

Fig. 1a

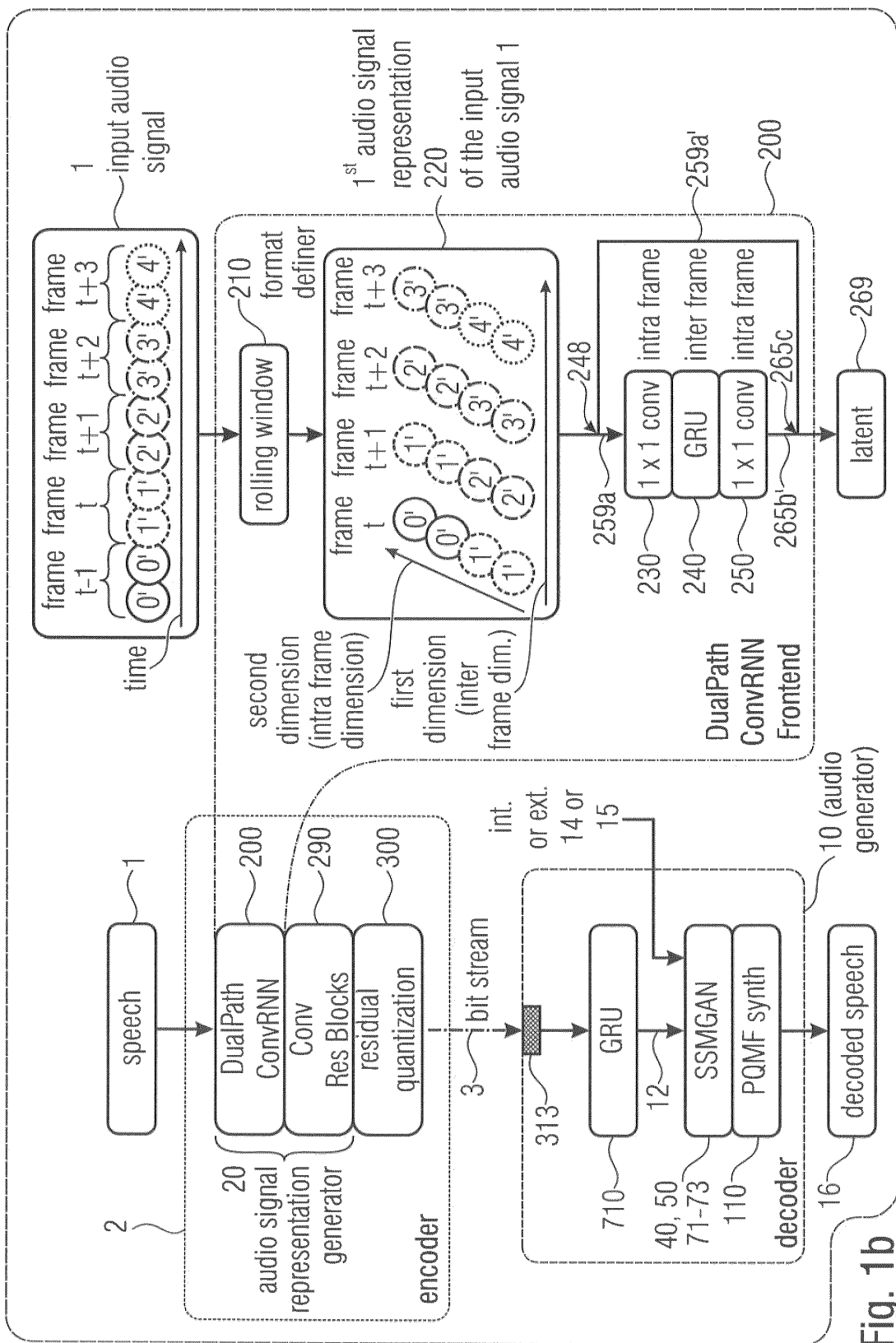


Fig. 1b

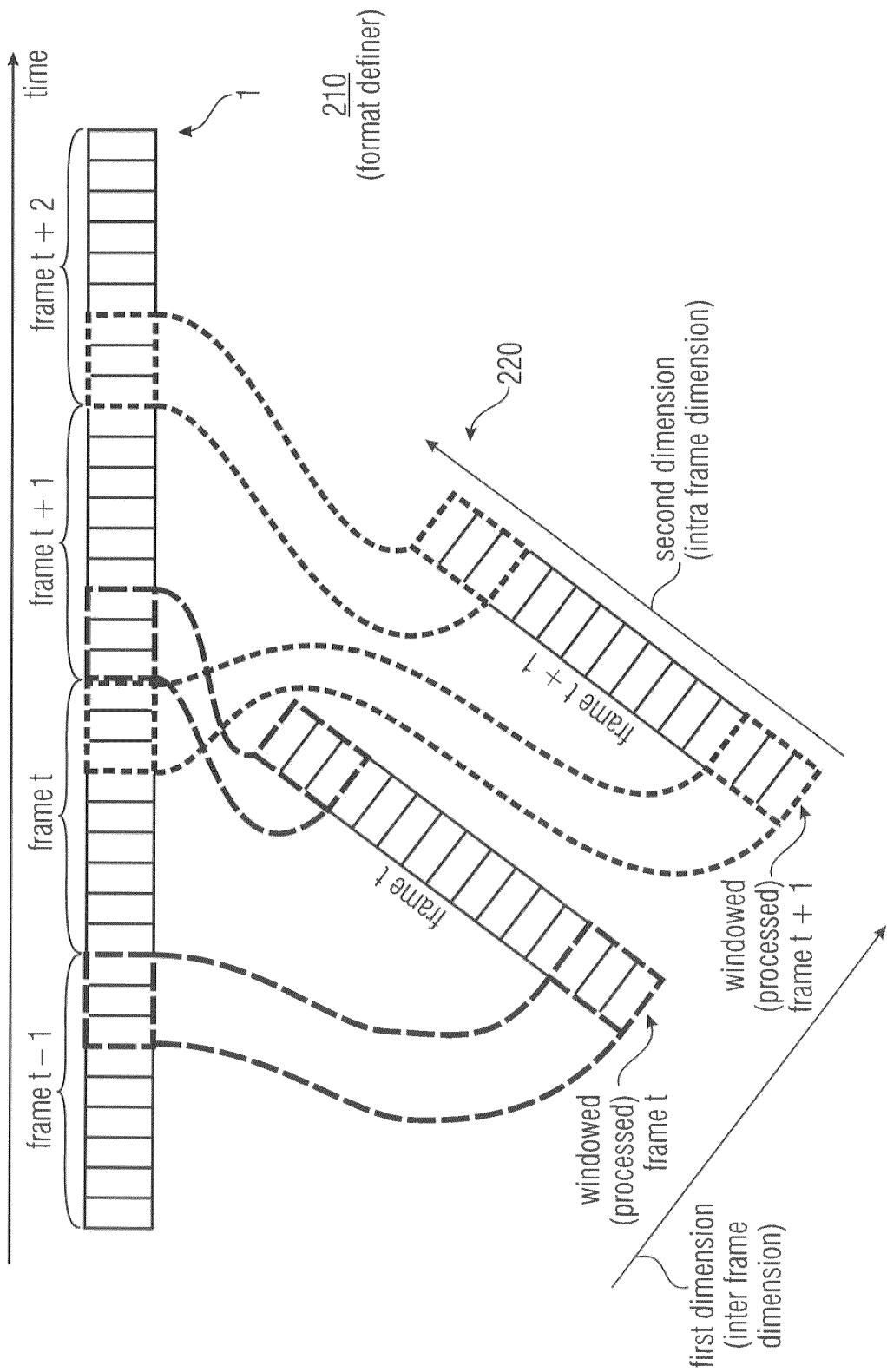
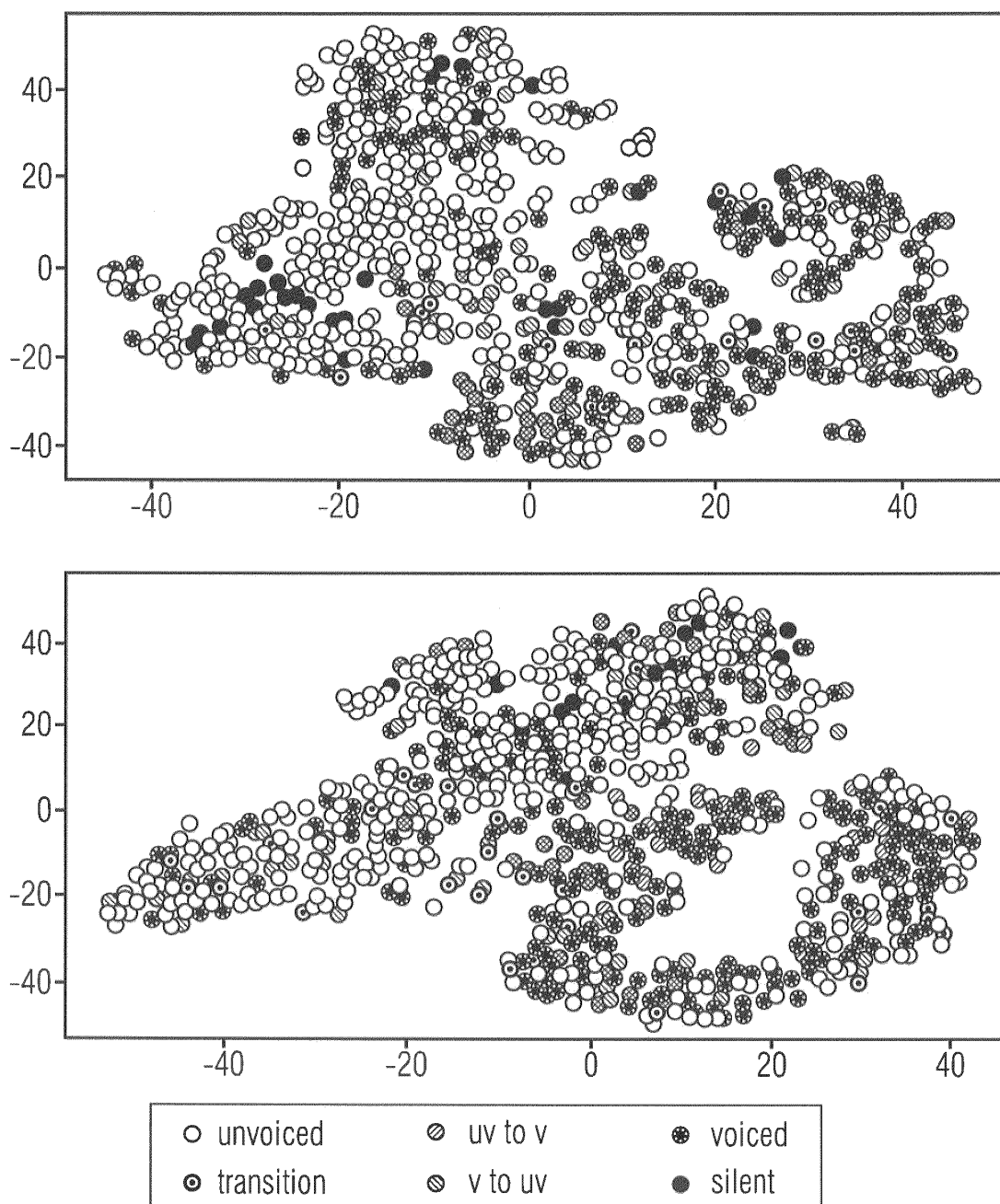


Fig. 1c

4/17

t-SNE projection of latent frames for voicing information

Fig. 2a
(Part 1)

5/17

t-SNE projection of latent frames for voicing information

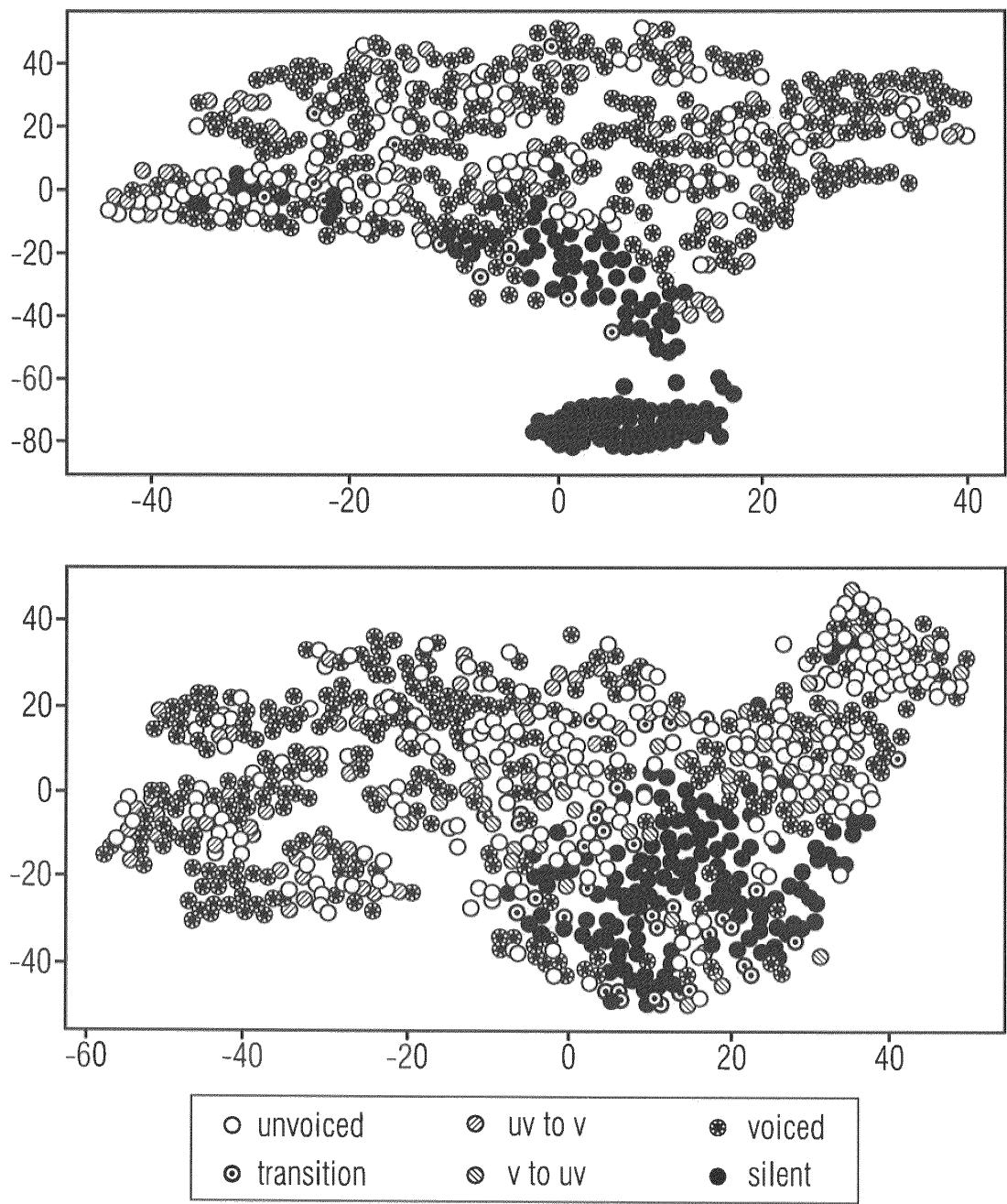


Fig. 2a
(Part 2)

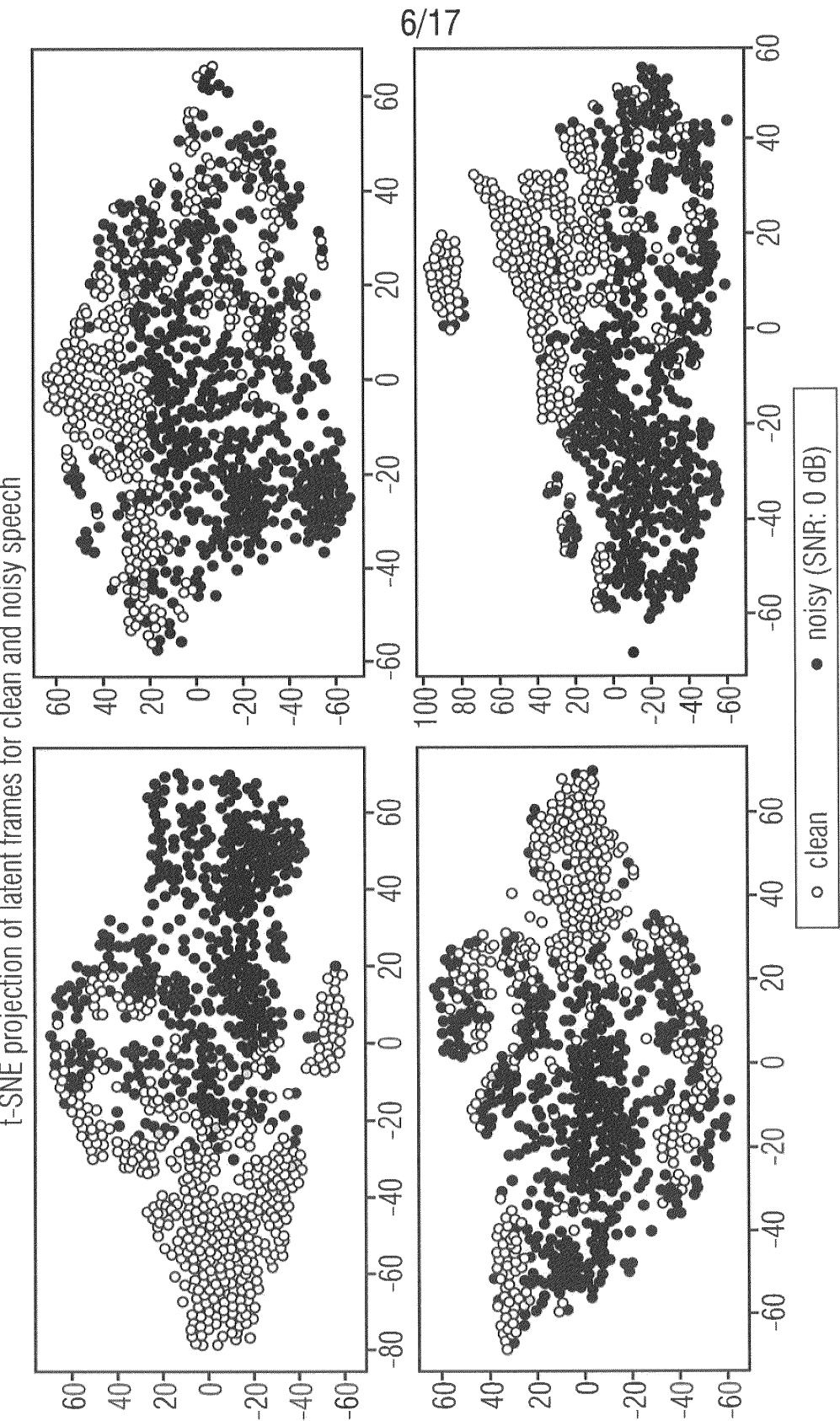


Fig. 2b

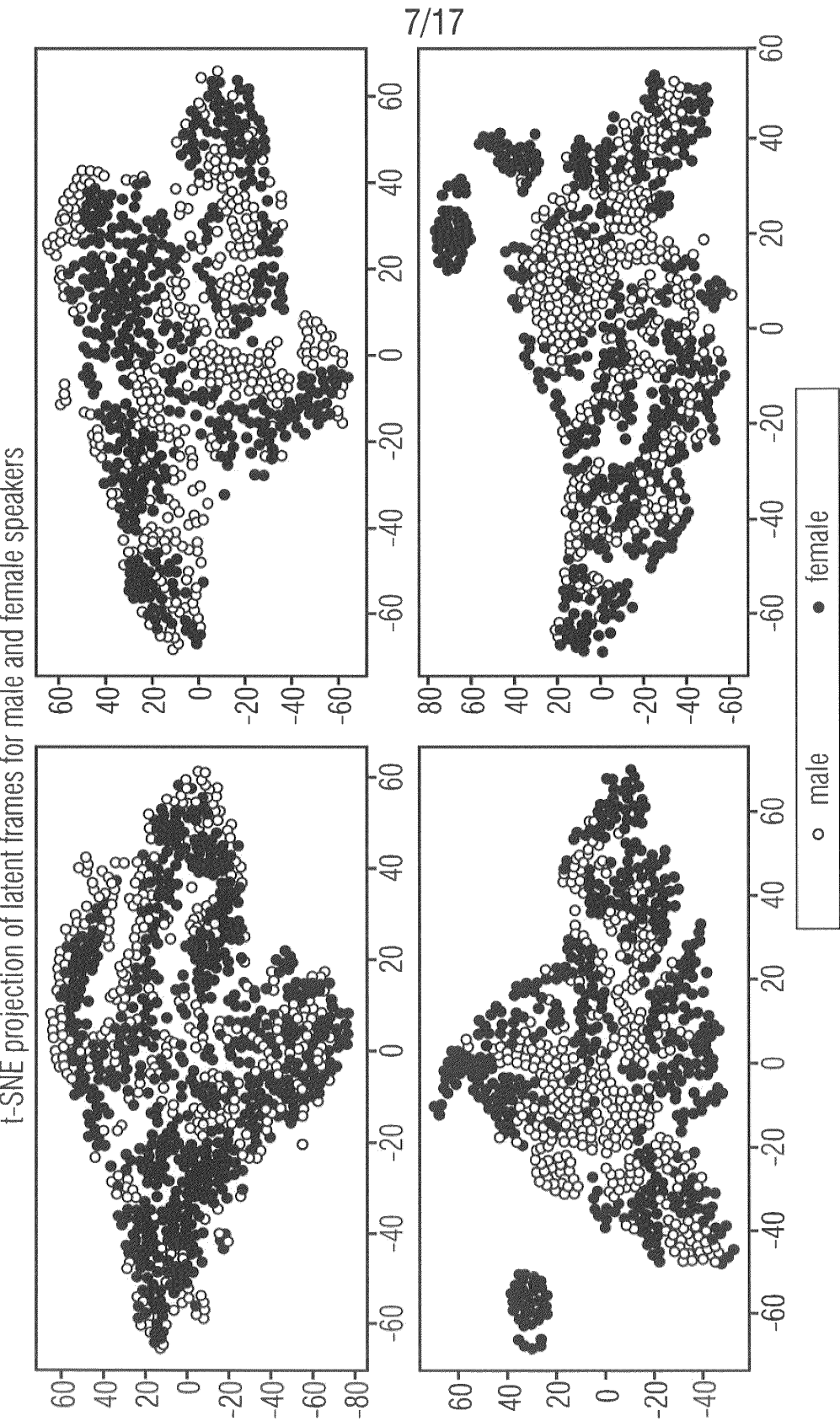


Fig. 2c

8/17

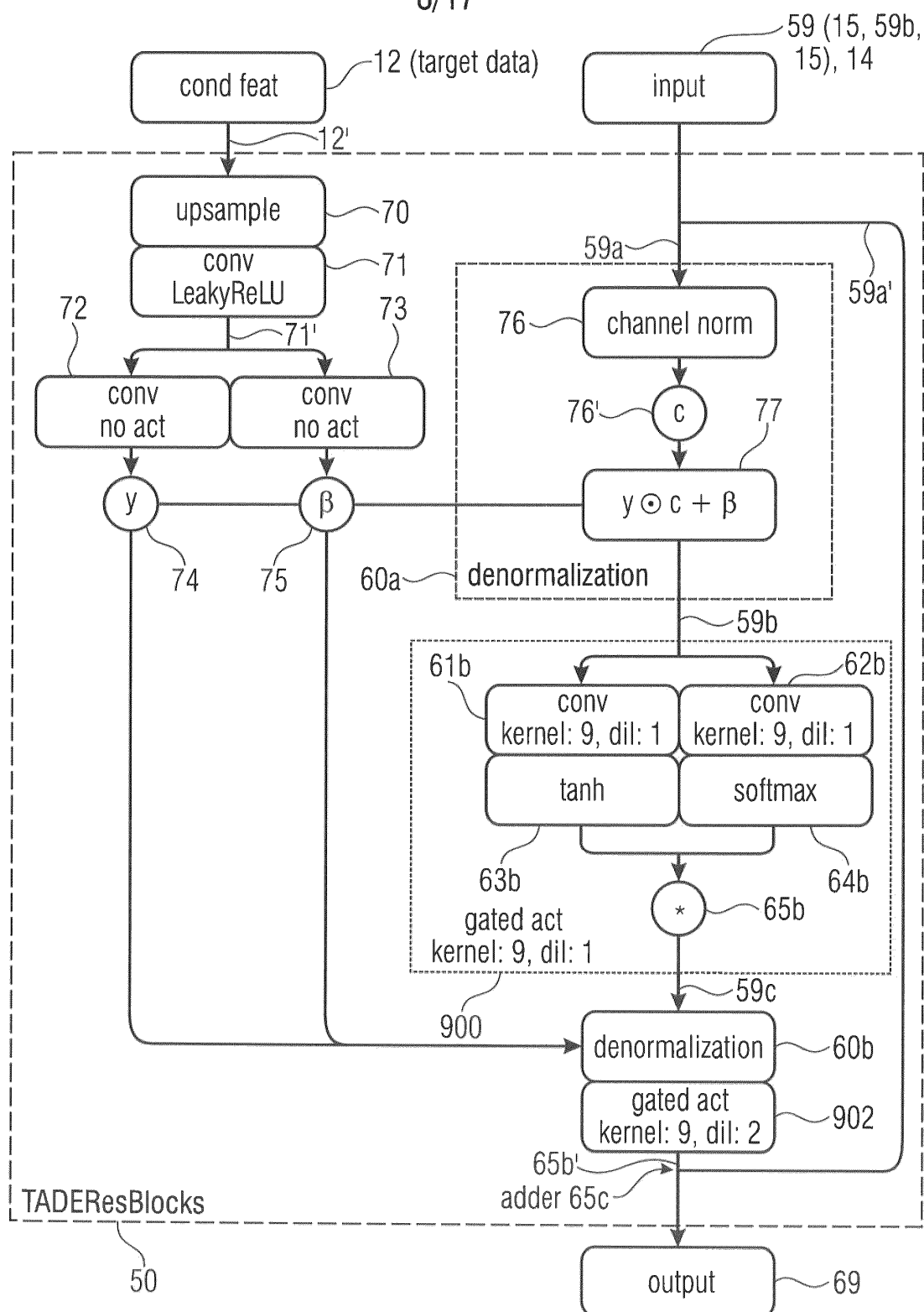


Fig. 3

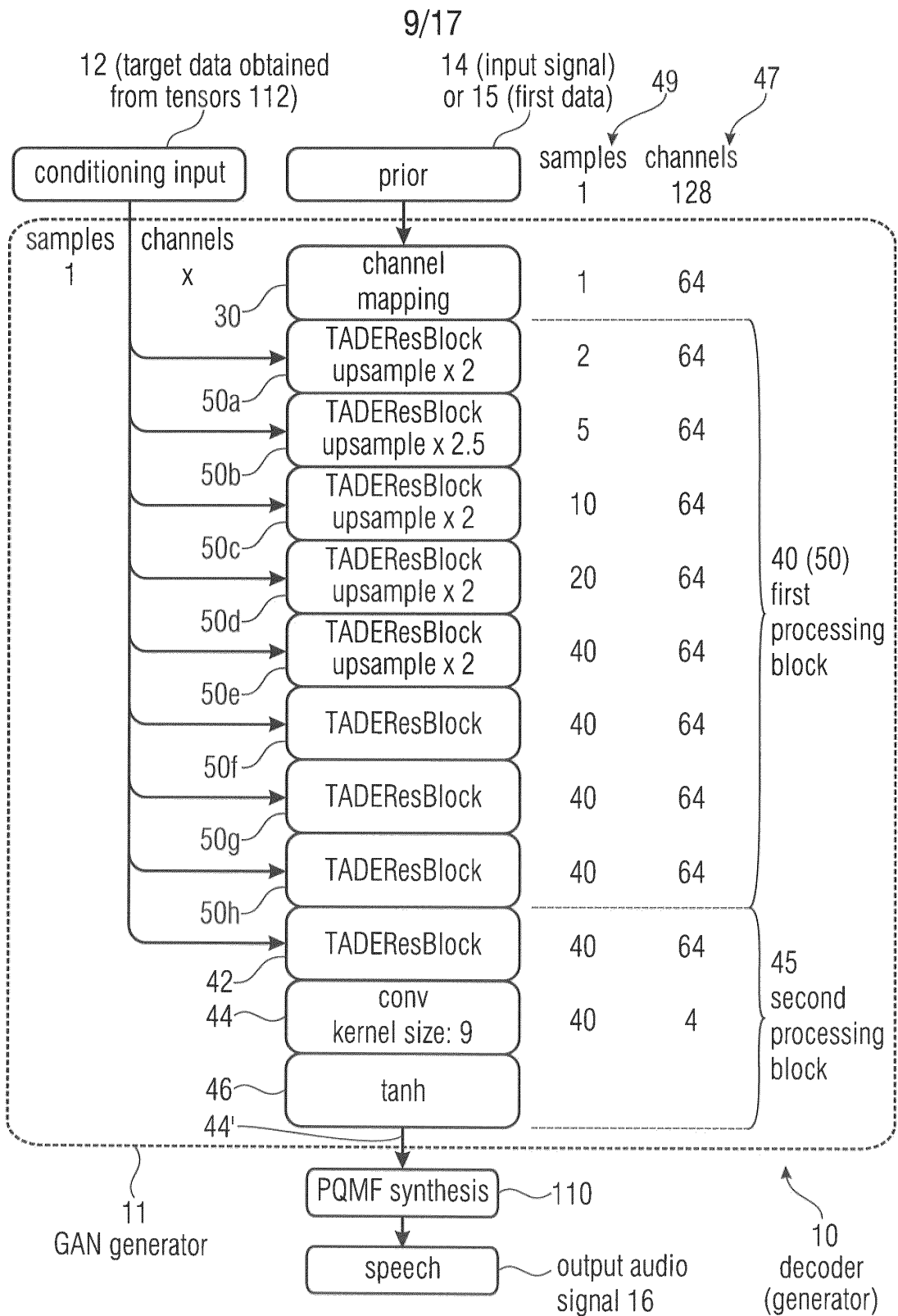


Fig. 4

10/17

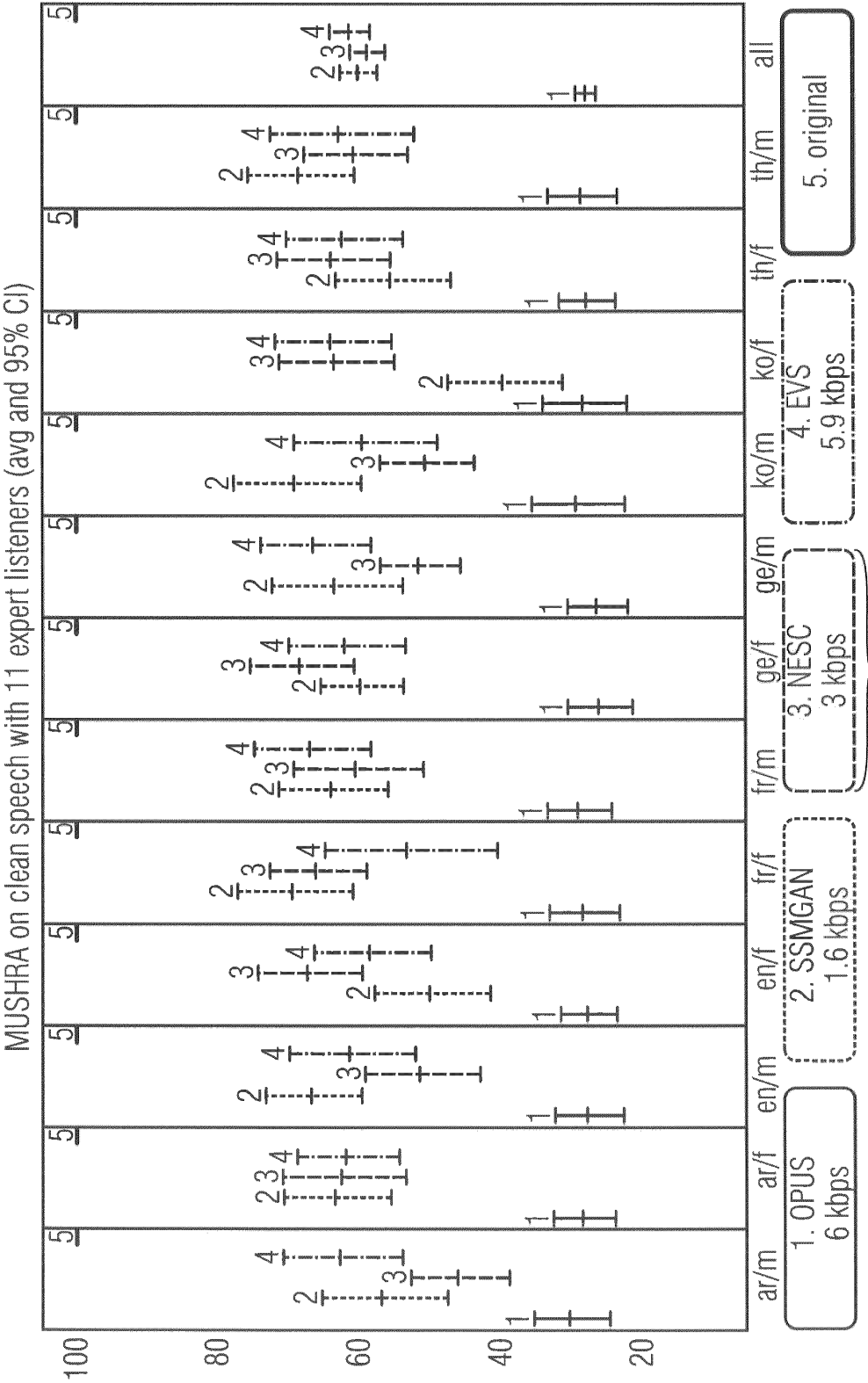


Fig. 5

11/17

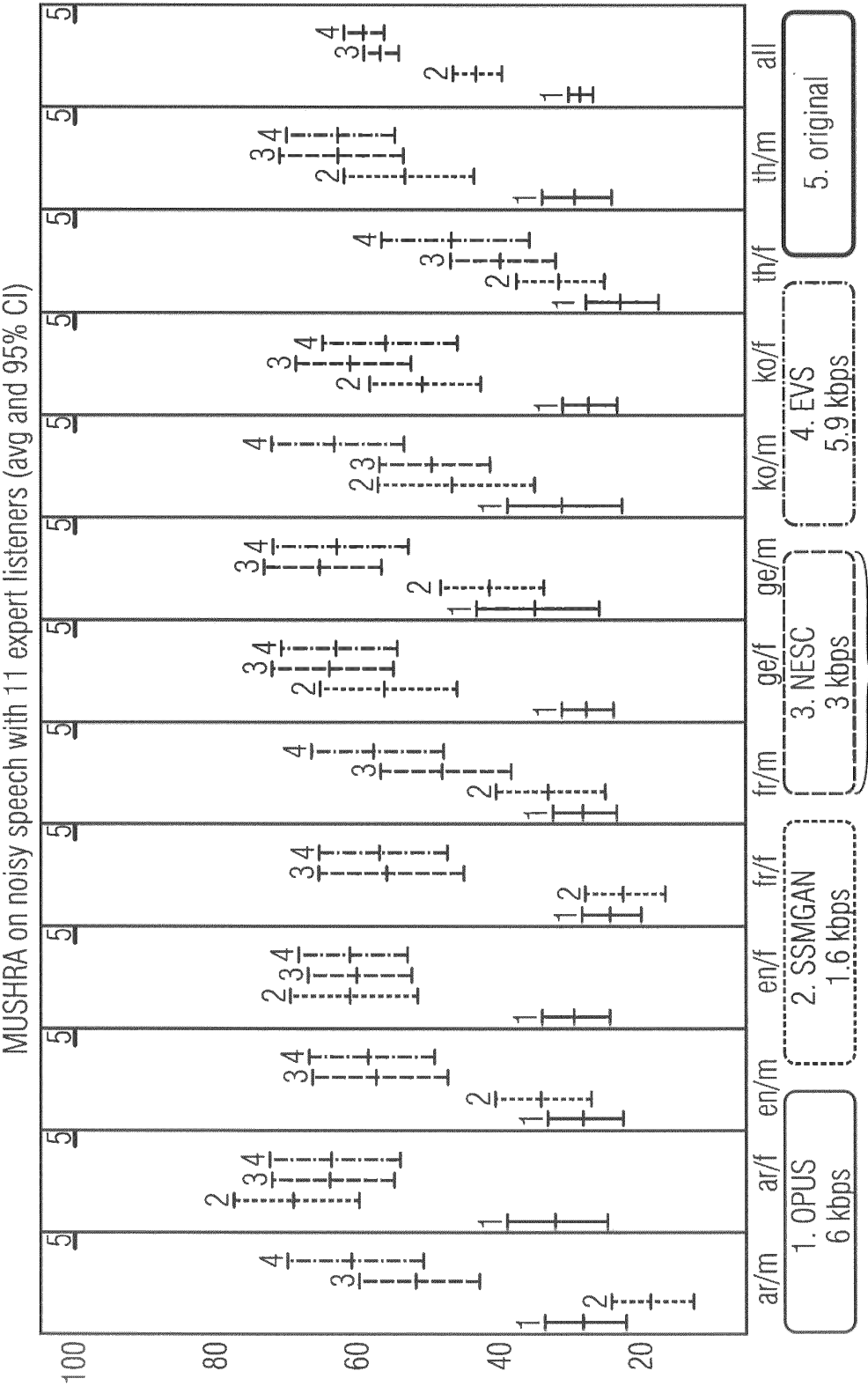


Fig. 6

12/17

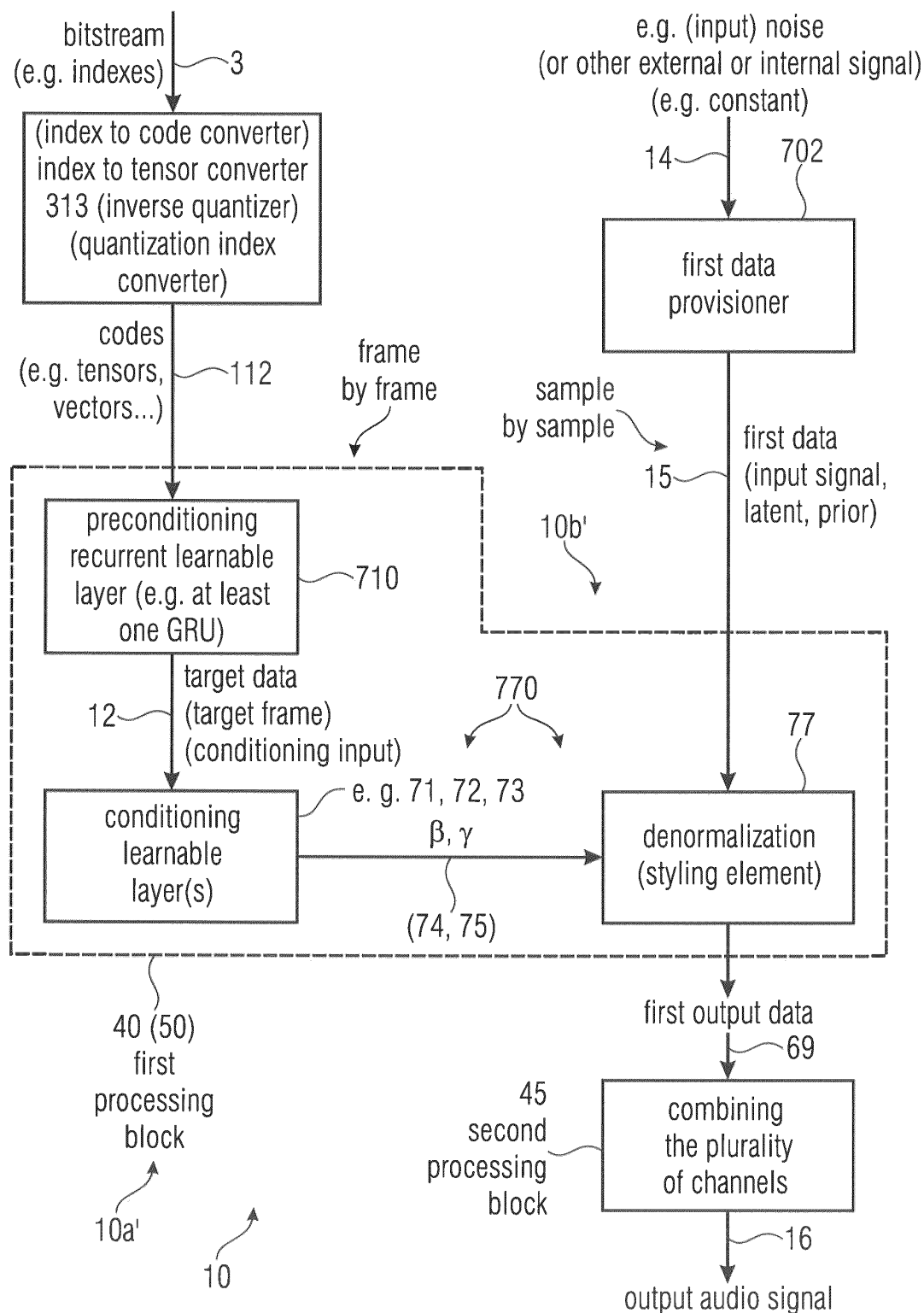


Fig. 7

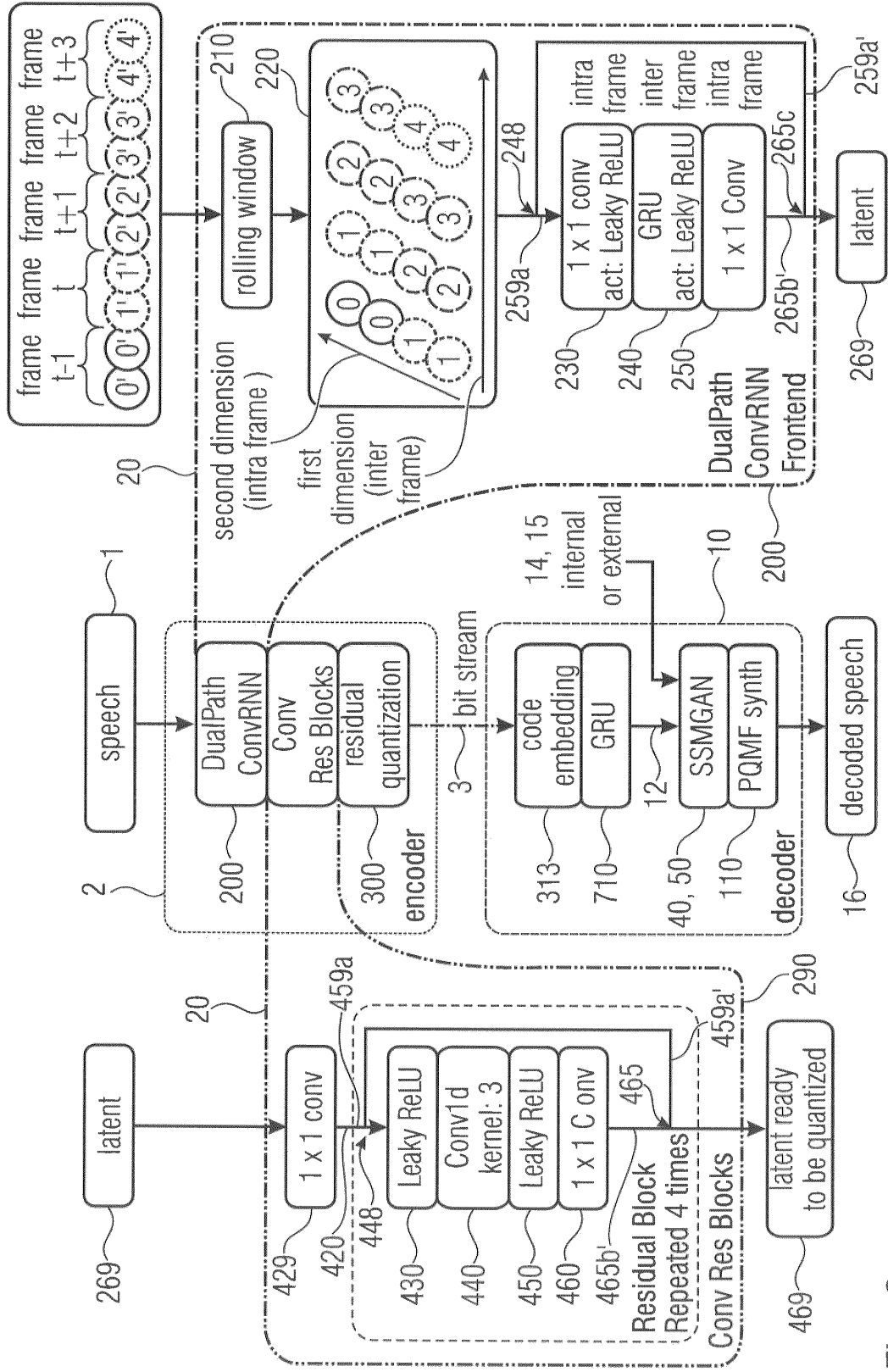


Fig. 8

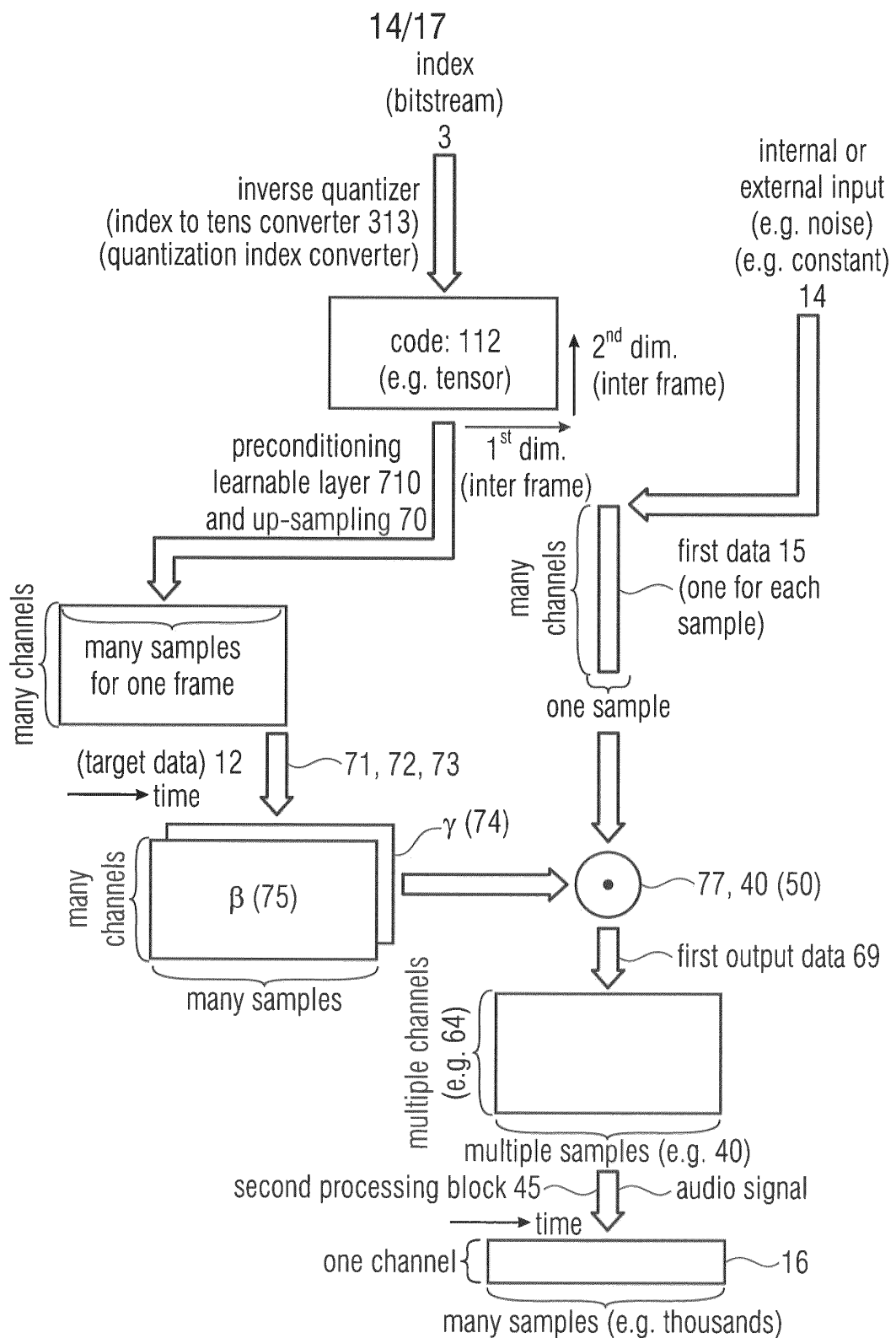


Fig. 9

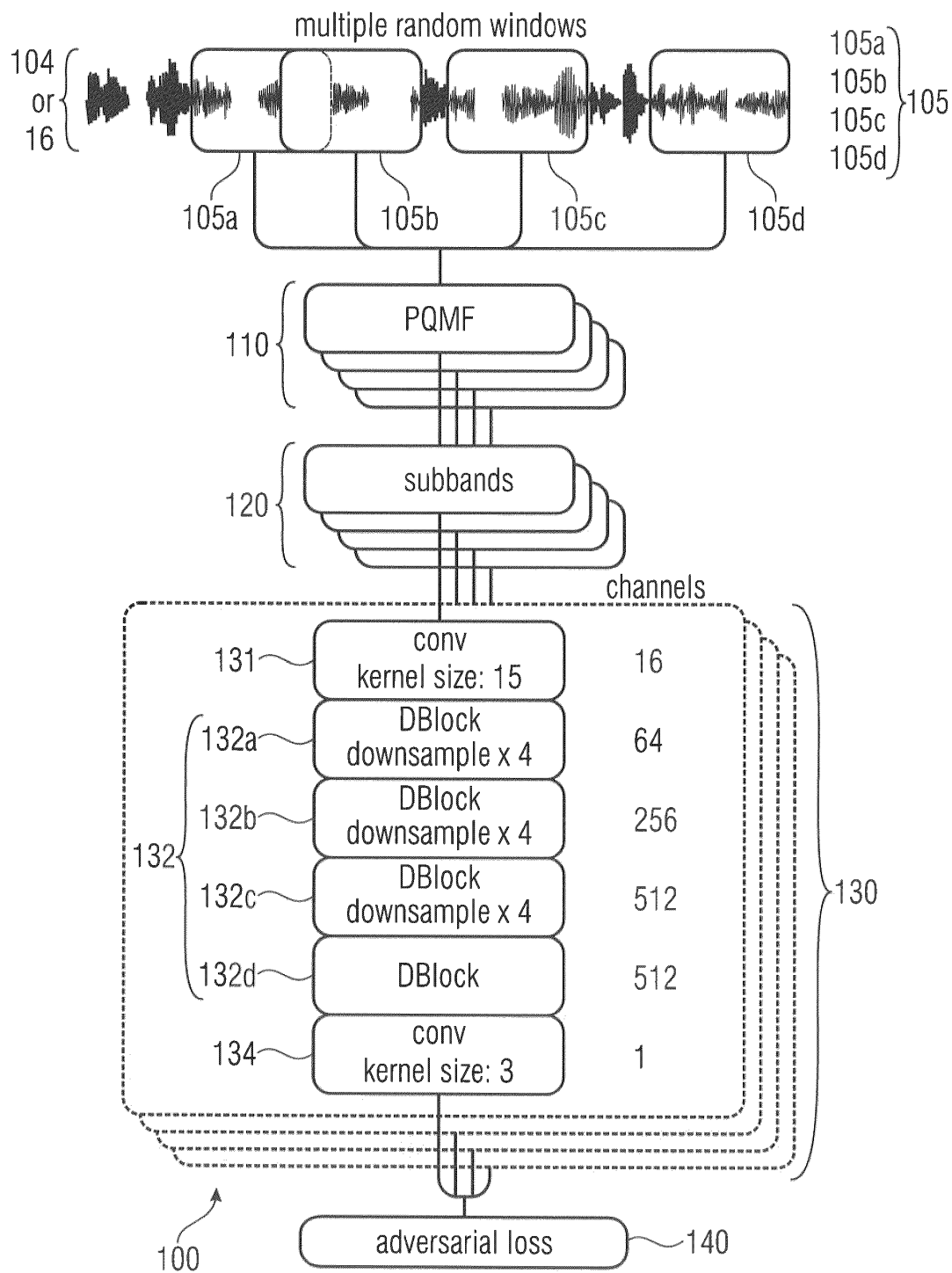
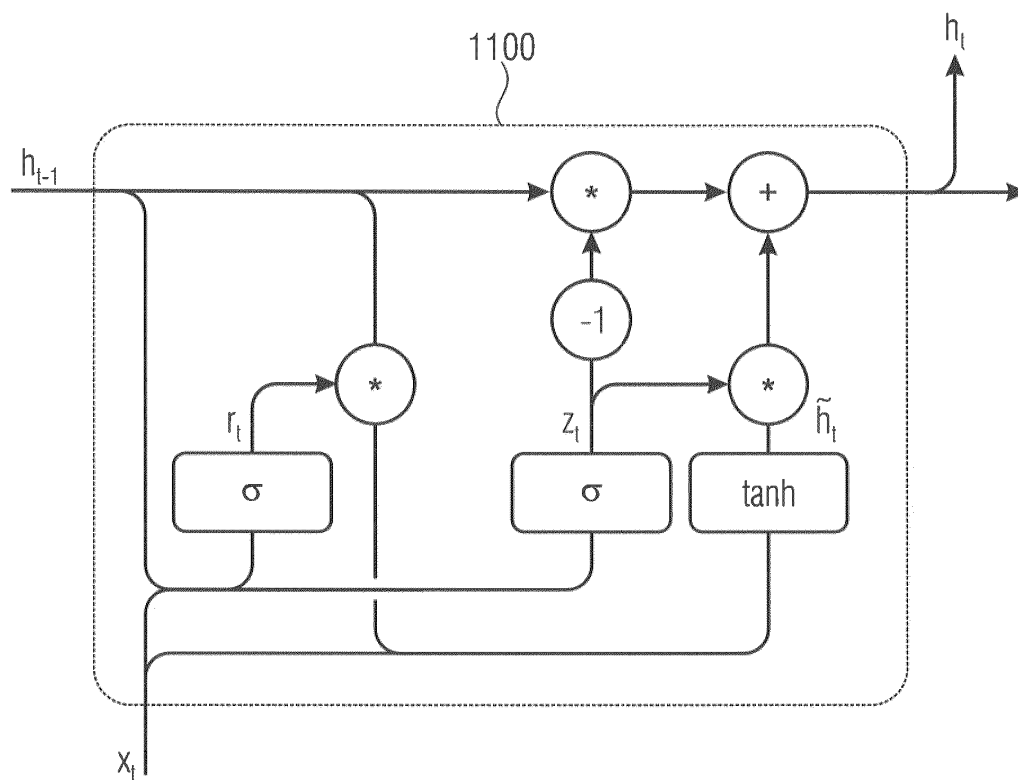


Fig. 10

16/17



$$z_t = \sigma(W_z \cdot (h_{t-1}, x_t))$$

$$r_t = \sigma(W_r \cdot (h_{t-1}, x_t))$$

$$\tilde{h}_t = \tanh(W \cdot (r_t * h_{t-1}, x_t))$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Fig. 11

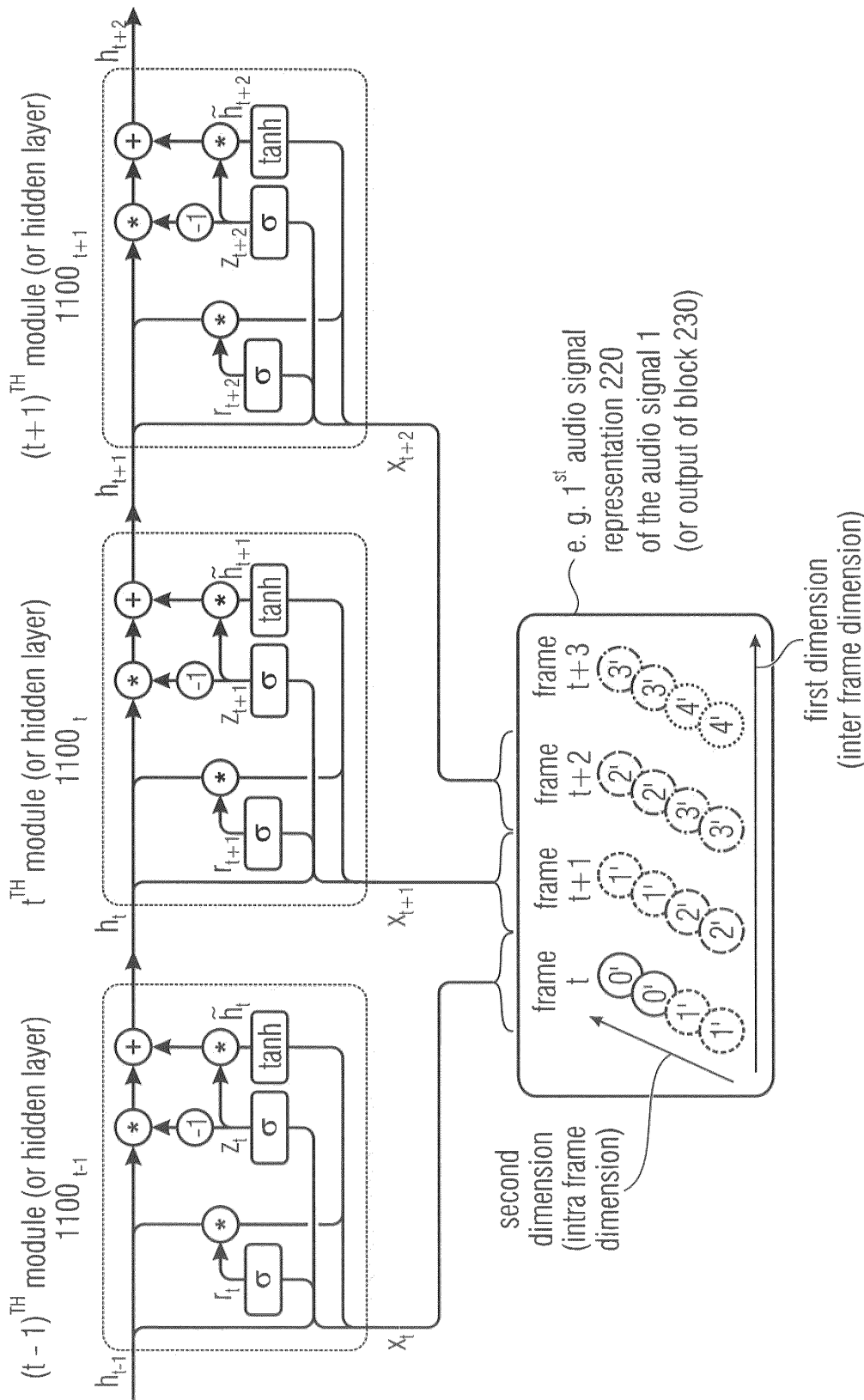


Fig. 12