Mismatched training and dev/test data

Addressing data mismatch

deeplearning.ai

Addressing data mismatch

→ • Carry out manual error analysis to try to understand difference between training and dev/test sets

E.g. noisy — car noise   street numbers

→ • Make training data more similar; or collect more data similar to dev/test sets

E.g. Simulate noisy in-car data

simulate noise in-car data.

Artificial data synthesis

Note: Limited amount of noise audio may cause an over-fitting to the noise audio.
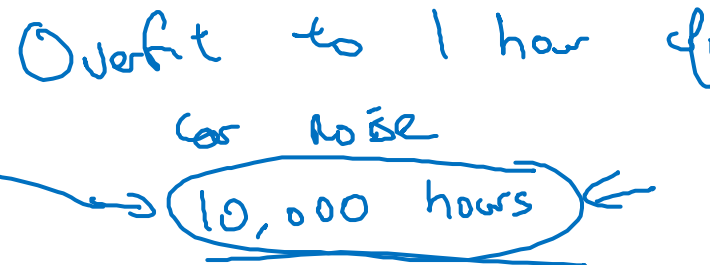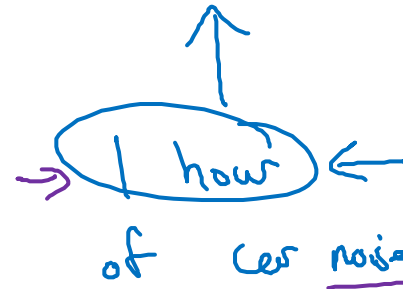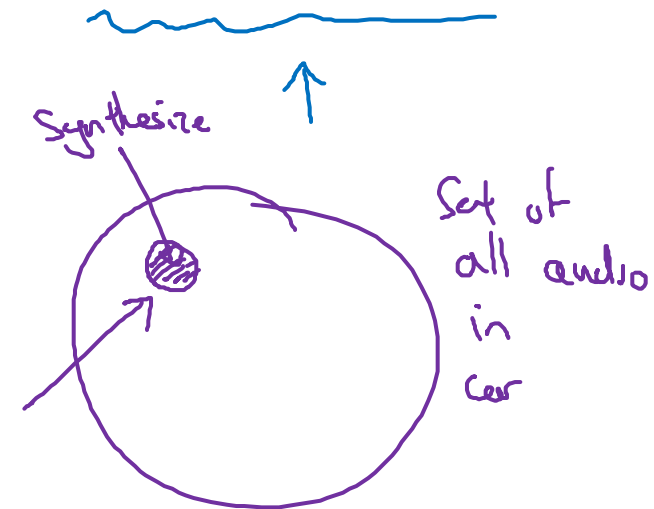
🔊 + 🔊 = 🔊

"The quick brown ← fox jumps over the lazy dog."

↑

10,000 hours

Car noise

↑

→ ⟨ 1 hour ⟩ ←

of   car noise

Overfit to 1 hour of

car noise

⟨ 10,000 hours ⟩ ←

Synthesized in-car audio

∿∿∿∿∿

↑

Synthesize

Set of all audio in car

Andrew Ng

Artificial data synthesis

Car recognition:

≈ 20 cars

Synthesized → All cars