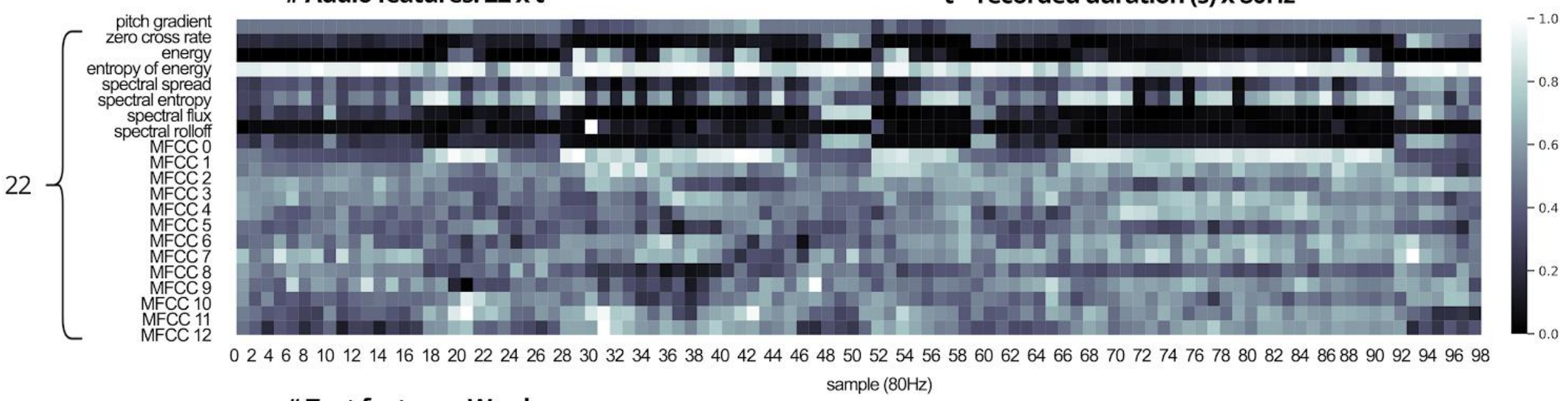
**A**

**# Video features: W x h x d x t**

**t = recorded duration (s) x 8Hz**

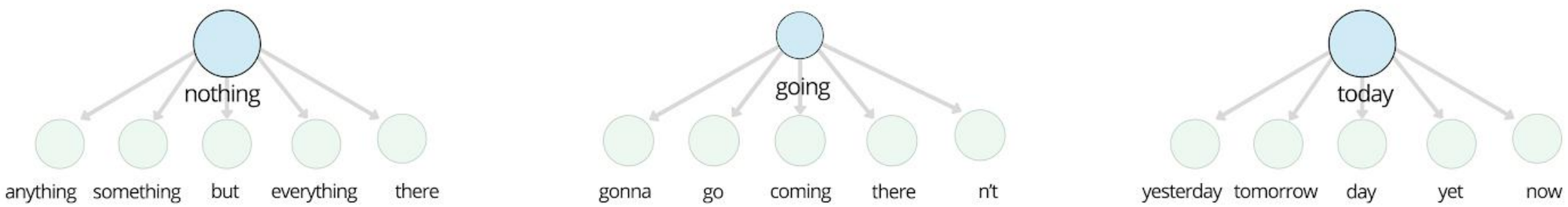

**# Audio features: 22 x t**

**t = recorded duration (s) x 80Hz**




**# Text features: W x d**

**w = number of words, d = vector size (e.g., 200)**




**B**