - Optimizer : Adam
  - Cost function : L1 cost function

< The structure of convolutional neural network >

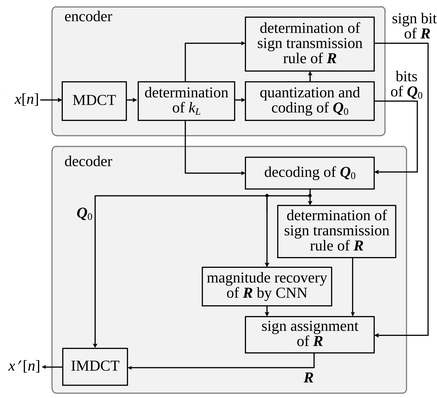
	layer	output shape	no. of filters	filter size	stride
encoder layer	1	[152, 8]	32	[5, 5]	[2, 1]
	2	[152, 4]	64	[5, 5]	[1, 2]
	3	[76, 4]	128	[5, 5]	[2, 1]
	4	[76, 2]	256	[5, 3]	[1, 2]
	5	[38, 2]	512	[5, 3]	[2, 1]
decoder layer	1	[76, 2]	256	[5, 3]	[2, 1]
	2	[152, 2]	128	[5, 3]	[2, 1]
	3	[304, 2]	64	[5, 3]	[2, 1]
	4	[608, 2]	32	[5, 3]	[2, 1]
	5	[608, 2]	1	[5, 3]	[1, 1]

Proposed method : Overall operation

- Quantization and coding of $\boldsymbol{Q}_0$
  - Using quantizer and arithmetic coder of the USAC
  - Converting 2D MDCT coefficients into 1D data using two scanning patterns for entropy coding



< Two scanning patterns >



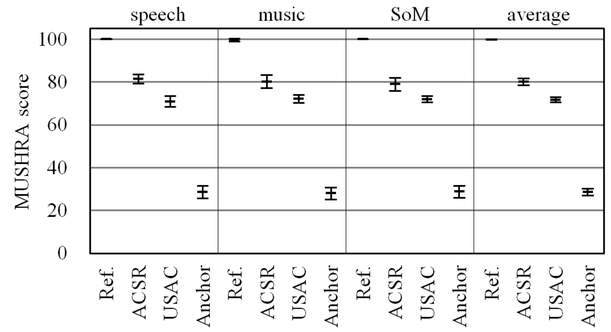
< Block diagram of overall operation >

Evaluation

- Database
  - Train/validation data : Beethoven sonata, VCTK dataset, RWC music database (total 57 hours)
  - Test data : audio sequence offered by MPEG audio group (12 items)
    - 3 categories : music, speech, speech over music (SoM)
- Using quantizer of USAC at 48 kbps and long window only
- Bit rate comparison with USAC

		Bit rate (kbps)		Reduction rate (%)
		proposed	USAC	
MDCT coeffs.	$\boldsymbol{Q}_0$	32.4	42.6	20.4
	Signs in ***R***	1.5		
Side information		5.5	5.4	-
Total		39.4	48.1	18.1

- Subjective performance evaluation by MUSHRA : 7 subjects participated
  - Comparing the proposed method with USAC at 39.4 kbps
    - Significantly better performance at the same bit rate



Conclusion

- The proposed method is a new audio coding method based on 2D spectral recovery by a convolutional neural network.
- The proposed method integrated with USAC provides higher coding performance than USAC.