Jan. 16th. CSC2518. LEC02

Phonemes.

phone → triphone → State
pentaphone

① replaced by DNNs.

$P(x|$ "t-d+uw1" $)$ → ~ Mixture of Gaussians (GMM)
$P(x|$ "d-ww-y3" $)$
→ Tied-state ← eg. pattern classification. / recognition.
(frame classification)

Lexical Model.

sequence → [ ] → language model (HMM)
acoustic model (HMM)

② replaced with end-to-end model

Decision-Tree-Based state clustering.

→ Fully connected Layers → 2×5, 10, 11 frames.
→ RNN (LSTM) → # of frame (context) can be dynamically adjusted.

→ Connectionist Temporal Classification (CTC)
Bi-directional LSTM is essential for CTC.

Attention

DNN

Speaker Adaptation eg. VTLN

↳ { Conservative Training.
Transformation methods.
Speaker-aware Training (noise/device-aware) only need acoustic info.