Sequence Discriminative Training; Robust Speech Recognition

Steve Renals

Automatic Speech Recognition – ASR Lecture 15
16 March 2017

Recall: Maximum likelihood estimation of HMMs

- Maximum likelihood estimation (MLE) sets the parameters so as to maximize an objective function F_{MLE}:

$$F_{\text{MLE}} = \sum_{u=1}^{U} \log P_\lambda(\mathbf{X}_u \mid M(W_u))$$

for training utterances $\mathbf{X}_1 \ldots \mathbf{X}_U$ where W_u is the word sequence given by the transcription of the uth utterance, $M(W_u)$ is the corresponding HMM, and λ is the set of HMM parameters

Maximum mutual information estimation

- Maximum mutual information estimation (MMIE) aims to directly maximise the posterior probability (sometimes called conditional maximum likelihood). Using the same notation as before, with $P(w)$ representing the language model probability of word sequence w:

$$
\begin{aligned}
F_{\text{MMIE}} &= \sum_{u=1}^{U} \log P_\lambda(M(W_u) \mid \mathbf{X}_u) \\
&= \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}
\end{aligned}
$$

Maximum mutual information estimation

- Maximum mutual information estimation (MMIE) aims to directly maximise the posterior probability (sometimes called conditional maximum likelihood). Using the same notation as before, with $P(w)$ representing the language model probability of word sequence w:

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log P_\lambda(M(W_u) \mid \mathbf{X}_u)$$

$$F_{\text{MLE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

Maximum mutual information estimation

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

Maximum mutual information estimation

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

- **Numerator**: likelihood of data given correct word sequence ("clamped" to reference alignment)

Maximum mutual information estimation

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

- **Numerator**: likelihood of data given correct word sequence ("clamped" to reference alignment)
- **Denominator**: total likelihood of the data given all possible word sequences – equivalent to summing over all possible word sequences estimated by the full acoustic and language models in recognition. ("free")

Maximum mutual information estimation

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

- **Numerator**: likelihood of data given correct word sequence ("clamped" to reference alignment)
- **Denominator**: total likelihood of the data given all possible word sequences – equivalent to summing over all possible word sequences estimated by the full acoustic and language models in recognition. ("free")
- The objective function F_{MMIE} is optimised by making the correct word sequence likely (maximise the numerator), and all other word sequences unlikely (minimise the denominator)

Sequence training and lattices

- Computing the denominator involves summing over all possible word sequences – estimate by generating lattices, and summing over all words in the lattice
- In practice also compute numerator statistics using lattices (useful for summing multiple pronunciations)
- Generate numerator and denominator lattices for every training utterance
- Denominator lattice uses recognition setup (with a weaker language model)
- Each word in the lattice is decoded to give a phone segmentation, and forward-backward is then used to compute the state occupation probabilities
- Lattices not usually re-computed during training

MMIE is sequence discriminative training

- **Sequence:** like forward-backward (MLE) training, the overall objective function is at the sequence level – maximise the posterior probability of the word sequence given the acoustics $P_\lambda(M(W_u) \mid \mathbf{X}_u)$
- **Discriminative:** unlike forward-backward (MLE) training the overall objective function for MMIE is discriminative – to maximise MMI:
  - Maximise the numerator by increasing the likelihood of data given the correct word sequence
  - Minimise the denominator by decreasing the total likelihood of the data given all possible word sequences

  This results in "pushing up" the correct word sequence, while "pulling down" the rest

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\text{MPE}} = \sum_{u=1}^{U} \log \frac{\sum_W P_\lambda(\mathbf{X}_u \mid M(W))P(W)A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W'))P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{\sum_W P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W'))P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\text{MPE}} = \sum_{u=1}^{U} \log \frac{\sum_W P_\lambda(\mathbf{X}_u \mid M(W))P(W)A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W'))P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u
- F_{MPE} is a weighted average over all possible sentences w of the raw phone accuracy
- Although MPE optimizes a phone accuracy level, it does so in the context of a word-level system: it is optimized by finding probable sentences with low phone error rates

HMM/DNN systems

- DNN-based systems are discriminative – the cross-entropy (CE) training criterion with softmax output layer "pushes up" the correct label, and "pulls down" competing labels
- CE is a frame-based criterion – we would like a sequence level training criterion for DNNs, operating at the word sequence level
- Can we train DNN systems with an MMI-type objective function?

Sequence training of hybrid HMM/DNN systems

- Can we train DNN systems with an MMI-type objective function? – **Yes**
- Forward- and back-propagation equations are structurally similar to forward and backward recursions in HMM training
- Initially train DNN framewise using cross-entropy (CE) error function
  - Use CE-trained model to generate alignments and lattices for sequence training
  - Use CE-trained weights to initialise weights for sequence training
- Train using back-propagation with sequence training objective function (e.g. MMI)

Sequence training results on Switchboard (Kaldi)

Results on Switchboard "Hub 5 '00" test set, trained on 300h training set, comparing maximum likelihood (ML) and discriminative (BMMI) trained GMMs with framewise cross-entropy (CE) and sequence trained (MMI) DNNs. GMM systems use speaker adaptive training (SAT).
All systems had 8859 tied triphone states.
GMMs – 200k Gaussians
DNNs – 6 hidden layers each with 2048 hidden units

	SWB	CHE	Total
GMM ML (+SAT)	21.2	36.4	28.8
GMM BMMI (+SAT)	18.6	33.0	25.8
DNN CE	14.2	25.7	20.0
DNN MMI	12.9	24.6	18.8

Veseley et al, 2013.

Robust Speech Recognition

Additive Noise

- Multiple acoustic sources are the norm rather than the exception
- From the point of view of trying to recognize a single stream of speech, this is additive noise
- Stationary noise: frequency spectrum does not change over time (e.g. air conditioning, car noise at constant speed)
- Non-stationary noise: time-dependent frequency spectrum (e.g. breaking glass, workshop noise, music, speech)
- Measure the noise level as SNR (signal-to-noise ratio), measured in dB
  - 30dB SNR sounds noise free
  - 0dB SNR has equal signal and noise energy

Feature normalization

- Basic idea: Transform the features to reduce mismatch between training and test
- *Cepstral Mean Normalization* (CMN): subtract the mean of the feature vectors from each feature vector, so each feature vector element has a mean of 0
- CMN makes features robust to some linear filtering of the signal — adds robustness to varying microphones, telephone channels, etc.
- *Cepstral Variance Normalization* (CVN): Divide feature vector by standard deviation of feature vectors, so each feature vector element has a variance of 1
- Cepstral mean and variance normalisation, CMN/CVN:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}(\mathbf{x})}{\boldsymbol{\sigma}(\mathbf{x})}$$

Feature compensation: Spectral subtraction

- Basic idea: Estimate the noise spectrum and subtract it from the observed spectra
- Any feature vector can then be computed from the noise-subtracted spectrum
- Problems:
  - Need to estimate noise spectrum from a period of non-speech: requires good speech/non-speech detection
  - Errors in the noise estimate (perhaps arising from speech/non-speech separation errors) result in over-/under-compensation of the spectrum
- Low computational cost, widely used in practice
- "ETSI adavanced front end" uses spectral subtraction and CMN

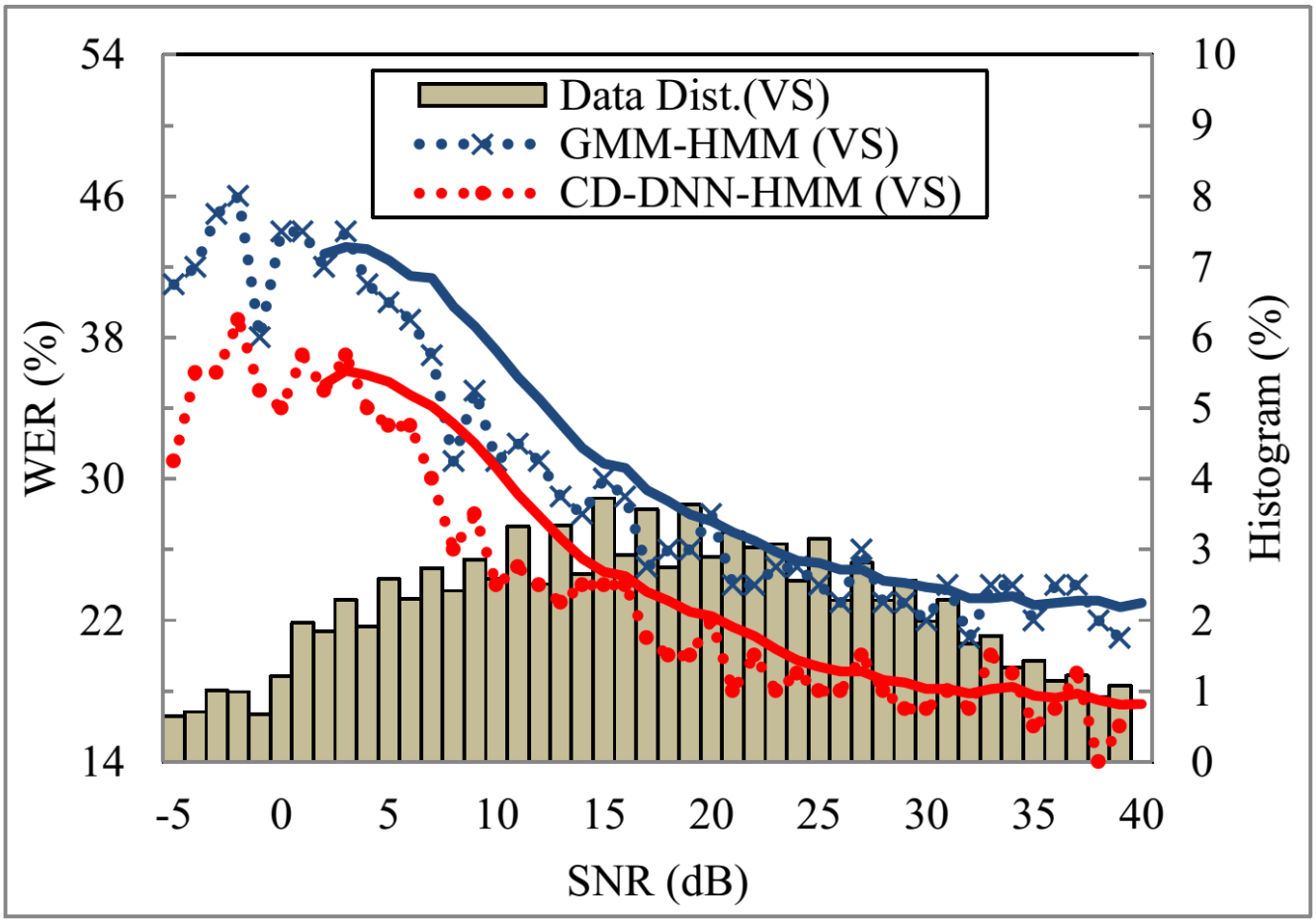
Multi-condition Training

- Basic idea: Don't train on clean speech, but train on speech with a similar noise level (and noise type)
- *Matched condition* — training in the same noise conditions as testing — is rarely possible since the test conditions are nearly always partly unknown
- *Multi-condition training* — train with speech data in a variety of noise conditions
- It is possible to artificially mix recorded noise with clean speech at any desired SNR to create a multi-style training set
- Advantage: training data much better matched to test conditions
- Disadvantage: acoustic model components become less discriminative and less well matched to the training data
- Model adaptation — can further reduce errors using an adaptation technique such as MLLR

Seltzer (2013)

GMMs and DNNs at varying SNRs

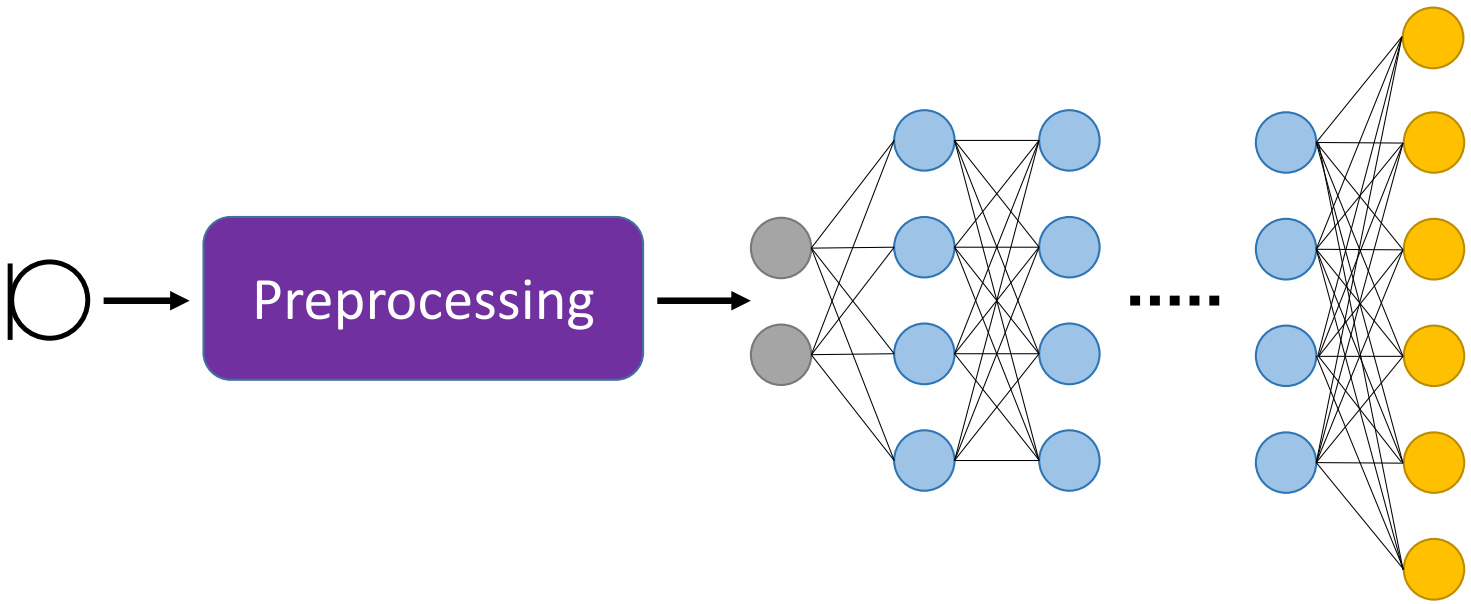
WERs on Microsoft voice search data at varying SNRs (Huang 2014)



Current approaches to robust speech recognition

Decoupled preprocessing: Acoustic processing independent of downstream activity

- Pro: simple
- Con: removes variability
- Example: beamforming for multi-microphone distant speech recognition

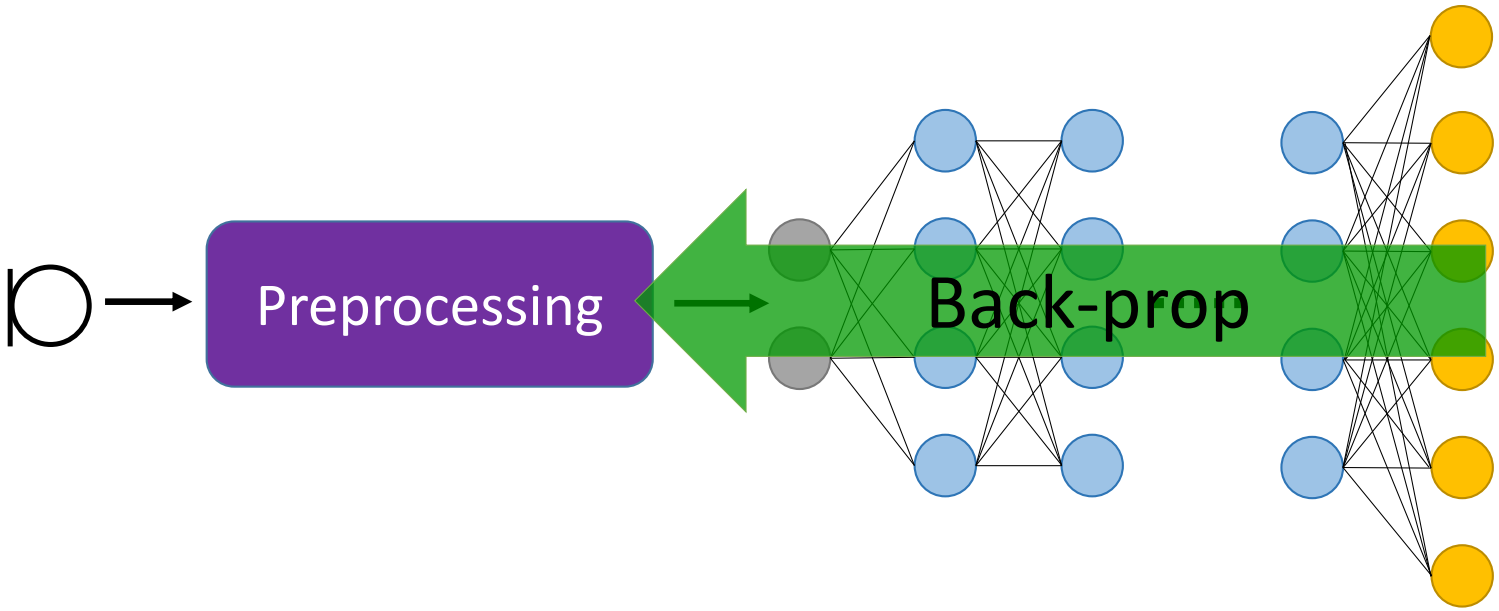


Slide from Mike Seltzer

Current approaches to robust speech recognition

Integrated processing: Treat acoustic processing as initial layers of the network – optimise parameters with back propagation

- Pro: should be "optimal" for the model
- Con: computationally expensive,
- Example: direct waveform systems

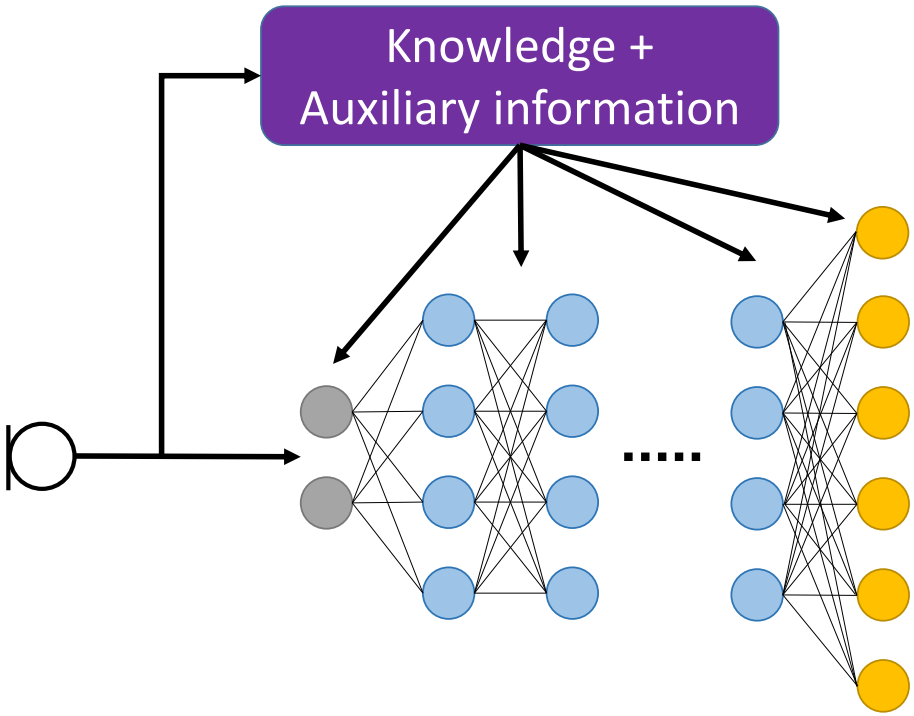


Slide from Mike Seltzer

Current approaches to robust speech recognition

Augmented information: Add additional side information to the network (additional nodes, different objective function, ...)

- Pros: preserves variability, adds knowledge, maintains representation
- Con: not a physical model
- Example: noise-aware training, factorised "noise codes" (iVectors)



Slide from Mike Seltzer

Summary

- Sequence training: discriminatively optimise GMM or DNN to a sentence (sequence) level criterion rather than a frame level criterion
- Noise robustness
  - Important for practical applications of speech recognition
  - Achieve robustness through feature invariance
  - Achieve invariance through large training sets and deep networks
  - Much active research in developiong architectures for robust ASR

Reading

- HMM discriminative training: Sec 27.3.1 of: S Young (2008), "HMMs and Related Speech Recognition Technologies", in *Springer Handbook of Speech Processing*, Benesty, Sondhi and Huang (eds), chapter 27, 539–557. http://www.inf.ed.ac.uk/teaching/courses/asr/2010-11/restrict/Young.pdf
- NN sequence training: K Vesely et al (2013), "Sequence-discriminative training of deep neural networks", Interspeech-2013, http://homepages.inf.ed.ac.uk/aghoshal/pubs/is13-dnn_seq.pdf
- DNNs for robust ASR: M Seltzer et al (2013), "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition", https://www.microsoft.com/en-us/research/publication/an-investigation-of-deep-neural-networks-for-noise-robust-speech-recognition/
  Y Huang et al (2014), "A comparative analytic study on the Gaussian mixture and context-dependent deep neural network hidden Markov models", Interspeech-2014. http://www.isca-speech.org/archive/interspeech_2014/i14_1895.html