Presentation to IBM

Cambridge University Engineering Department

Jan 30th 2003

Dan Povey

Minimum Phone Error – better than MMI
Practical Issues for Implementation

Theory of Optimisation (MPE)

Theory of Optimisation (MMI)

Introduce MPE, give results on various corpora

Four sections (each followed by questions):

Summary
\[ P(\mathcal{M}_\text{PE}) = \sum_R \left( \gamma \left( R \right) \right)^{\text{phone}\text{accuracy}(s, \mathcal{M}_\text{PE})} \]

Errors where \( \text{phone}\text{accuracy}(s, \mathcal{M}_\text{PE}) \) is \# phones in reference, minus \# phone

i.e., an average of phone accuracy, weighted by sentence likelihood

\text{Minimum Phone Error (MPE)}
Maximum Mutual Information (MMI)
Difference is not so simple for more complex examples.

\[
\frac{\lambda q + \lambda p}{0 \times q + 1 \times p} = \begin{align*}
\text{MCE objective function} \\
\text{MPE objective function for other training files.}
\end{align*}
\]

\[
\frac{q + p}{0 \times q + 1 \times p} = \begin{align*}
\text{MPE objective function} + \text{other training files.}
\end{align*}
\]

\[
\text{MNI} \text{ objective function} = \text{log} (\frac{q + p}{p}) + \begin{align*}
\text{other training files.}
\end{align*}
\]

\[
\text{ML objective function} = \text{log} (\frac{q + p}{p}) + \begin{align*}
\text{other training files.}
\end{align*}
\]

Let \( a \) be same for "b".

\[
(\text{acoustic \& \ LMs \ likelihood}) \quad \text{b is same for} \quad \text{"b".}
\]

Suppose correct sentence is "a", only alternative is "b".

**Comparison of objective functions**
Criterion

log p(Correct)−log p(Incorrect)

MPE criterion
MCE criterion
MMI criterion

Comparison of objective functions (cont'd)
On baseline systems on various corpora (no MLLR), relative improvement of MPE vs ML:

- Switchboard
- NAB/WSJ
- RM
- BN

![Graph showing relative % improvement vs log (#frames/Gaussian)]
Comparison of MPE with MM1, I-smoothed MM1
phone-marked lattices
... cannot release HTK code before we release a recogniser which can produce

We are using MPE in evaluations
(depends on #Gaussians in HMM set)

With all bells and whistles, perhaps 5% relative improvement from MPE

... but each tends to reduce the relative improvement due to MPE

WLLR, HLD, SAT, gender adaptation, improve absolute results

On a baseline evaluation systems, typically about 10-12% relative improvement

Combination with other techniques
\[ \mathcal{X} = \mathcal{X} \left| (\mathcal{Y}) \frac{\chi_0}{\Theta} = \mathcal{X} = \mathcal{X} \left| (\mathcal{X}, \mathcal{Y}) \delta \frac{\chi_0}{\Theta} \right. \]

iff \( (\mathcal{Y}) \mathcal{X} \) is a weak-sense auxiliary function for \( (\mathcal{X}, \mathcal{Y}) \delta \)

iff \( (\mathcal{Y}) \mathcal{X} - (\mathcal{X}) \mathcal{Y} > (\mathcal{X}, \mathcal{Y}) \delta - (\mathcal{X}, \mathcal{Y}) \delta \) is a strong-sense auxiliary function for \( (\mathcal{X}, \mathcal{Y}) \delta \)

Definitions:

**Auxiliary Functions**
Weak sense auxiliar has same differential around local point $x = \chi'$ at a local point $x = \chi$, but $\nabla \Phi$ everywhere else.

Strong sense auxiliar function: has the same value as real objective function

Auxiliar functions cont'd
Say, a variance matrix or a probability, etc. Useful where "natural" form of parameters is not a normal linear variable, but, where gradient descent in functional form of update is more efficient than gradient descent (which will only converge for correct speed of optimization). Similar level of guarantee to gradient descent (which will only converge for update). ... But if it does converge it will converge to a local maximum (only fixed point). ... A weak sense auxiliary function does not give such a guarantee. Strong sense auxiliary functions give a guarantee of convergence.

Auxiliary functions & function maximization
Optimising Gaussian Likelihoods
\[
\chi = \mathcal{L} \frac{\partial}{\partial \mathcal{L}} \mathcal{L} = \mathcal{L} \frac{\partial}{\partial \mathcal{L}} \mathcal{L}
\]

Adding this does not change the gradient where \( \chi \).

\[
\frac{\partial}{\partial \mathcal{L}} \mathcal{L} \mathcal{L} \mathcal{L} \frac{\partial}{\partial \mathcal{L}} \mathcal{L} \mathcal{L} \mathcal{L}
\]

This has zero differential where \( \chi \).

In order to make sure aux function is convex, add

Optimising Gaussian Likelihoods cont'd
Optimizing Gaussian likelihoods cont'd

Note: optimization techniques affect recognition results independently of criterion value.

For good convergence set $D_{\text{den}} = \mathcal{H}$. e.g. for $E = 1$ or $2$

\[
\frac{w_{\text{f}}}{D} + \left\{ \frac{w_{\text{f}}}{\text{F}_{\text{f}}} \right\} = w_{\text{f}}
\]

the mean): Solving this leads to the Extended Baum-Welch update equations, e.g. (for

$\text{Pover}$: Minimum Phone Error
Optimising MPE objective function

Parameter values for MPE objective function in terms of log-likelihoods around current \((b')d\log\theta^\mathcal{H}\). Use intermediate [weak-sense] auxiliary function which is linear expansion about phone arcs in the lattice for phone models. Lattice likelihood computation for MPE uses fixed start \& end points.

\[(b)d\log\left(\gamma=\gamma\right)\mathcal{H} = (\gamma, \gamma)\in\mathcal{H}\]

\((b)\) is a shorthand for the acoustic likelihood for arc.
Leads to Extended Baum-Welch equations

auxiliary function, for it is known

Since $H_{MPE}(\lambda, \lambda')$ is basically the same form as M1 opt, a weak sense

\[
\begin{align*}
(b) & \log \left( \frac{b}{d \log (c)} \right) | (b) d \log (c) | (b) d \log (c) \\
& = \frac{b}{d \log (c)} - \max_{1=b} \frac{b}{d \log (c)} = 0
\end{align*}
\]

This is very similar to the M1 objective function, separate out +ve and -ve

Optimising MPE objective function
is average correctness of sentences in the speech file

is average correctness of sentences including arc $\tilde{h}$, and

is the occupation probability of the arc (as in MLE),

where

\[
(b) \omega \big| \mu = \frac{b\mu}{\sum_{\mu} b\mu}
\]

Can be calculated as:

... Trivial to calculate sufficient statistics once this is calculated

... which is scaled differential of objective function w.r.t. log likelihood of arc:

\[
\frac{(b) d \log Q}{d b\mu} = \frac{b\mu}{\sum_{\mu} b\mu}
\]

Important definition MPE objective function, cont'd
Approximate one is simpler to implement.

Similar in terms of performance and time taken.

Two techniques used: approximate and exact.

Calculations specific to MPE are to calculate $c(b)$ and $c_{avg}$. Optimising MPE objective function, cont'd.
Approximate MPE

$\text{Povm: Minimum Phone Error}$

- Respectively
- Tradeoffs between insertion & correct phone, and insertion & deletion,

\[
\begin{cases}
-1 + e \text{ if different phone} \\
-1 + 2e \text{ if same phone}
\end{cases} = (b)_{\text{PhoneAcc}} \quad \cdots
\]

extents of overlap as proportion of extent of phone $z$ is 0 $\leq e \leq 1$.

- Approximation: if $h$ overlaps with a reference (time-marked) phone $z$, and

\[
\begin{cases}
-1 \text{ if insertion} \\
0 \text{ if substitution} \\
1 \text{ if correct phone}
\end{cases} = (b)_{\text{PhoneAcc}} \quad \text{sum over phones in sentence s, of}
\]

which equals # correct phones - # insertions

- Function $\text{RawPhoneAccurate(s)}(s')$ equals # phones in $s'$ minus # errors
Algorithm to calculate $c(h)$ for each $h$

- Easy to integrate each phone's contribution into a forward-backward like

Approximate MPE cont'd
Use a traceback algorithm to find alignment.

View each hypotheses phone as $P + 1$ separate arcs, depending on position.

If reference transcription has $P$ phones,

Illustrate this consideration of hypotheses sentonce (i.e. single sentonce in the lattice) to optimization.

Considers a single hypotheses sentonce except for

Doesn’t rely on time alignment of reference transcription.

Exact MPE
Exact MPE cont'd
Forward-backward algorithm can be done for lattice

This makes it possible to integrate the traceback into a forward-backward type

... so only the traced-back path has a nonzero probability

This traceback algorithm can be formulated in terms of transition probabilities

\text{Exact MPE cont'd}
for MMI) Generally more iterations than MMI to reach optimum WER (e.g. 6-8 vs. 4

... and use \(E=2\) to control optimization speed

As MMI, Generate lattice just once (at start)

Optimization Schedule
(Named in reference to H-criterion, turned out not to be a criterion)

Very important if MPE is to give test-set improvements

... which is set to the ML estimates of the Gaussian parameters

... which is log likelihood of \( \mathcal{L}(\eta) \) points from the distribution

\((\eta \mid \theta, \pi) \propto \pi \times \mathcal{L}(\eta) \)

Log prior distribution is \( \mathcal{O} \) the use of a prior over the Gaussian parameters

\( I \)-smoothing
... ●

\[ T = \frac{f_0}{2} \]  Generally the best value

- Generally optimisation affects recognition independently of criterion value

- Generally 8 iterations

\[ \theta \]  Optimisation speed: constant 

\[ \frac{1}{10} \]  Generally in range \( \frac{1}{10} \)...

- Probability scale: \( R \) best value is Generally the inverse of the normal LM scale,

\[ \text{BN WSJ} \]

- Implementation issues have mostly been tested on 3 corpora (Switchboard,

Other Implementation Issues
beam? 150 or so (MPE is less sensitive to this than MML)

- Training lattices can be generated just once but need to be big enough, e.g.

- Approximate and exact MPE give similar results

- MPE better than Minimum Word Error (calculate error on a word level)

unigram

... note that we use bigram to generate actual words in lattice, haven't tried

- Language model in lattices: unigram better than bigram, zero-gram

Other Implementation Issues
Should be easy to combine this with EMLT etc.

• Of prior
  – Smooth off-diagonal parts of the ML variance estimates which form center
  – Use a larger value of \( \tau \) (for smoothing) for variance, e.g. 50–500

To get this to work, it’s necessary to

We don’t do this because we don’t have working full-cov MLR

• best diagonal-covariance systems
  – Relative improvement from MPE is less, but absolute results better than the
    MPE also gives improvements for full-covariance systems
  – Not scalars in update equations
  – Optimization technique is trivial to extend to full covariance: Just use vectors

Full covariances
Second case easier, I think it can help but I haven’t done proper comparisons.

First case gave impressive gains on I Class-state W5J system, but didn’t work well on larger systems. With MI it gave a degradation criterion

- Assume parameters are MPE-trained, train transforms to maximise MPE
- Assumption (can do MPE training later)

Two separate cases:

- MPE can be used to train HLD A transforms or “semi-trained” transforms (Not published or in my PhD)

Transforms and MPE
What does the function $B_{\text{est}}(N)$ look like?

be described in $N$ bits

let $B_{\text{est}}(N)$ equal the WER of the best speech recognition system that can

Suppose the recognition system is described in $N$ bits

I believe complexity is potentially good in terms of recognition rate

complexity (of the system) good or bad?

Grand Theory of Speech Recognition
Grand Theory of Speech Recognition (2)
But don't want to lose control of the code base with or added on, etc.

Want to increase $N$ (have many adjustable parameters, bits that can be played

Grand Theory of Speech Recognition (3)
with adjustable properties

| favour neural-net type architectures but with many different “neuron-tybes”

• Evolution, DNA, etc...

them

issue of being able to implement them does not arise, just have to duplicate numbers

• So parts of the system description are not human-understandable, they are just parameters

• Have parts of the system defined by some kind of code, or by numerical

Potential solution...

Grand Theory of Speech Recognition (4)