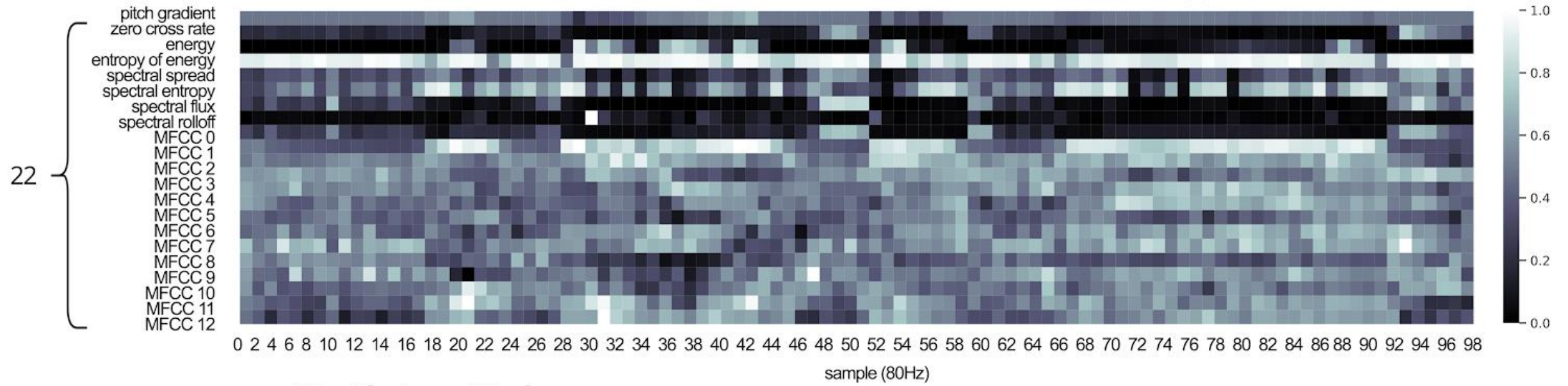
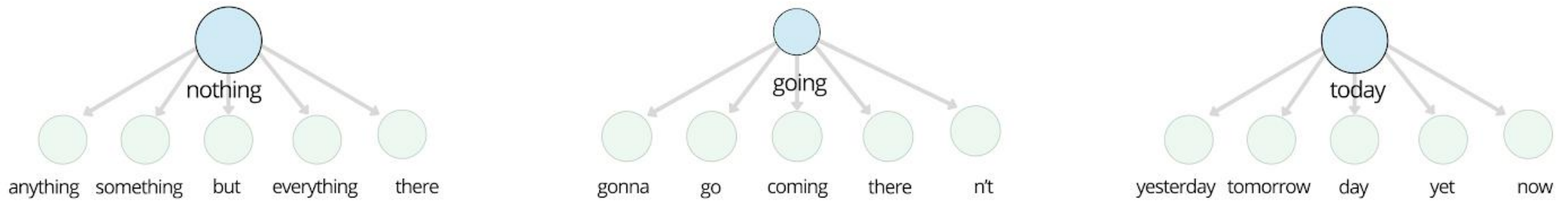
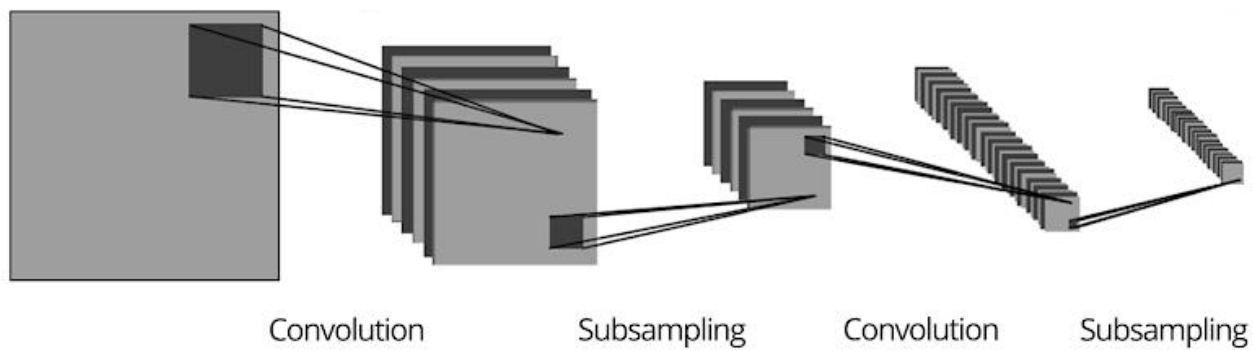


A# Video features: $W \times h \times d \times t$ $t = \text{recorded duration (s)} \times 8\text{Hz}$  $t = 0$ t # Audio features: $22 \times t$ $t = \text{recorded duration (s)} \times 80\text{Hz}$ # Text features: $W \times d$ $w = \text{number of words, } d = \text{vector size (e.g., 200)}$ **B**

Intermediate feature maps

Video input layer

Audio and Text
Input LayersFinal feature
vector