Discriminative training and Feature combination

Steve Renals (Hiroshi Shimodaira)

Automatic Speech Recognition— ASR Lecture 13
January-March 2012

Overview

Hot topics in ASR

- Discriminative training
- Combining multiple streams of features

Discriminative training

- Basic idea Estimate the parameters of a speech recognizer so as to make the fewest classification errors (optimize the word error rate)
- Generative model: estimate the parameters so that the model reproduces the training data with the greatest probability (maximum likelihood)
- Generative modelling only results in minimum classification error if certain conditions are met, including
  - the model is correct (i.e. the true data source is an HMM)
  - infinite training data

  This never happens in practice
- Discriminative training:
  - Focus on learning *boundaries* between classes
  - Consider incorrect word sequences as well as correct word sequences
  - This is related to direct optimisation of the posterior probability of the words given the acoustics $P(W \mid \mathbf{X})$

Discriminative training criteria

- Minimum classification error (MCE)
  - Correct parameters if misrecognition occurs
  - Discriminant function is the difference between the log likelihood of the correct sentence and the average likelihood of incorrect competitors
  - Used mainly for small vocabularies
  - Uses training data inefficiently (only considers misrecognised examples)
- Maximum mutual information estimation (MMIE)
  - Maximise the mutual information between the acoustics and word sequence
  - Variant of conditional maximum likelihood
- Minimum Bayes Risk (MBR)
  - Optimise the word error rate rather than a likelihood ratio
  - Use the string edit distance between competing and reference utterances
  - Minimum phone error (MPE) training

MLE and MMIE

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

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

- Maximum mutual information estimation (MMIE) (also referred to as conditional maximum likelihood, CML) aims to directly maximise the posterior probability:

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

$\mathbf{X}_u$ is the acoustic observation sequence for the uth utterance

MMIE

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

- The denominator sums over all possible word sequences estimated by the full acoustic and language models in recognition
- The first term of the numerator is identical to the MLE objective function
- MMIE training corresponds to maximizing the likelihood, while simultaneously minimizing the denominator term
- Discriminative criterion: maximize the probability of the correct sequence (as in MLE) while simultaneously minimizing the probability of all possible word sequences

Optimizing the MMIE objective function

- No straightforward efficient optimization approach for F_{MMIE}
- Gradient-based approaches are straightforward but slow
- Extended Baum-Welch (EBW) algorithm provides update formulae similar to forward-backward recursions used in MLE
- Extended by Povey (PhD thesis, 2003) using notions of strong-sense and weak-sense auxiliary functions
- For large vocabulary tasks, estimating the denominator is expensive (an unpruned decoding!)—in practice it is estimated using word lattices to restrict the set of words sequences that are summed over

MPE: Minimum phone error

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

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

- $A(W, W_u)$ 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

Example: meeting speech recognition

System	Training criterion	PLP
Baseline	ML	28.7
SAT	ML	27.6
SAT	MPE	24.5

Combining multiple feature streams

- Basic idea Different representations of the speech signal are possible: if they result in complementary errors than it may reduce error rates to combine them
- Combination at the feature level: linear discriminant analysis (and related methods) to combine feature streams
- Combination at the acoustic model level: combine frame-level probability estimates (multi-stream methods)
- Combination at the system level: combine the word sequence outputs of different recognizers (ROVER)

Feature combination

- Basic idea Compute different feature vectors for each frame and train acoustic models on all of them
- Simplest approach: concatenate feature vectors at each frame
  - Increases the dimensionality
  - May be strong correlations between the feature streams (can cause problems for diagonal covariance Gaussians)
- Transform concatenated feature vectors (linear discriminant analysis (LDA), principal component analysis (PCA))
  - dimension reduction
  - decorrelation
- PCA estimates a global transform; LDA estimates a transform per-class / per-state / per-component

LDA: Linear discriminant analysis

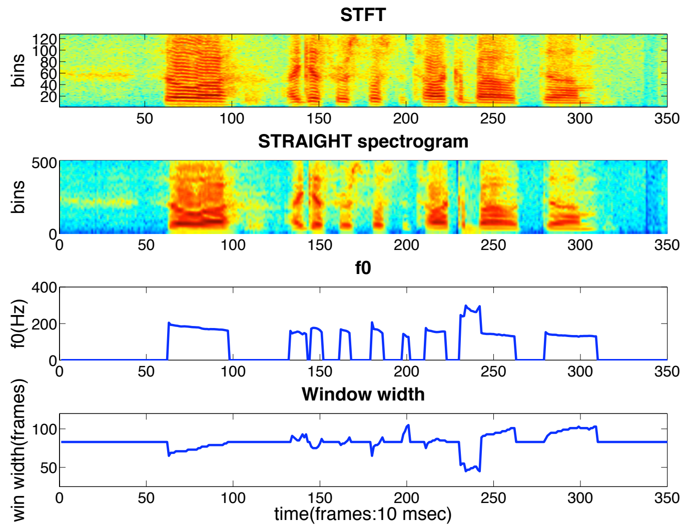
- LDA aims to find a linear transformation (from d dimensions to p dimensions, $p \leq d$) given by a matrix $\boldsymbol{\theta}^T$:
  $$\mathbf{z} = \boldsymbol{\theta}^T \mathbf{x}$$
  $\boldsymbol{\theta}^T$ projects $\mathbf{x}$ to a vector $\mathbf{z}$ in a lower dimension space
- The LDA transform $\boldsymbol{\theta}^T$ is chosen to simultaneously
  - maximise the between class covariance Σ_b
  - minimise the within class covariance Σ_w

  using the eigenvectors corresponding to the p largest eigenvalues of $\Sigma_b \Sigma_w^{-1}$
- HLDA: Heteroscedastic Linear Discriminant Analysis
  - In LDA classes share the same within-class covariance matrix
  - In HLDA a different covariance matrix is estimated for each class
- Both HLDA and LDA assume a Gaussian distribution
- NB: "class" may be a phone, a state or a Gaussian component, depending on the amount of data

Example: STRAIGHT features

- Conventional PLP and MFCC computation use a fixed size analysis window
- STRAIGHT spectral representation (Kawahara et al, 1999): smoothed spectral representation computed using a pitch adaptive window
- Requires a use of a pitch tracker to obtain F_0
- Resolution of STRAIGHT spectrogram follows the values of the fundamental frequency
- Can use STRAIGHT spectral analysis to obtain STRAIGHT MFCCs (and STRAIGHT PLPs)
- For recognition, combine STRAIGHT and conventional MFCCs using HLDA, reducing from 78 dimensions (39+39) to 39

STRAIGHT Spectral Analysis



Results on CTS

	TOTAL	Female	Male	SW1	S23	Cell
MFCC (no CMN/CVN)	**42.7**	41.8	43.6	36.5	43.3	47.9
STRAIGHT (no CMN/CVN)	**45.7**	44.5	46.9	40.0	46.6	50.3
MFCC+CMN/CVN+VTLN	**37.6**	37.0	38.3	31.8	37.1	43.5
STRAIGHT +CMN/CVN+VTLN	**39.2**	38.2	40.1	33.6	39.0	44.5
MFCC + STRAIGHT +CMN/CVN+VTLN+HLDA	**34.7**	33.8	35.6	28.6	34.7	40.5

Results on Meetings

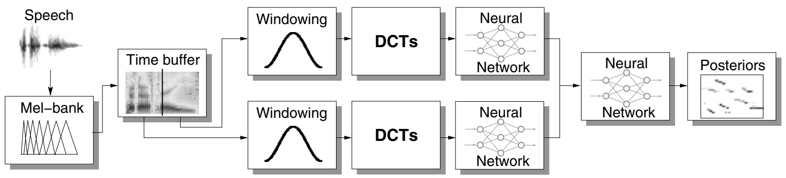
	TOTAL	Female	Male	CMU	ICSI	LDC	NIST
MFCC+VTLN	**38.4**	38.5	38.3	42.7	23.9	52.1	30.9
STRAIGHT+VTLN	**39.3**	38.3	39.7	44.7	24.8	53.1	31.2
MFCC+STRAIGHT +VTLN	**42.1**	44.4	41.0	45.6	28.5	55.4	37.0
MFCC+STRAIGHT VTLN+HLDA	**36.6**	36.3	36.7	41.0	22.5	51.2	28.5

Example: Discriminative features

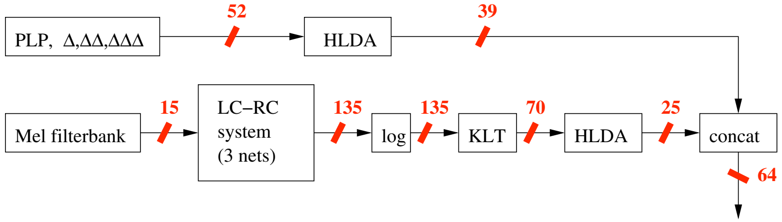
- Can also use the outputs of other statistical models as a feature stream
- Neural networks (eg multi-layer perceptrons – MLPs) when trained as a phone classifier output a posterior probability $P(\text{phone}|\text{data})$
- This is a locally discriminative model
- MLP probability estimates can be used as an additional feature stream, modelled by the HMM/GMM system (*Tandem*)
- Advantages of discriminative features
  - can be estimated from a large amount of temporal context (eg ± 25 frames)
  - encode phone discrimination information
  - only weakly correlated with PLP or MFCC features

Tandem features

- Tandem features



- LCRC features



Example: meeting speech recognition

- Tandem (LCRC – left context, right context) features (Karafiat, 2007)
- Derived from multiple stages of MLPs that try to estimate phoneme state posterior probabilities
- Wide context:input to these is not only the feature vector at the current time, but 25 surrounding frames as well
- Separate MLPs for left and right context

System	Training criterion	PLP	LCRC+PLP
Baseline	ML	28.7	25.2
SAT	ML	27.6	23.9
SAT	MPE	24.5	21.7

Summary

- Discriminative methods optimize a criterion other than maximum likelihood (eg more directly related to the error rate)
- But, we still want to optimize all parameters according to a consistent criterion
- Combining features can take advantage of approaches which are complementary, but still make different errors
- Increasing emphasis on approaches which view the features as another model to be optimized