

Error approximation and minimum phone error acoustic model estimation

Matthew Gibson and Thomas Hain

Abstract—Minimum phone error (MPE) acoustic parameter estimation involves calculation of edit distances (errors) between correct and incorrect hypotheses. In the context of large vocabulary continuous speech recognition, this error calculation becomes prohibitively expensive and so errors are approximated. This paper introduces a novel error approximation technique. Analysis shows that this approximation yields a higher correlation to the Levenshtein error metric than a previously-used approximation. Experimental evaluations on a large vocabulary recognition task demonstrate that the novel approximation also delivers significant performance improvements over the previously-used approximation when applied to MPE acoustic model estimation.

Index Terms—Discriminative training, acoustic modelling, minimum phone error.

I. INTRODUCTION

DISCRIMINATIVE training of acoustic models has yielded significant classification performance improvements over maximum likelihood (ML) trained models for the task of automatic speech recognition [1; 2]. The minimum phone error (MPE, [3]) or minimum word error (MWE, [4]) method is an example of a discriminative approach to model estimation. Acoustic models estimated using the MPE technique have not only displayed significant classification performance improvements over ML-estimated models, but have also yielded significant improvements over models learned using other discriminative approaches [3].

The implementation of MPE acoustic model estimation involves approximation of errors associated with a set of transcriptions of the training data, as will be discussed in Section III-A. This paper introduces a novel error approximation method and demonstrates how it addresses limitations of a previously used technique. While it is certainly not evident that more accurate error approximation during model estimation leads to improved model generalisation [5], the novel error approximation method introduced here is found to yield significant performance improvements when deployed for MPE acoustic model estimation.

The paper is structured as follows. Sections II and III respectively explain the theory and implementation of MPE acoustic model estimation. Section IV describes previously introduced approaches to error approximation. Limitations of an alignment-based approach, referred to as the baseline approximate error, are explained in Section V. Novel alignment-based approximations are proposed in Section VI. The experimental system is described in Section VII. The accuracy of the novel approximations is compared to the accuracy of the baseline approximate error in Section VIII. The effect of the novel approximations upon MPE parameter

re-estimation is experimentally measured in Section IX. A concluding discussion is found in Section X.

II. MINIMUM PHONE ERROR THEORY

The MPE criterion $R_{\text{MPE}}(\theta)$, also referred to as the overall risk criterion [4; 6], is defined by Equation 1.

$$R_{\text{MPE}}(\theta) = \frac{1}{R} \sum_{r=1}^R \sum_{w_1^N \in \mathcal{W}} p(w_1^N | \mathbf{o}_r, \theta) L(w_1^N, \hat{w}_1^{M(r)}) \quad (1)$$

The set \mathcal{W} comprises all possible phoneme transcriptions of the acoustic data \mathbf{o}_r , $\hat{w}_1^{M(r)}$ is the correct transcription of \mathbf{o}_r , $M(r)$ is the length of the correct transcription, and $L(w_1^N, \hat{w}_1^{M(r)})$ is the Levenshtein distance between the correct transcription and hypothesis w_1^N . The set \mathcal{W} is called the hypothesis space, θ represents the model parameters and R is the number of training set examples. The symbol w_1^N denotes a hypothesis, where N is the number of labels in the hypothesis.

Adjustment of the acoustic model parameters such that the MPE criterion is minimised is generally performed by iterative updates of the model parameters. In the case of continuous density hidden Markov model (HMM) acoustic models, these updates are given by the extended Baum-Welch (EBW) update formulae. Different versions of the EBW formulae have been introduced, corresponding to the conditional maximum likelihood (CML, also known as maximum mutual information, MMI), minimum classification error and MPE criteria. In the case of the MPE criterion, the EBW update for the mean μ_s of a Gaussian state output distribution is given by Equation 2. Note that only mean updates are used in this work.

$$\hat{\mu}_s = \frac{\theta_s + D\mu_s}{\gamma_s + D} \quad (2)$$

In the above equation, $\hat{\mu}_s$ is the updated mean, γ_s is described by Equation 3 and θ_s is given by Equation 4. The quantity D is a learning rate discussed in Section III-B.

$$\gamma_s = \sum_{r=1}^R \sum_{w_1^N \in \mathcal{W}} K(r, w_1^N, \theta) \sum_{t=1}^{T(r)} \gamma_s(t | w_1^N, \mathbf{o}_1^{T(r)}, \theta) \quad (3)$$

$$\theta_s = \sum_{r=1}^R \sum_{w_1^N \in \mathcal{W}} K(r, w_1^N, \theta) \sum_{t=1}^{T(r)} \gamma_s(t | w_1^N, \mathbf{o}_r, \theta) \mathbf{o}_t(r) \quad (4)$$

The occupancy $\gamma_s(t | w_1^N, \mathbf{o}_r, \theta)$ is the conditional probability that state s is the t -th element of the hidden state sequence, given observation sequence \mathbf{o}_r , hypothesis w_1^N and model

parameters θ . The term $K(r, w_1^N, \theta)$ is described by Equation 5.

$$K(r, w_1^N, \theta) = p(w_1^N | \mathbf{o}_r, \theta) [L_{\text{av}}(r) - L(w_1^N, \hat{w}_1^{M(r)})] \quad (5)$$

In Equation 5, $L_{\text{av}}(r)$ is the average error of all hypotheses, given by Equation 6.

$$L_{\text{av}}(r) = \sum_{w_1^N \in \mathcal{W}} p(w_1^N | \mathbf{o}_r, \theta) L(w_1^N, \hat{w}_1^{M(r)}) \quad (6)$$

There have been several different justifications of the EBW update equations for continuous density HMMs. These justifications include:

- use of a discrete approximation to a continuous probability distribution [4; 7; 8].
- use of a weak-sense auxiliary function to derive the update equation [3; 9].
- proof of existence of a sufficiently large learning rate D such that iterative updates result in optimisation of the criterion [10].
- use of a standard auxiliary function to derive the update equation [11].

The statistics necessary for MPE model estimation, namely the quantities θ_s and γ_s , are calculated via a lattice-based implementation discussed in Section III. The learning rate D is discussed in Section III-B.

III. IMPLEMENTATION OF MPE PARAMETER UPDATES

This section summarises the state of the art implementation of MPE training used in this work. When using small vocabulary systems it is possible to calculate the statistics required to perform the parameter update specified by Equation 2 without approximation. However, in the context of large vocabulary continuous speech recognition, a prohibitively large amount of computation is required to gather these statistics. This is due to the size of the hypothesis space \mathcal{W} . A practical solution to this problem is to approximate the hypothesis space, and hence the resulting statistics, using an n-best list of the hypotheses of highest posterior [4]. In the context of large vocabulary continuous speech recognition, use of a lattice [2] to represent the hypothesis space is favoured because it is a more compact representation of such a list. For this reason, a lattice representation of the hypothesis space is used in this work.

A. Lattice-based MPE

A lattice-based implementation of MPE estimation is introduced in [3; 12]. Lattices which include temporal alignment information, i.e. label start and end times, are used, and the lattice encodes the alignments of the acoustic data of highest posterior [13]. A lattice is generated via a recognition pass of a speech utterance. Additionally, the alignments of the correct label sequence of highest posterior, generated using a constrained recognition pass, are added to the lattice produced by recognition. The resulting lattice represents a set of alternative alignments of the acoustic data associated with an utterance.

The idea behind lattice-based MPE is not only to use the lattice as an approximation to the hypothesis space, but also to use the alignment information which is present in the lattice to save computation. Note that Equation 3 can be re-phrased as a sum over all possible alignments of the acoustic data as in Equation 7. A similar re-formulation for Equation 4 is given by Equation 8.

$$\gamma_s = \sum_{r=1}^R \sum_{z \in \mathcal{Z}_r} K(r, z, \theta) \sum_{t=1}^{T(r)} \gamma_s(t|z, \mathbf{o}_r, \theta) \quad (7)$$

$$\theta_s = \sum_{r=1}^R \sum_{z \in \mathcal{Z}_r} K(r, z, \theta) \sum_{t=1}^{T(r)} \gamma_s(t|z, \mathbf{o}_r, \theta) \mathbf{o}_t(r) \quad (8)$$

The set \mathcal{Z}_r comprises all possible alignments of the utterance \mathbf{o}_r and $K(r, z, \theta)$ is given by Equation 9.

$$K(r, z, \theta) = P(z | \mathbf{o}_r, \theta) [L_{\text{av}}(r) - L(w_z, \hat{w}_1^{M(r)})] \quad (9)$$

In the above equation, w_z is the hypothesis associated with alignment z . Notice also that the average error $L_{\text{av}}(r)$ may also be expressed as a sum over alignments as in Equation 10.

$$L_{\text{av}}(r) = \sum_{z \in \mathcal{Z}_r} P(z | \mathbf{o}_r, \theta) L(w_z, \hat{w}_1^{M(r)}) \quad (10)$$

Substituting the set of all possible alignments \mathcal{Z}_r with the set of alignments specified by the lattice, Equations 7 and 8 yield practical approximations to the statistics required for MPE model estimation. Further, since an alignment is a sequence of lattice arcs, Equation 7 can be expressed in terms of lattice arcs as in Equation 11. A similar rearrangement of Equation 8 in terms of lattice arcs may be performed.

$$\gamma_s = \sum_{r=1}^R \sum_{a \in \mathcal{A}_r} K(r, a, \theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_s(t|a, \mathbf{o}_r, \theta) \quad (11)$$

The symbol a represents a lattice arc which in turn represents a label, its start time a_{start} and end time a_{end} . The set \mathcal{A}_r contains all arcs in the lattice and $K(r, a, \theta)$ is expressed by Equation 12.

$$K(r, a, \theta) = p(a | \mathbf{o}_r, \theta) [L_{\text{av}}(r) - L(a, \hat{w}_1^{M(r)})] \quad (12)$$

In the above equation, $p(a | \mathbf{o}_r, \theta)$ is the posterior probability that arc a is included in any path, i.e. any contiguous sequence of arcs from the lattice start node to the lattice end node. The quantity $L(a, \hat{w}_1^{M(r)})$ is the posterior-weighted sum of the Levenshtein error of all the lattice paths which include arc a , while $L_{\text{av}}(r)$ is the posterior-weighted sum of the Levenshtein error of all the lattice paths.

Calculation of the Levenshtein distance between a path in the lattice and the reference label sequence $\hat{w}_1^{M(r)}$ is non-trivial. This involves a dynamic programming alignment of the label sequence associated with the path and the reference label sequence. Since a lattice encodes many such paths, calculation of the quantities $L(a, \hat{w}_1^{M(r)})$ and $L_{\text{av}}(r)$ becomes computationally expensive. This costly computation is avoided by approximating the Levenshtein distance between a lattice path and the reference label sequence. This approximation

is the focus of this paper. Section IV reviews previously-used approaches to error approximation. These approximations assign an error $l(a)$ to each lattice arc a such that the overall approximate error of each path is the sum of the errors associated with its composite arcs.

The state occupancies $\gamma_s(t|a, \mathbf{o}_r, \theta)$ of Equation 11 are calculated via a forward-backward pass over the models defined by each lattice arc a using the segment of acoustic data aligned with arc a . This is a standard forward-backward procedure, as implemented in the Baum-Welch algorithm. In order to calculate the quantities $K(r, a|\theta)$ for each arc a , a single lattice-level forward-backward pass suffices when an error $l(a)$ is assigned to each lattice arc. This forward-backward algorithm is introduced in [3].

B. Setting the learning rate

The choice of an appropriate learning rate D in Equation 2 has mainly been studied in the context of the CML (or MMI) discriminative criterion [3; 14; 15]. An occupancy-dependent scheme for determining the learning rate in the case of CML training is adopted for MPE training in [3]. The procedure is as follows. For each mixture component m :

- 1) Calculate D_m^{\min} , the minimum D required to ensure all variance updates are positive for component m .
- 2) Define γ_m^{den} as follows.

$$\gamma_m^{\text{den}} = \sum_r \sum_{a \in \mathcal{A}_r^{\text{den}}} K(r, a, \theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_s(t|a, \mathbf{o}_r, \theta) \quad (13)$$

where $\mathcal{A}_r^{\text{den}}$ denotes the subset of lattice arcs in the set of all lattice arcs \mathcal{A}_r for which $K(r, a, \theta)$ is negative. The symbols a_{start} and a_{end} denote the start and end time of arc a , respectively.

- 3) Set the learning rate D_m to $\max\{2D_m^{\min}, E\gamma_m^{\text{den}}\}$ where E is a configurable constant which typically assumes a value in the interval $[1, 2]$.

Since it has been shown to yield reasonably quick convergence of the MPE criterion, the above procedure used for calculating the learning rate in the experimental work of this paper.

C. Acoustic scaling and language model specificity

It has been noted [2] that use of LM scaling for the purposes of calculation of the posterior probabilities of a set of state sequences results in a posterior probability distribution which is sharply-peaked at the state sequence of maximal posterior probability. When using acoustic probability scaling, this posterior probability distribution is broader i.e. competing state sequences have larger posterior probabilities and it has been observed [2] that use of acoustic scaling leads to improved generalisation of discriminatively-estimated acoustic models.

Similarly, the use of lower-order LMs during training, for example a zero-gram or unigram, is another method of generating stronger (i.e. of higher posterior probability) competing hypotheses. It has been observed [16] that use of a unigram LM during training yields acoustic models which generalise better than those estimated using LMs of lower or higher

order. Acoustic scaling and unigram LMs are deployed in the experimental work of this paper. Note that acoustic scaling is applied to lattice arc acoustic likelihoods.

D. I-smoothing

Model parameters which optimise the MPE criterion often overfit the training data. The I-smoothing technique [3; 12] is used to alleviate this undesirable overfitting. In the case of the MPE criterion, a prior probability distribution over the acoustic parameters is subtracted from the criterion function.

When using I-smoothing, the ML estimates of state means and covariances are used to define the mode of this prior distribution. Using this prior, the I-smoothed MPE criterion is optimised using a weak sense auxiliary function in [3]. It is shown that the adjusted EBW formulae shown in Equation 14 can be used to optimise the I-smoothed MPE criterion.

$$\hat{\boldsymbol{\mu}}_s = \frac{\boldsymbol{\theta}_s + \frac{\tau^I}{\gamma_s^{ml}} \boldsymbol{\theta}_s^{ml} + D \boldsymbol{\mu}_s}{\gamma_s + \tau^I + D} \quad (14)$$

The quantities γ_s^{ml} and $\boldsymbol{\theta}_s^{ml}$ are given by Equations 15 and 16 respectively.

$$\gamma_s^{ml} = \sum_{r=1}^R \sum_{t=1}^{T(r)} \gamma_s(t|\hat{w}_1^{M(r)}, \mathbf{o}_1^{T(r)}, \theta) \quad (15)$$

$$\boldsymbol{\theta}_s^{ml} = \sum_{r=1}^R \sum_{t=1}^{T(r)} \gamma_s(t|\hat{w}_1^{M(r)}, \mathbf{o}_1^{T(r)}, \theta) \mathbf{o}_t(r) \quad (16)$$

IV. ERROR APPROXIMATION

Two types of approximation are used to avoid the need for a dynamic programming alignment of each label sequence in the lattice to the reference label sequence. The first type of approximation, discussed in Section IV-A, will be referred to as alignment-based error approximation. This type of approximation is the focus of this paper. A second type of approximation, referred to as lattice manipulation, is discussed in Section IV-B.

A. Alignment-based error approximation

The accuracy of a label sequence is approximated using aligned sequences in [3]. This approximation technique is shown in Figure 1. A path in a lattice defines what will be referred to as a hypothesis alignment. The most likely alignment of the reference transcription is called the reference alignment. Using the hypothesis and reference alignments, the accuracy of the hypothesis label sequence is approximated using the following procedure.

For each label q in the hypothesis alignment, the set of reference labels which overlap temporally with q is identified. Then, for each overlapping reference label z , the proportion of z which overlaps with q , $e(q, z)$, is calculated. Then the accuracy of q , $A(q)$, is given by Equation 17. The overall accuracy of the hypothesis is equal to the sum of the accuracies of the comprising labels.

$$A(q) = \max_z \left\{ \begin{array}{ll} -1 + 2e(q, z) & \text{if } q \text{ and } z \text{ same label} \\ -1 + e(q, z) & \text{if } q \text{ and } z \text{ different} \end{array} \right\} \quad (17)$$

Reference	A		B
Hypothesis	A	C	
Length (frames)	80	20	100
Overlap proportion $e(q,z)$	0.8	0.2	1.0
$\left\{ \begin{array}{l} \text{correct: } 2e(q,z) - 1 \\ \text{incorrect: } e(q,z) - 1 \end{array} \right\}$	0.6	-0.8	0.0
$A(q)$ (max of overlapping)	0.6	0.0	
Approximate accuracy	0.6		
Approximate error	1.4		
Levenshtein error	1		

Fig. 1. Alignment-based error approximation. Detailed in Section IV-A.

In the case of the hypothesis in Figure 1 the approximated overall accuracy of the hypothesis is 0.6. The actual accuracy is 1; the number of correct labels (1) minus the number of insertions (0).

The approximate Levenshtein distance $L_{\text{approx}}(\hat{w}_1^M, w_1^N)$ between the reference sequence \hat{w}_1^M and hypothesis sequence w_1^N is related to the approximate accuracy of w_1^N , $A_{\text{approx}}(\hat{w}_1^M, w_1^N)$, via Equation 18, where M is the number of labels in the reference sequence. This alignment-based error approximation will be referred to as the baseline approximate error.

$$L_{\text{approx}}(\hat{w}_1^M, w_1^N) = M - A_{\text{approx}}(\hat{w}_1^M, w_1^N) \quad (18)$$

The baseline approximate error is desirable because it circumvents the need for a dynamic programming step for each lattice path. However this error approximation has some limitations. Section V highlights several theoretical disadvantages of the approximated error and Section VI proposes alternative approaches to address some of these limitations. Note that this error approximation may be applied to sequences of words, phonemes or labels of other speech units.

As an alternative to the above approximation, a frame-level error metric is introduced in [17]. Improvements in the generalisation of the resulting models are reported.

B. Lattice manipulation

A technique is introduced in [18] which manipulates a lattice such that each label present in the lattice is aligned with zero or one labels of the reference transcription. Zero-length placeholders are used to manage insertions and deletions present in lattice label sequences. Errors between lattice paths and the reference transcription subsequently respect this alignment, reducing the complexity of error calculation. This lattice manipulation technique was used for MWE parameter estimation in [19].

In [5] the lattice topology is altered to enable efficient exact calculation of word errors. No significant improvement is noted when using the exact error for the purposes of MWE training. While these lattice manipulation techniques are not used in this work, they are mentioned for the sake of completeness.

V. LIMITATIONS OF BASELINE APPROXIMATE ERROR

A. Error overestimation

The baseline approximate error, described in Section IV, often overestimates the error of the hypothesis label sequence. This error overestimation occurs when the label alignment times of the hypothesis and reference alignments are not identical. Figure 1 shows how this overestimation arises. In the example illustrated the Levenshtein error is 1, a single substitution. However the approximate error, 1.4, is larger than the actual error because the end time of the first labels of the reference and hypothesis disagree.

B. Asymmetry

Figure 2 shows the baseline approximate error calculation when the reference and hypothesis alignments of Figure 1 are swapped. The approximate error is smaller due to the correct label of the hypothesis sequence overlapping with a larger portion of the reference label in this case. This example illustrates an asymmetry in the error approximation. This asymmetry is not present in the Levenshtein error metric, and is therefore difficult to motivate in an approximation to the Levenshtein metric. Since the Levenshtein error is symmetric, it is preferable for an approximation to retain this property.

Reference	A	C	
Hypothesis	A	B	
Length (frames)	80	20	100
Overlap proportion $e(q,z)$	1.0	0.17	0.83
$\left\{ \begin{array}{l} \text{correct: } 2e(q,z) - 1 \\ \text{incorrect: } e(q,z) - 1 \end{array} \right\}$	1.0	-0.83	-0.17
$A(q)$ (max of overlapping)	1.0	-0.17	
Approximate accuracy	0.83		
Approximate error	1.17		
Levenshtein error	1		

Fig. 2. Asymmetry of alignment-based error approximation. The reference and hypothesis label sequences of Figure 1 are swapped, resulting in a different approximate error.

C. Insertion to deletion bias

The asymmetry of the baseline approximate error is the source of an undesirable bias with regard to the approximation of insertion and deletion errors. Figure 3 shows the approximate error when an insertion error occurs while Figure 4 displays the approximate error when the scenario is reversed and a deletion error occurs. In both cases the Levenshtein error is 1, a single insertion or deletion. Note however that the approximate error is larger in the case of the insertion error. This example illustrates a bias in the error approximation resulting in a larger error being assigned to the sequence with an insertion.

VI. ALTERNATIVE ERROR APPROXIMATIONS

The asymmetries discussed in Section V can be addressed by using an error approximation called the frame error [17].

Reference	A		B	
Hypothesis	A	C	B	
Length (frames)	80	20	20	80
Overlap proportion $e(q,z)$	0.8	0.2	0.2	0.8
$\left\{ \begin{array}{l} \text{correct: } 2e(q,z) - 1 \\ \text{incorrect: } e(q,z) - 1 \end{array} \right\}$	0.6	-0.8	-0.8	0.6
$A(q)$ (max of overlapping)	0.6	-0.8		0.6
Approximate accuracy	0.4			
Approximate error	1.6			
Levenshtein error	1			

Fig. 3. Approximate error in case of insertion. Compare with Figure 4 which shows the approximate error in the case of a deletion.

Reference	A	C	B	
Hypothesis	A		B	
Length (frames)	80	20	20	80
Overlap proportion $e(q,z)$	1.0	0.5	0.5	1.0
$\left\{ \begin{array}{l} \text{correct: } 2e(q,z) - 1 \\ \text{incorrect: } e(q,z) - 1 \end{array} \right\}$	1.0	-0.5	-0.5	1.0
$A(q)$ (max of overlapping)	1.0			1.0
Approximate accuracy	2.0			
Approximate error	1.0			
Levenshtein error	1			

Fig. 4. Approximate error in case of deletion. The reference and hypothesis of Figure 3 have been swapped. A smaller approximate error is yielded in this case.

The frame error between two aligned label sequences is the number of frames at which the labels in the alignments differ, as shown in Figure 5. It is clear that this measure is symmetric since the frame error between two alignments is identical regardless of which alignment is the reference.

Reference	A		B	
Hypothesis	A	C	B	
Length (frames)	80	20	20	80
Frame error	0	20	20	0
Overall frame error	40			

Fig. 5. Frame error metric. The number of frames at which the aligned reference and hypothesis labels differ.

One limitation of the frame error approximation is that an incorrect label which spans a long period of time contributes more to the overall error than an incorrect label which spans a shorter period. A second limitation is illustrated in Figure 6, which shows a hypothesis containing an insertion error. However, since the inserted label agrees with the reference label in the region of overlap, no error is incurred when using the frame error metric and the error and the insertion error is overlooked. The first of these limitations can be addressed

Reference	A		B	
Hypothesis	A	A	B	
Overall frame error	0.0			

Fig. 6. Frame error metric fails to capture insertion error.

Reference	A	C	B		Overall error
Hypothesis	A		B		
Length (frames)	80	20	20	80	
Frame error	0	20	20	0	40
Reference normalised frame error	0	0.5	0.5	0	1.0
Hypothesis normalised frame error	0	0.2	0.2	0	0.4
Symmetrically normalised frame error	0	0.5	0.5	0	1.0

Fig. 7. Frame error normalisation in case of deletion error. In this case the reference and symmetrically normalised frame error approximations yield accurate estimates of the Levenshtein error of 1. The hypothesis normalised frame error underestimates the Levenshtein error.

via a normalisation scheme. The following section describes several such schemes.

A. Frame error normalisation

Figures 7 and 8 illustrate some different approaches to the normalisation of the frame error. Firstly the temporal region of each label of the hypothesis alignment is divided into segments corresponding to regions of overlap with different labels of the reference alignment. For example, in Figure 8 the label C is split into two segments, the first being associated with the label A of the reference alignment and the second associated with label B. The segment boundaries are illustrated by vertical dashed lines in Figures 7 and 8. Each segment has a corresponding frame error; the number of frames within the segment at which the hypothesis label differs from the label specified by the reference alignment. Then for each segment a normalisation factor is defined. The frame error for each segment is divided by this factor to yield a normalised frame error for each segment. The overall approximate error for the hypothesis is the sum of the normalised frame error over all segments.

The reference normalised frame error (RNFE) of Figures 7 and 8 is the result of defining the normalisation factor for a

Reference	A	C	B		Overall error
Hypothesis	A	C	B		
Length (frames)	80	20	20	80	
Frame error	0	20	20	0	40
Reference normalised frame error	0	0.2	0.2	0	0.4
Hypothesis normalised frame error	0	0.5	0.5	0	1.0
Symmetrically normalised frame error	0	0.5	0.5	0	1.0

Fig. 8. Frame error normalisation in case of insertion error. In this case the hypothesis and symmetrically normalised frame error approximations yield accurate estimates of the Levenshtein error of 1. The reference normalised frame error underestimates the Levenshtein error.

segment as the length (in frames) of the overlapping reference label. In the example of Figure 7, a deletion error, the normalisation factor is 40 for each of the segments corresponding to the label C of the reference transcription, the length of the reference label corresponding to these segments. This leads to an accurate approximation to the Levenshtein error of 1 in this case. However in the example of Figure 8, an insertion error, this normalisation method yields an underestimate (0.4) of the Levenshtein error. This is because the segments corresponding to the label C of the hypothesis transcription are normalised by 100, the length of the reference label corresponding to these segments.

The hypothesis normalised frame error (HNFE) defines the normalisation factor as the length of the overlapping hypothesis label. This method leads to accurate approximation of the insertion error of Figure 8 but an underestimate of the deletion error of Figure 7.

The third normalisation technique is to normalise the frame error of each segment by the length of the shorter of the overlapping labels. This method leads to an error approximation with the desirable symmetric property and yields accurate approximations for the deletion and insertion errors illustrated in Figures 7 and 8. This approximation is referred to as the symmetrically normalised frame error (SNFE).

These approximations to the Levenshtein error are expressed by Equation 19. The approximate error between hypothesis label sequence w_1^N and reference label sequence \hat{w}_1^M is denoted by $L_{\text{approx}}(w_1^N, \hat{w}_1^M)$, $\hat{\mathcal{A}}$ represents the set of aligned reference labels corresponding to sequence \hat{w}_1^M , \mathcal{A} is the set of aligned hypothesis labels corresponding to sequence w_1^N , and $l(a, \hat{a})$ is the normalised frame error between the aligned labels a and \hat{a} .

$$L_{\text{approx}}(w_1^N, \hat{w}_1^M) = \sum_{\hat{a} \in \hat{\mathcal{A}}} \sum_{a \in \mathcal{A}} l(a, \hat{a}) \quad (19)$$

The normalised frame error between two aligned labels $l(a, \hat{a})$ is defined by Equation 20.

$$l(a, \hat{a}) = \frac{e(a, \hat{a})}{n(a, \hat{a})} \quad (20)$$

The quantity $e(a, \hat{a})$ is the number of frames at which the aligned labels a and \hat{a} differ (defined as zero if no temporal overlap exists between the aligned labels). The divisor $n(a, \hat{a})$ is the normalisation term, defined as:

- the length of \hat{a} in the case of the reference normalised frame error.
- the length of a in the case of the hypothesis normalised frame error.
- the length of the shorter of a and \hat{a} in the case of the symmetrically normalised frame error.

B. Using multiple reference alignments

The SNFE described in Section VI-A addresses the asymmetry of the baseline approximate error. However the error overestimation limitation, illustrated in Figure 1, applies also to the SNFE. This overestimation is a consequence of differing alignment times in the hypothesis and reference label

sequences. This effect may be limited by using not only one, but multiple alignments of the reference label sequence and minimising the approximate error over this set of reference alignments. This set of multiple alignments is generated using a recogniser constrained to align only the reference word sequence.

Figure 9 illustrates this calculation. Two reference alignments and a single hypothesis alignment are shown. The SNFE of the hypothesis sequence is calculated with respect to each of the reference alignments to give a set of errors. The minimal symmetrically normalised frame error (MSNFE) is the minimum value in this set. In the example of Figure 9 the SNFE with respect to Reference1 is 1.4 and the SNFE with respect to Reference2 is 1.5. Therefore the MSNFE is 1.4.

Reference1	A		B		C		
Reference2	A	B		C			
Hypothesis	A		D		C		
Length (frames)	50	50	80	20	30		Overall Error
Symmetrically normalised frame error (1)	0		1.0	0.4	0		1.4
Symmetrically normalised frame error (2)	0	0.5	1.0		0		1.5
Minimum symmetrically normalised frame error							1.4
Approx. min. symmetrically normalised frame error	0		1.0		0		1.0

Fig. 9. Minimal symmetrically normalised frame error (MSNFE) and approximate minimal symmetrically normalised frame error (AMSNFE). MSNFE minimises the SNFE of the entire hypothesis sequence over the set of reference alignments. AMSNFE minimises the SNFE of each hypothesis label individually over the set of reference alignments.

In practice a large amount of reference alignments are encoded by a reference lattice. It is therefore impractical to enumerate this set of alignments and explicitly calculate the hypothesis error with respect to each one. So a further approximation is made. This approximation minimises the error of each hypothesis label over the set of all reference alignments (instead of minimising the error of the entire hypothesis label sequence over all reference alignments). This approximation avoids the need to explicitly enumerate all reference alignments. The last row of Figure 9 illustrates how this approximation is applied using the two reference alignments. For the initial hypothesis label A, the upper reference alignment Reference1 provides the minimal error. In the case of the second hypothesis label D, both reference alignments induce the same error. The error for the third hypothesis label C is minimised using the lower reference alignment Reference2. Summing the minimal error over all hypothesis labels gives the approximate minimal symmetrically normalised frame error (AMSNFE). Note that in the example illustrated in Figure 9 the AMSNFE of 1.0 is a closer approximation to the Levenshtein error (1) of the hypothesis sequence than the MSNFE of 1.4. This is a consequence of the flexibility of the former technique to select different reference alignments to minimise the error of different labels of the same hypothesis.

VII. EXPERIMENTAL SYSTEM

The system used for the experimental work of this paper is based upon the individual headset microphone (IHM) 2005 AMI meeting speech transcription system [20; 21]. Two recognition passes are used. The output of the first recognition pass is used in the second pass for unsupervised estimation of speaker-specific vocal tract length normalisation [22] as well as speaker-specific cepstral mean (CMN) and variance (CVN) normalisation [23]. Both passes use the same trigram language model and a language model scaling factor of 14.0. The 1-best output of the second pass is the recognised hypothesis. The pronunciation dictionary contains 50000 words and is based on the UNISYN lexicon [24].

The speech signal is represented using 39-dimensional features. In the first pass this feature vector comprises 13 Mel-frequency-based perceptual linear prediction coefficients (MF-PLP, [25]) and the first and second time derivatives of these features. In the second pass the feature vector is extracted via smoothed heteroscedastic linear discriminant analysis (SHLDA, [26]) from a 52-dimensional vector comprising 13 MF-PLP coefficients and their first, second and third time derivatives.

The training dataset comprises 104 hours of transcribed speech in meetings from a selection of corpora. The National Institute of Standards and Technology (NIST) 2005 conference meeting evaluation dataset *rt05seval* is used as test data¹. This comprises 1.9 hours of speech, 53 speakers and 24776 words (including hesitations, partial words and fillers).

A. MPE estimation details

The system used to evaluate MPE-estimated acoustic models in Section IX is based upon the ML system described above. The only difference is that the MPE-estimated acoustic models replace the ML-estimated models in the second recognition pass.

The ML-estimated models used in the second recognition pass of the baseline system are used as a starting point for MPE re-estimation. Using these models, a lattice (so-called denominator lattice) is generated for each training set utterance using a recognition pass and a bigram language model. The bigram language model is derived from the trigram used in recognition. A language model scaling factor of 13.0 and an insertion penalty of -10.0 are used at this stage.

The same ML models are again used in a recognition pass to generate a lattice (so-called numerator lattice) comprising the most likely alignments of the correct transcription for each training utterance. These lattices are referred to as the numerator lattices. The most likely alignment is chosen as the reference alignment in MPE parameter estimation. Note that the lattices are phoneme-marked to accommodate the optimisation of the MPE criterion. This means that each arc (which represents an aligned word) has an associated time-aligned sequence corresponding to the most likely phoneme alignment of the aligned word. The lattice can thus be considered as a sample of alignments corresponding to the phoneme-level hypothesis space.

The numerator and denominator lattices are rescored with a unigram language model (again, derived from the trigram used in recognition) to improve the MPE-estimated model generalisation (Section III-C). The denominator lattices are subsequently pruned to a maximum density of 200.0 arcs per second to reduce the computational cost of the MPE criterion optimisation process. The alignments present in the unigram numerator lattices are then added into the corresponding unigram denominator lattices to yield the final lattices used in MPE parameter re-estimation.

During parameter estimation, the probability distribution described by the unigram LM is further broadened by using an LM scaling factor of 0.5. An acoustic scaling factor of $\frac{1}{14}$ is used. The occupancy-dependent scheme for setting the learning rate (Section III-B) is used, and the constant E is set to 0.5.

VIII. ERROR APPROXIMATION ANALYSIS

The theoretical arguments presented in Sections V and VI are experimentally tested in this section. Approximately one hour of speech comprising 1851 utterances is selected from the training corpus of spontaneous meeting speech described in Section VII. Each utterance is recognised using the first and second pass of the recognition system described in Section VII to produce a phoneme-level hypothesis alignment for each utterance. The reference phoneme sequence and reference phoneme-level alignment for each utterance are produced by aligning the correct word sequence using the same recognition system. To measure the AMSNFE approximation, a lattice of reference alignments is also generated by the recognition system at this stage.

For each utterance, the phoneme-level Levenshtein error is calculated using dynamic programming alignment, and the phoneme-level approximate error is calculated using each of the techniques described in Sections V and VI.

A. Raw error approximation

From the dataset described above, the utterances which are transcribed with non-zero phoneme-level error are selected. This subset is then further split into three smaller subsets, one which contains those utterances transcribed with substitution errors only (\mathcal{S}), one which contains those utterances transcribed with insertion errors only (\mathcal{I}) and one which contains those utterances with deletion errors only (\mathcal{D}).

Table I presents some analysis of the approximate error techniques. Each row of Table I corresponds to the subset of utterances indicated in the first column, where the notation of the previous paragraph has been used. The number of utterances in each subset is displayed and the sum of the Levenshtein errors of each transcribed utterance are indicated in the second and third columns respectively. The remaining columns display the sum of the approximate errors of each transcribed utterance in each set for each approximation technique. The acronyms BAE and FE are used to represent the baseline approximate error and the frame error, respectively.

The results presented in Table I are used to test the theoretical arguments presented earlier. Firstly, it is clear that

¹Details of these datasets are found at <http://www.nist.gov/speech/tests/rt/>.

Set	# Utt	Lev	Approximate error					
			BAE	FE	HNFE	RNFE	SNFE	AMSNFE
\mathcal{S}	110	245	560.9	3211	345.4	359.1	426.4	290.8
\mathcal{I}	43	82	208.2	1317	115.8	70.9	128.7	80.3
\mathcal{D}	184	327	552.0	2297	193.9	340.5	370.2	181.2

TABLE I

Error approximations for substitution, insertion and deletion errors.

the baseline approximate errors are greater than the true Levenshtein errors for each utterance subset. This confirms that the BAE approximation overestimates errors, as discussed in Section V-A. Note further that the ratio of the BAE to the Levenshtein error is 2.54 (208.2/82) in the case of insertion errors, and 1.68 (552/327) in the case of deletion errors. The higher ratio in the case of insertions is evidence of the insertion to deletion bias discussed in Section V-C.

Examination of the same ratios in the case of the FE approximation reveals that the FE also possesses an insertion to deletion bias. The ratio is larger (16.1 (1317/82)) in the case of insertion errors than in the case of deletions (7.02 (2297/327)). This somewhat surprising result is due to the fact that, on average, for the dataset considered, deletion errors correspond to phonemes of shorter duration than phonemes associated with insertion errors.

In concurrence with the arguments of Section VI-A, it is clear from Table I that the RNFE approximation underestimates insertion errors (an approximation of 70.9, in comparison with a true insertion error of 82), while the HNFE approximation underestimates deletion errors (an approximation of 193.9, in comparison with a true deletion error of 327).

Again, in accordance with the discussion of Section VI-A, the results of Table I show that the SNFE does not underestimate either deletion or insertion errors. The ratio of the SNFE approximation to true Levenshtein error is 1.57 (128.7/82) in the case of insertions and 1.13 (370.2/327) in the case of deletions, showing that some insertion to deletion bias is still present even when using the SNFE. This effect is due to the fact that insertion errors sometimes correspond to phonemes of longer duration than the overlapping phonemes of the reference transcription. Such insertion errors are then normalised by the length of a relatively short reference phoneme, resulting in an overestimation of the insertion error. Although the analogous phenomenon occurs also in the case of deletion errors (in this case, normalisation by the length of a relatively short hypothesis phoneme occurs), it happens less often than in the case of insertions. This in turn is because, on average, for the dataset considered, deletion errors correspond to phonemes of shorter duration than phonemes associated with insertion errors.

Comparing the entries in the last and second last columns of Table I, it can be seen that the AMSNFE approximation always yields a lower value than that of the SNFE approximation. However, a large underestimation of deletion errors occurs in this case. Examination of a typical set of reference alignments provides an explanation of this phenomenon. The set of reference alignments often contains two alternative alignments for a word, one of which contains some word-end silence

and the other which contains little or none, as shown in Figure 10. For the dataset considered, the most likely (i.e. the 1-best) alignment is usually the reference which contains silence. When using the SNFE approximation, the most likely alignment is used as the reference.

Reference1 (more likely)	h	ay	sil
Reference2	h	ay	
Hypothesis	h	ay	
Length (frames)	80	50	30
Symmetrically normalised frame error (1)	0	0	1
Symmetrically normalised frame error (2)	0	0	0
Approx. min. symm. norm. frame error	0	0	0

Fig. 10. Silence handling using the approximate minimal symmetrically normalised frame error.

In the example of Figure 10, the SNFE of the hypothesis is 1 due to the deletion of a silence label. When using multiple reference alignments, as in the case of the AMSNFE approximation, a hypothesis which erroneously deletes word-end silence may be assigned zero error due to the presence of the additional, less likely, reference alignment. In the example of Figure 10, the AMSNFE approximation assigns zero error to the hypothesis due to the presence of the reference alignment Reference2. This accounts for the underestimation of deletion errors witnessed in the case of the AMSNFE approximation.

B. Error approximation accuracy

The accuracy of the error approximations described in Sections V and VI is now measured. Note that there exist several ways to quantify this accuracy, for example the summed squared difference between the approximate errors and the true errors. In this section, the correlation between the true Levenshtein error and the approximate error is used as a measure of the accuracy of the approximation [27]. This correlation is used because it is a measure of the similarity between the MPE criterion which uses the error approximation (the approximate MPE criterion, say) and the MPE criterion which uses the true Levenshtein error (the true MPE criterion, say). An approximation which correlates perfectly with the Levenshtein error yields an approximate MPE criterion which, in turn, yields the same model parameter updates as the true MPE criterion.

To measure the correlation of the approximate error with the true Levenshtein error, the entire 1851-utterance subset of the training corpus described above is used. The correlation of the phoneme-level Levenshtein and approximate errors is then calculated for each of the error approximations. Table II records this correlation for the different error approximations discussed above.

It is firstly worth noting that the differences between all pairs of correlation coefficients displayed in Table II are all significant. A significant difference between two correlation coefficients is indicated by the procedure described in [28] for correlation coefficients derived from the same sample.

Approximation method	Correlation coefficient
Frame error	0.909
Baseline approximate error	0.959
Reference normalised frame error	0.968
Hypothesis normalised frame error	0.938
Symmetrically normalised frame error	0.976
Approx. min. symm. normalised frame error	0.986

TABLE II

Correlation of error approximations with Levenshtein error.

This procedure assumes that the samples are independently selected, a reasonable assumption for the dataset considered here.

Table II shows that the frame error displays weaker correlation with the Levenshtein error than all other approximations. This observation indicates that the unnormalised frame error is a relatively inaccurate approximation of the Levenshtein error.

The RNFE has a higher correlation than the HNFE approximation. This result is not immediately understandable, but is due to the nature of the ASR system used to provide hypotheses in this evaluation. The ASR system is designed to output more deletion errors than insertion errors. As argued in Section VI-A, the RNFE captures deletion errors more effectively than the HNFE approximation. Consequently, the RNFE is a better approximation of the Levenshtein error for this particular dataset.

The SNFE approximation displays a higher correlation with the Levenshtein error than both the hypothesis and reference normalised frame error. So, with respect to this measurement of accuracy, the SNFE is a more accurate approximation of the Levenshtein error than both the RNFE and HNFE.

Despite its underestimation of deletion errors, as discussed in Section VIII-A, the AMSNFE metric yields a correlation coefficient of 0.986, greater than that of the SNFE metric. This result suggests that incorporation of knowledge of multiple reference alignments improves the SNFE approximation of the Levenshtein distance.

IX. EVALUATION: MPE-ESTIMATED ACOUSTIC MODELS

The impact of each of the error approximations described in Section VI upon the behaviour of MPE-estimated acoustic models is now measured. Section VII-A described the recognition system used to evaluate the MPE-estimated acoustic models and the MPE parameter estimation procedure. To evaluate the effect of the different error approximations, the MPE-estimated acoustic models are substituted into the second recognition pass of the transcription system. The *rt05seval* NIST evaluation dataset is used as test data.

A. Unsmoothed MPE

Figure 11 displays how the recognition WER yielded by the MPE-estimated models varies with each iteration of parameter estimation. No smoothing of the MPE criterion is performed. As shown in the legend of Figure 11, each plotted curve corresponds to the use of a different error approximation within the MPE implementation. The zeroth iteration corresponds to the performance of the baseline ML-estimated models.

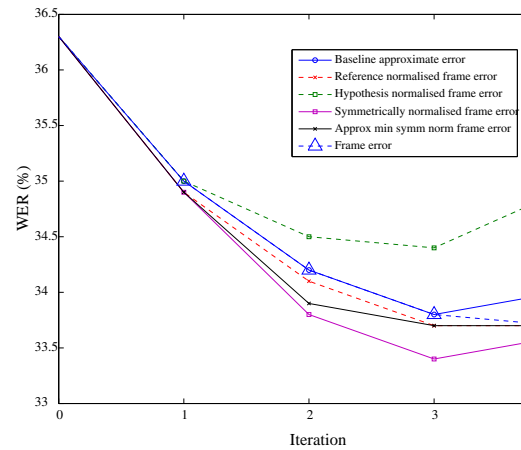


Fig. 11. Performance of unsmoothed MPE-estimated models using different error approximations (*rt05seval* dataset). The error approximation used is indicated in the legend.

All of the error approximations used result in MPE re-estimated models which display improved classification performance over the ML-estimated models. The models which generalise best are generally yielded after three iterations of MPE re-estimation after which the effects of overfitting the training data become evident. Table III provides some analysis of the errors made by the models yielded after three MPE iterations.

1) *Discussion:* The initial row of Table III displays the error analysis for the ML-estimated models used as the starting point for MPE estimation. The remaining rows analyse the errors committed by MPE-estimated models. Note firstly that, with the exception of the RNFE approximation method, all MPE-estimated models display deletion rates greater than that of the ML-estimated models. This is because, as discussed in Section VIII-A, for the dataset considered, an insertion to deletion bias is manifested when using all error approximations other than the RNFE approximation.

The analysis of Table III shows that the RNFE approximation technique yields the least number of deletion errors but the greatest number of insertion errors of all the error approximations used. This is due to the underestimation of insertion errors when using the RNFE technique, as explained in Section VI-A.

The HNFE method gives rise to acoustic models which yield the highest deletion rate and lowest insertion rate of all the approximations considered. This is due to the systematic underestimation of deletion errors when using this technique, again explained in Section VI-A.

The BAE technique yields models which have a relatively high deletion rate and low insertion rate. This is explained by the insertion to deletion bias discussed in Section V-C.

The use of the SNFE approximation is inherently unbiased with regard to the correction of insertion and deletion errors. Moreover, this technique yields WER improvements over all other approximation methods including the baseline error approximation. A statistical test reveals that these improvements

Criterion	Approx.	Sub (%)	Del (%)	Ins (%)	WER (%)
ML	-	18.1	13.6	4.7	36.4
MPE	FE	15.9	14.3	3.7	33.8
	BAE	15.7	14.6	3.5	33.8
	RNFE	16.2	13.5	4.1	33.7
	HNFE	16.0	15.2	3.2	34.4
	SNFE	15.6	14.2	3.6	33.4
	AMSNFE	15.7	14.4	3.6	33.7

TABLE III

Performance analysis of models yielded after three iterations of unsmoothed MPE estimation (rt05seval dataset).

are significant. Note that a significant improvement is defined as significant at the 95% confidence level using the matched pairs sentence segment word error test (MPSSWE, [29; 30]).

Despite having displayed a higher correlation with the Levenshtein error (see Table II), the AMSNFE technique is significantly out-performed in terms of WER by the SNFE method. This is due to the underestimation of deletion errors when using the AMSNFE technique, as discussed in Section VIII-A. This is reflected in the comparatively high deletion rate of the AMSNFE technique, as displayed in Table III.

B. I-smoothed MPE

To investigate if the classification performance improvements yielded by the SNFE approximation over the baseline approximate error persist when using smoothed MPE criteria, the experiment described above is repeated using I-smoothed MPE (see Section III-D). An I-smoothing factor (τ^I) equal to 50 is used.

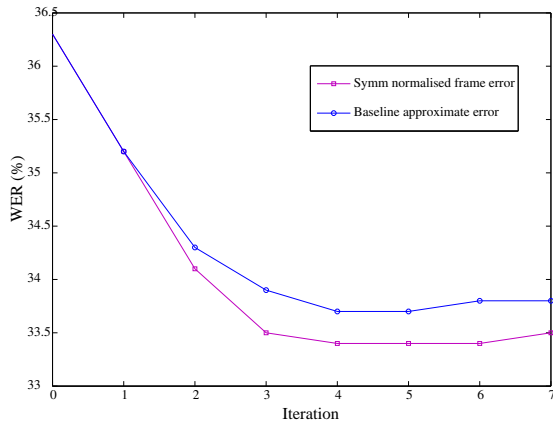


Fig. 12. Performance of I-smoothed MPE-estimated models using different error approximations (rt05seval dataset). The error approximation used is indicated in the legend and the I-smoothing factor τ^I is set to 50.

Approx.	Sub (%)	Del (%)	Ins (%)	WER (%)
BAE	15.8	14.6	3.3	33.7
SNFE	15.7	14.3	3.4	33.4

TABLE IV

Performance analysis of models yielded after five iterations of I-smoothed MPE estimation (rt05seval dataset). An I-smoothing factor τ^I of 50 is used.

Figure 12 displays how the recognition WER yielded by the smoothed MPE-estimated models varies with each iteration of parameter estimation in the cases of the baseline approximate error and SNFE approximations. Again, the zeroth iteration corresponds to the performance of the baseline ML-estimated models. Table IV analyses the errors made by the models yielded after five smoothed MPE iterations in the case of each error approximation technique.

1) Discussion: Although the effects of overfitting the training data are alleviated in later iterations (compare the performance of the fourth iteration models in the smoothed and unsmoothed cases), the performance of the fifth iteration models in the smoothed case is similar to the performance of the third iteration models in the unsmoothed case.

Examination of Table IV reveals that, as in the case of unsmoothed MPE model estimation, the baseline approximate error technique yields models which have a relatively high deletion rate and relatively low insertion rate. As in the case of unsmoothed MPE model estimation, the SNFE approximation yields models which in turn yield a lower WER than those models estimated using the baseline error approximation. The MPSSWE test reveals that the WER improvements yielded by models estimated with the SNFE approximation are significant.

X. SUMMARY AND FUTURE WORK

Limitations of a previously introduced alignment-based error approximation technique have been highlighted in this paper. Alternative approximations based on the frame error metric have been introduced. The accuracy of these approximations has been evaluated, as well as their impact upon the performance of MPE acoustic model re-estimation. Significant improvements over the previously introduced error approximation have been recorded for a large vocabulary recognition task when the symmetrically normalised frame error approximation is deployed for MPE acoustic parameter re-estimation.

Future work should compare use of the approximate methods introduced in this paper with lattice manipulation approaches [5; 19] and the minimum phone frame error introduced in [17].

REFERENCES

- [1] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Speech and Audio Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.
- [2] P. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–37, 2002.
- [3] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2003.
- [4] J. Kaiser, B. Horvat, and Z. Kacic, "Overall risk criterion estimation of hidden Markov model parameters," *Speech Communication*, vol. 38, no. 3-4, pp. 383–398, 2002.

- [5] G. Heigold, W. Macherey, R. Schluter, and H. Ney, "Minimum exact word error training," in *Proceedings ASRU*, 2005, pp. 186–190.
- [6] K. Na, B. Jeon, D. Chang, S. Chae, and S. Ann, "Discriminative training of hidden Markov models using overall risk criterion and reduced gradient descent method," in *Proceedings Eurospeech*, Madrid, 1995, pp. 97–100.
- [7] P. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "A generalization of the Baum algorithm to rational objective functions," in *Proceedings ICASSP*, Glasgow, 1989, pp. 631–634.
- [8] Y. Normandin, "Hidden Markov models, maximum mutual information estimation and the speech recognition problem," Ph.D. dissertation, McGill University, 1991.
- [9] D. Povey, M. Gales, D. Y. Kim, and P. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Proceedings Eurospeech*, Geneva, Switzerland, 2003, pp. 1981–1984.
- [10] D. Kanevsky, "Extended Baum transformations for general functions," in *Proceedings ICASSP*, vol. 1, 2004, pp. 821–824.
- [11] M. Gibson, "Minimum Bayes risk acoustic model estimation and adaptation," Ph.D. dissertation, Sheffield University, 2008.
- [12] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings ICASSP*, Orlando, Florida, 2002, pp. 105–108.
- [13] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book for HTK version 3.2.1*, 2003.
- [14] V. Valtchev, J. Odell, P. Woodland, and S. Young, "Mmie training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.
- [15] R. Schluter, W. Macherey, S. Kanthak, , H. Ney, and L. Welling, "Comparison of optimization methods for discriminative training criteria," in *Proceedings Eurospeech*, 1997, pp. 15–18.
- [16] R. Schluter, B. Muller, F. Wessel, and H. Ney, "Interdependence of language models and discriminative training," in *Proceedings IEEE ASRU Workshop*, Keystone, Colorado, 1999, pp. 119–122.
- [17] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proceedings Interspeech*, Lisbon, Portugal, 2005, pp. 2125–2128.
- [18] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum Bayes-risk decoding for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 234–249, 2004.
- [19] V. Doumpiotis and W. Byrne, "Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition," *Speech Communication*, vol. 48, no. 2, pp. 142–160, 2005.
- [20] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordeman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," *Proceedings MLMI*, 2005.
- [21] T. Hain, L. Burget, J. Dines, I. McCowan, G. Garau, M. Karafiat, M. Lincoln, D. Moore, V. Wan, R. Ordeman, and S. Renals, "The development of the AMI system for the transcription of speech in meetings," *Proceedings MLMI*, 2005.
- [22] L. Lee and R. Rose, "Speaker normalisation using efficient frequency warping procedures," in *Proceedings ICASSP*, vol. 1, Atlanta, 1996, pp. 353–356.
- [23] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 6, pp. 1304–1312, 1974.
- [24] S. Fitt, "Documentation and user guide to UNISYN lexicon and post-lexical rules," Centre for Speech Technology Research, Edinburgh University, Tech. Rep., 2000.
- [25] P. Woodland, M. Gales, D. Pye, and S. Young, "Broadcast news transcription using HTK," in *Proceedings ICASSP*, vol. 2, Munich, 1997, pp. 719–722.
- [26] L. Burget, "Combination of speech features using smoothed heteroscedastic linear discriminant analysis," in *Proceedings Interspeech*, Jeju Island, Korea, 2004, pp. 2549–2552.
- [27] F. Wessel, R. Schluter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," in *Proceedings ICASSP*, Salt Lake City, Utah, 2001, pp. 33–36.
- [28] J. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological Bulletin*, vol. 87, pp. 245–251, 1980.
- [29] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings ICASSP*, Philadelphia, 1989, pp. 532–535.
- [30] D. S. Pallett, W. M. Fisher, and J. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proceedings ICASSP*, Albuquerque, 1990, pp. 97–100.



Matthew Gibson received the B.Sc(Hons) in 1994 from Glasgow University, M.Sc. in 1996 from Oxford University, M.Phil. in 2004 from Cambridge University and PhD in 2008 from Sheffield University. He is currently a research associate at Cambridge University engineering department. His main research interests are machine learning, automatic speech recognition and speech synthesis.



Thomas Hain holds the degree Dipl.-Ing. in EEE from the University of Technology, Vienna and a PhD from Cambridge University. After working at Philips Speech Processing he joined the Speech, Vision and Robotics group at Cambridge University Engineering Department. He currently is a Senior Lecturer at the Department of Computer Science at Sheffield University. His research interests are in speech and audio processing, machine learning and optimisation. He is a member of the IEEE SLT Committee and the Editorial Board of CSL.