LOW-BITRATE REDUNDANCY CODING OF SPEECH FOR PACKET LOSS CONCEALMENT IN TELECONFERENCING

Marcin Ciołek, Michał Sulewski, Mihailo Kolundzija, Rafał Pilarczyk, Raul Casas, Samer Hijazi
Cisco Systems, Inc., Babblelabs team
marciole@cisco.com, msulewsk@cisco.com

Abstract

The proposed system significantly improves quality and intelligibility under packet loss in video-conferencing applications. We introduce a 16 kHz novel neural codec for low-bitrate speech coding at 6 kbit/s, with long 1 kbit/s redundancy, that also enhances speech by suppressing noise and reverberation. Transmitting large amounts of redundant information allows for speech reconstruction on the receiver side during severe packet loss. [Paper number: 7175]

1. Webex conferencing platform

I. Neural codec features

The Webex neural audio codec brings a wide variety of features:
A) removes background noise
B) reduces the amount of reverberation
C) uses low-bitrate speech coding at 6 kbit/s with long 1 kbit/s redundancy
D) reconstructs speech on the receiver side during packet loss

II. Neural codec components

Our neural audio codec is an autoencoder trained jointly with a vector quantizer:
A) **Encoder** – transforms input audio into embeddings
B) **Vector quantizer** – quantizes the embeddings
C) **Decoder** – reconstructs output audio using quantized embeddings as input

III. Problem statement

1) Speech intelligibility can be reduced during long packet losses at the receiver side
2) Speech reconstruction based on redundant information has a high bitrate cost

IV. Demo goals

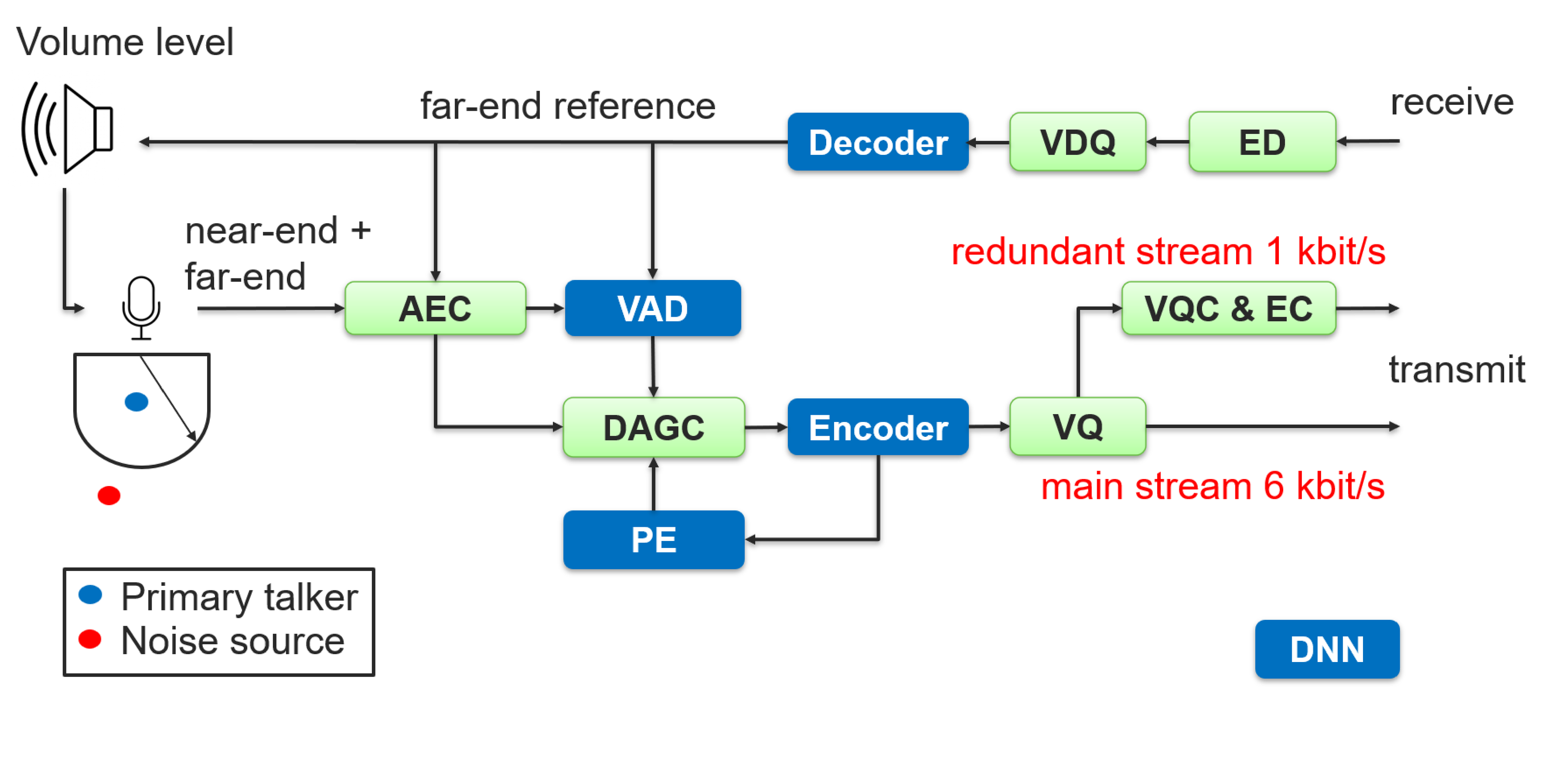
Compression of redundant information can help with:
1) Increasing transmission redundancy and improving speech quality during severe packet loss
2) Lowering the bitrate cost

	Bits per 10 ms frame	Without Compression	Lossy Compression
Speech		**Bitrate**	
Low quality	10	1 kbit/s	~0.45 kbit/s
High quality	60	6 kbit/s	6 kbit/s
Total bitrate		56 kbit/s	28.5 kbit/s

Example packet structure:

20 ms	1 sec of past audio
High quality	Low quality

2. System diagram



AEC – Acoustic Echo Cancellation
VAD – Voice Activity Detection
DAGC – Digital Adaptive Gain Control
PE – Power Estimator
VQ – Vector Quantizer
VDQ – Vector Dequantizer
VQC – Vector Quantizer Correction
EC/ED – Entropy Coding/Entropy Decoding

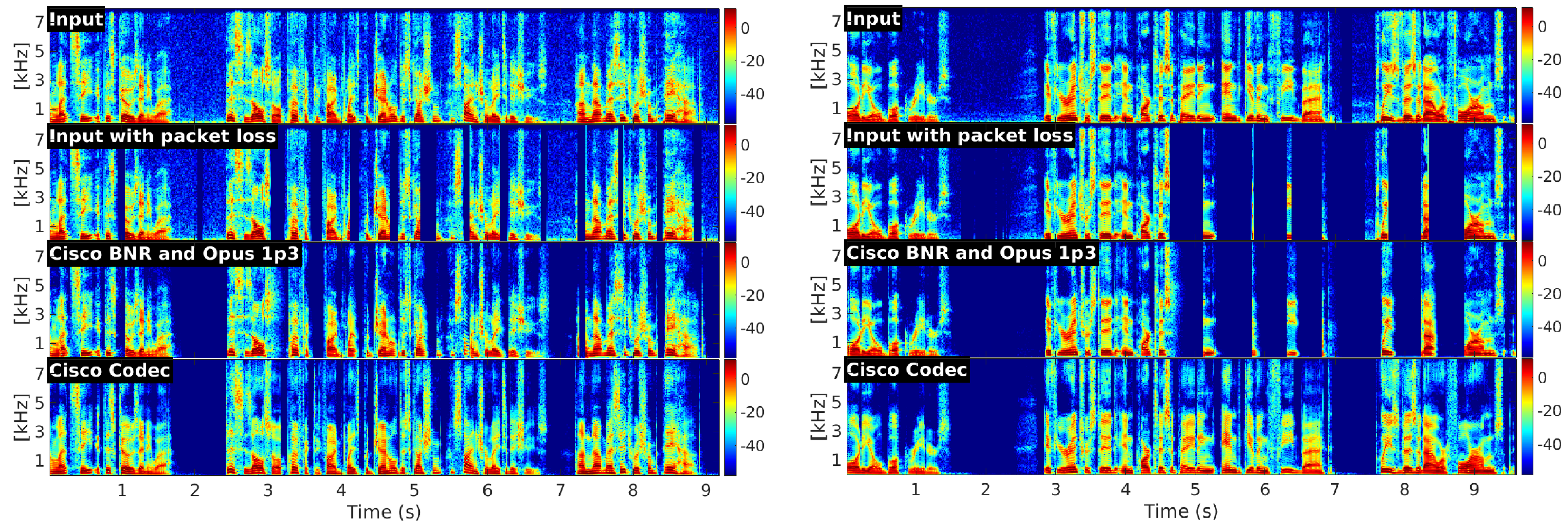
4. Demo scenarios

I. Medium packet loss

- Packet length = 20 ms
- Max packet burst length = 220 ms
- Average packet burst length = 117 ms
- Packets lost = 19 %

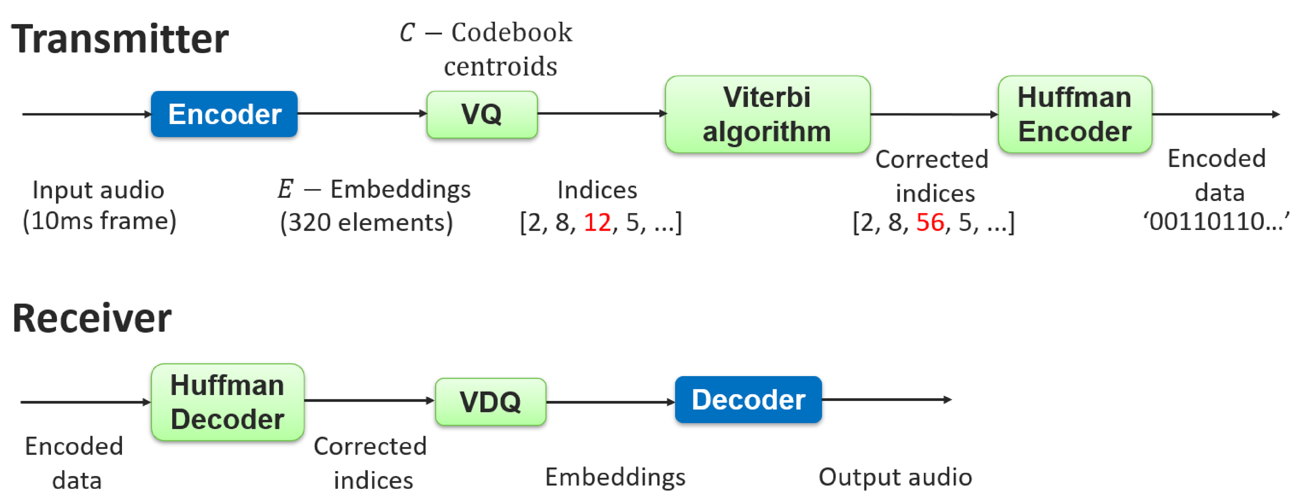
II. Severe packet loss

- Packet length = 20 ms
- Max packet burst length = 520 ms
- Average packet burst length = 310 ms
- Packets lost = 42 %



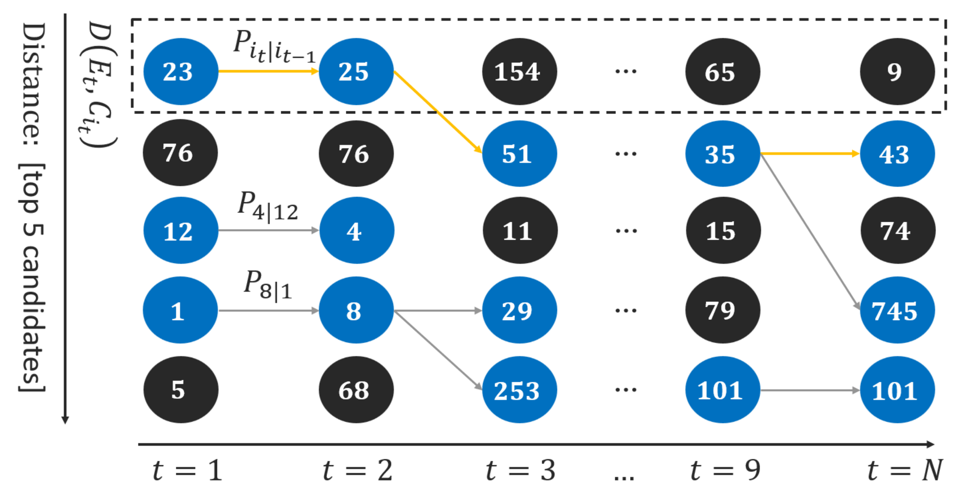
3. Vector Quantizer Correction & Entropy Coding

I. Low-bitrate speech coding



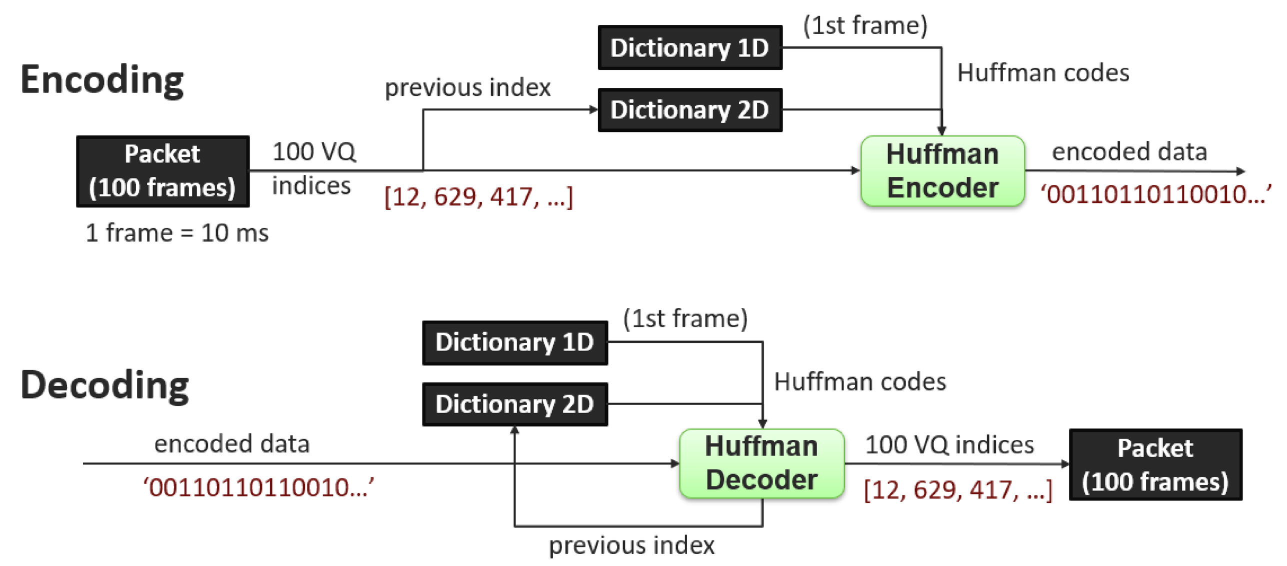
II. Viterbi algorithm

$$x_t^* = \arg\min_{x_t} L(x_t) \quad \text{where } x_t = [i_1, \dots, t_N] \quad L(x_t) = \sum_{t=1}^{N} \alpha D(E_t, C_{i_t}) - \sum_{t=2}^{N} \log P_{i_t|i_{t-1}}$$



III. Huffman coding

Dictionary 1D: 1 dictionary of 1024 indices
Dictionary 2D: 1024 dictionaries of all possible transitions



IV. Lossy compression with Viterbi and Huffman

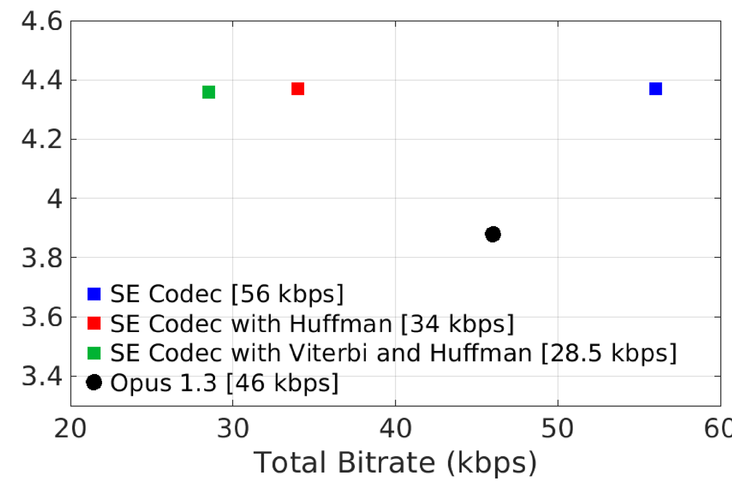
1200 hours training data from LibriVox
30 minutes test data from LibriSpeech

α	Compression rate [%]	Replaced ind. [%]	Avg WMOS	Avg POLQA
α = 100	48%	2%	2.50	1.85
α = 5	56%	22%	2.48	1.64
α = 4	57%	26%	2.47	1.60
α = 3	58%	30%	2.47	1.53

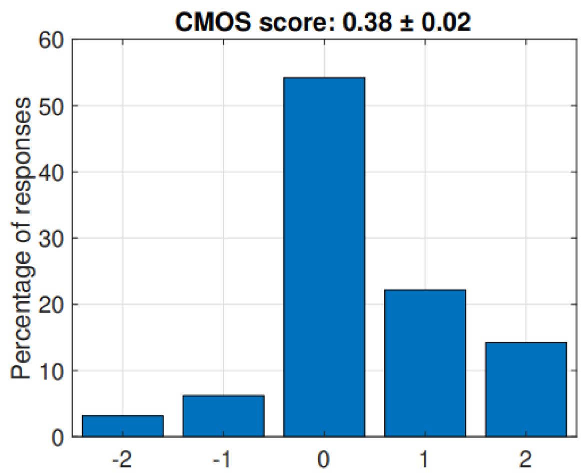
5. System results: PLC Challenge 2023

I. PLC-MOS scores

Opus 1.3 DRED vs. SE Codec RED, both with 1s redundancy. Total bitrate = main stream + redundant stream.



II. Subjective listening test results



Opus DRED at 46 kbps (**A**) vs. SE Codec RED at 28.5 kbps (**B**), both with 1 s long redundancy.
- −2: **B** is much worse than **A**
- −1: **B** is slightly worse than **A**
- 0: there is no difference
- 1: **B** is slightly better than **A**
- 2: **B** is much better than **A**

6. Future work

Improving speech quality at low-bitrate.
Lowering the cost in bitrate and complexity.