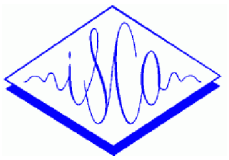

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# Complementary Phone Error Training

*F. Diehl & P.C. Woodland*

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PW U.K.
{fd257, pcw}@eng.cam.ac.uk

## Abstract

This paper introduces a novel method for the training of a complementary acoustic model with respect to set of given acoustic models. The method is based upon an extension of the Minimum Phone Error (MPE) criterion and aims at producing a model that makes complementary phone errors to those already trained. The technique is therefore called Complementary Phone Error (CPE) training. The method is evaluated using an Arabic large vocabulary continuous speech recognition task. Reductions in word error rate (WER) after combination with a CPE-trained system were obtained with up to 0.7% absolute for a system trained on 172 hours of acoustic data and up to 0.2% absolute for the final system trained on nearly 2000 hours of Arabic data.

**Index Terms**: Speech recognition, acoustic model training, discriminant training, complementary models, system combination

10.21437/Interspeech.2012-499

## 1. Introduction

State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems often use system hypothesis combination techniques between sub-systems. The hypotheses from these sub-systems are then combined with techniques such as ROVER [11].

A key issue is how to design the recognition sub-systems to give the largest improvements from combination. Typically the system designer will choose a range of different system attributes such as the use of different acoustic features, different acoustic modelling units, different tokenizations or different acoustic model training paradigms, see for example [1]. However the success of such methods relies on expert knowledge from a designer and the complementary outputs are a side-effect of design decisions rather than being explicitly built in to the model training procedure.

Hence there has been interest in ways to construct complimentary recognition systems in a more systematic way. Approaches to this problem have included the use of randomised or directed decision trees [2, 3], and boosting schemes for building the hidden Markov model (HMM) based acoustic models.

Boosting is a technique which aims to construct a strong classifier by combining a set of consecutively built complementary weak classifiers. Recent progress reported in boosting ASR systems has focused on frame-level weighting schemes during the re-estimation of HMMs [4, 5, 6, 7]. A notable exception is [8] which also includes the rebuilding of the decision trees used in state clustering.

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

The present work proposes a novel method for the construction of a complementary acoustic model given a set of previously trained acoustic models. The method is an extension of standard Minimum Phone Error (MPE) training [9]. Standard MPE maximises the expected phone accuracy with respect to a gold-standard reference. However in Complementary Phone Error (CPE) training the expected accuracy with respect to alternative reference outputs produced by pre-existing systems is also simultaneously minimised. Therefore the resulting system will have a low error rate but will tend to produce different errors to the pre-existing systems.

The paper is organised as follows. In Section 2 a more detailed description of CPE is given, and some comments on the relationship to boosting are discussed. In Section 3 the experimental setup is described using an Arabic broadcast transcription task. Initially experiments are given with a 172 hour training set for fast turn-around. Later experiments use the full 1831 hour training set used in the full Cambridge Arabic LVCSR system.

## 2. Complementary phone error training

To train a complementary acoustic model $\mathcal{M}_\mathcal{C}$ given one or more pre-existing acoustic models $\mathcal{M}_i$, we propose to use a modified MPE objective function defined as follows

$$\mathcal{F}_{\mathcal{M}_\mathcal{C}}(\boldsymbol{\lambda}) = \mathbb{E}\left[\alpha_0 \mathcal{A}(\boldsymbol{w}, \boldsymbol{w}_{\text{ref},0}^{(r)}) - \sum_i \alpha_i \mathcal{A}(\boldsymbol{w}, \boldsymbol{w}_{\text{ref},i}^{(r)})\right] \quad (1)$$

The $\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}_{\text{ref},i}^{(r)})$ terms denote phone accuracies and the interpolation weights $\alpha_i$ are positive scalars (and typically sum to one). The phone accuracies $\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}_{\text{ref},i}^{(r)})$ in the CPE objective function (1) are defined with respect to the true reference transcription $\boldsymbol{w}_{\text{ref},0}$ (as for standard MPE [9]) for the first term. For the subsequent terms the phone accuracy is defined with respect to phone level transcripts produced by pre-existing systems $\boldsymbol{w}_{\text{ref},i}^{(r)}$. Standard MPE is obtained by setting $\alpha_0 = 1$ and all other $\alpha_i = 0$.

As for standard MPE training, the expectation in the CPE objective function is taken with respect to the scaled sentence posterior probability $P^\kappa(\boldsymbol{w} \mid \boldsymbol{O}^{(r)}; \boldsymbol{\lambda})$ which is conditioned by the $r^{th}$ training observation sequence $\boldsymbol{O}^{(r)}$. The expectation sums over all possible transcription hypotheses $\boldsymbol{w}$ (typically represented by a lattice) and observations $\boldsymbol{O}^{(r)}$ and CPE training consists of maximising $\mathcal{F}_{\mathcal{M}_\mathcal{C}}(\boldsymbol{\lambda})$ with respect to the model parameters $\lambda$.

The phone accuracies are defined between a transcription hypothesis $\boldsymbol{w}$ and a reference transcription $\boldsymbol{w}_{\text{ref},i}^{(r)}$. However, in contrast to the standard MPE case where only one reference transcription, the so-called *ground truth* is used, the CPE objective function is defined with respect to two or more ref-

erence transcriptions. Though the first of the accuracy terms, $\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}^{(r)}_{\text{ref},0})$, is always taken with respect to the ground truth $\boldsymbol{w}^{(r)}_{\text{ref},0}$, the other terms are contributions from so-called *cross-systems* for which a complementary system is to be estimated. That is, the *cross-reference* $\boldsymbol{w}^{(r)}_{\text{ref},i}$ denotes the transcription which is obtained by decoding the acoustic training material by the $i^{th}$ *cross-system*.

In order to understand the motivation behind the CPE objective function, it is useful to consider the standard MPE criterion. MPE training, which maximises the expected phone accuracy (or minimises the expected phone error rate) between the reference transcriptions and possible alternative transcriptions, can be regarded as estimating parameters such that the target system is as close as possible to a hypothetical error-less reference system. The phone accuracy for each training utterance is needed and it is obtained by comparing the reference transcription (at each iteration) to a set of reasonable alternative transcriptions (typically stored in a lattice). These training lattices are obtained by decoding the training material by an initial version of the target system (often the result of maximum likelihood training).

An objective function which generates a complementary target system $\mathcal{M}_{\mathcal{C}}$ to one or more pre-existing systems $\mathcal{M}_i$ would therefore replace the hypothetical error-less reference system by the existing cross-systems $\mathcal{M}_i$ for which a complementary system $\mathcal{M}_{\mathcal{C}}$ is to be designed. Correspondingly, the ground truth reference transcriptions are replaced by a set of cross-reference transcriptions which are obtained by decoding the training material by the cross-systems $\mathcal{M}_i$. The phone accuracies $\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}^{(r)}_{\text{ref},i})$ calculated between the target system transcriptions $\boldsymbol{w}$ and the cross-system references $\boldsymbol{w}^{(r)}_{\text{ref},i}$ therefore provide a measure of the similarity of these systems. Therefore, the basic idea for the design of a training criterion for a complementary target system $\mathcal{M}_{\mathcal{C}}$ consists in minimizing the expected cross-system phone accuracies $\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}^{(r)}_{\text{ref},i})$ with respect to the model parameters $\lambda$. Such a training criterion will obviously aim to force the target system $\mathcal{M}_{\mathcal{C}}$ to produce a different phone sequence output when compared to the cross-systems $\mathcal{M}_i$.

However, it should also be noted that for an effective complementary system the error rate of each system should be broadly similar. Therefore as well as maximising the difference to other cross-system outputs, it is also necessary to simultaneously maximise the expected phone accuracy with respect to the ground truth reference transcription. The weighted combination of the standard MPE objective function along with the minimisation of the expected cross-system accuracies gives rise to the CPE objective function given in (1). Given that the weighting parameters $\alpha_i$ defined as positive, the minus signs assigned to the cross phone accuracies serve to convert the minimisation of the associated expectation for the cross phone accuracies to a maximisation which can therefore be used for all the terms in the CPE objective function. This then allows the extended Baum-Welch algorithm to be used in parameter estimation as for standard MPE [9].

### 2.1. Relationship to Boosting

The iterative application of the CPE criterion shares a number of attributes with frame-level boosting. Both techniques construct one system based on a pre-existing system and also perform a data weighting. The weights are obtained by comparing the current system with the pre-existing system and the choice of weights are the key to ensure the desired system diversity.

While there are similarities between CPE and boosting, there are also conceptual differences. The key difference consists in CPE encoding the generation of a complementary model directly in its discriminative optimisation criterion. As a consequence the resulting weights are conditioned by the phone-arcs of the competing hypotheses for the acoustic data. Thus, at each time instance there are several weights active, one for each phone instance within the hypothesis space. This is in contrast to the boosting schemes mentioned in section 1. There the weights are conditioned by the frames of the acoustic data. That is, at each time instance just one weight applies.

## 3. Experimental setup

In this paper, the CPE criterion is investigated within a two-system setup. The sum of cross phone accuracies in expression (1) reduces therefore to just one term from the system $\mathcal{M}$ for which a complementary system $\mathcal{M}_{\mathcal{C}}$ should be constructed. We obtain thus the two-system CPE objective function (2)

$$\mathcal{F}_{\mathcal{M}_{\mathcal{C}}}(\boldsymbol{\lambda}) = \mathbb{E}\left[(1-\alpha)\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}^{(r)}_{\text{ref}}) - \alpha\mathcal{A}(\boldsymbol{w}, \boldsymbol{w}^{(r)}_{\text{xref}})\right] \quad (2)$$

where the interpolation weights were set to $1-\alpha$ and $\alpha$ and the reference transcriptions are just marked 'ref' for the transcription from the reference system and 'xref' for the cross-reference transcription from the cross-system $\mathcal{M}$.

In addition, in this paper we concentrate on the design of complementary models within one training configuration. First, a standard MPE system is trained. Next, the CPE criterion is applied with the MPE system used as the cross-system to generate the cross-references $\boldsymbol{w}^{(r)}_{\text{xref}}$. Therefore this paper investigates if the CPE criterion is able to find a different, complementary solution to an MPE optimisation problem although keeping the structure of the acoustic model constant as well as not changing the training data.

The baseline for all experiments is a system built using MPE training. For this MPE system and all CPE systems, the lattices representing the hypothesis space $\boldsymbol{w}$ are kept the same and are obtained by decoding the training material by the corresponding maximum likelihood (ML) system. The ML model also served as the initial model for both the MPE and the CPE training. The cross-references $\boldsymbol{w}^{(r)}_{\text{ref},i}$ were obtained by decoding the training material by the MPE system. This involved the same weak bi-gram language model and decoder settings as used for generating the transcription hypotheses $\boldsymbol{w}$.

Two sets of experiments were carried out. The first set used 172 hours of acoustic training data to allow for rapid turnaround times for basic development purposes. The second set of experiments used 1831 hours of training data and a more sophisticated acoustic model design. Both test series are based on the Arabic system environment developed by Cambridge University within the DARPA GALE program.

All models use a PLP-based front-end with a 39-dimensional feature vector after an HLDA transform. Cross-word decision-tree state clustered triphones were built using the acoustic training data. For simplicity, graphemes were used as the acoustic units in all experiments. The 172 hour system used 6k tied states and an average of 16 Gaussian components per state (about 7.5M model parameters). The 1831 hour system applied a more sophisticated acoustic model setup. The feature vector was augmented by multilayer perceptron (MLP) features giving a features vector dimension of 65 [1]. Position dependent acoustic modelling was used [10] and the number of tied

states as well as the average number of Gaussian components per state was increased to 12k and 36, respectively. This 7.5-fold increase in the number of model parameters to 56.2M goes along with a 10-fold increase of acoustic training material.

Testing for both configurations was performed within the same unadapted decoding framework. It involved lattice generation with a trigram language model (LM) followed by confusion network (CN) decoding. The LM and the associated dictionary used a 350k word list and 1.2G tokens of LM training material. The systems were evaluated on two test sets defined by the GALE project: dev08 (3.04 hours) and dev09[1] (2.93 hours). The out-of-vocabulary (OOV) rates for these test sets were in the range of 1–2%.

### 3.1. The 172 hour system

Initial experiments were carried out on the 172 hour acoustic training corpus. For testing only dev08 was used. Column *CN-WER* of Table 1 compares the WER of the MPE system with five CPE trained systems for varying $\alpha$.

| System | $\alpha$ | CN-WER dev08 | cross-WER dev08 |
|---|---|---|---|
| **MPE** | - | 21.6 | - |
| **CPE** | 0.1 | 21.6 | 3.5 |
| **CPE** | 0.2 | **21.5** | 5.1 |
| **CPE** | 0.3 | 21.8 | 7.7 |
| **CPE** | 0.4 | 22.7 | 11.0 |
| **CPE** | 0.5 | 28.4 | 22.1 |

Table 1: CN decoding results (CN-WER) and cross-scoring results (cross-WER) for the MPE system and various CPE systems after the first CPE iteration, WER in %. Cross-scoring scores the CPE systems against the MPE system.

From Table 1 it is interesting to note that for $\alpha = 0.1$ and $\alpha = 0.2$ the system WER is either the same as or lower than that for MPE. For $\alpha = 0.4$ the performance of the CPE system starts to degrade. However, for $\alpha = 0.3$ the system performance is still very close to the MPE performance, with a WER increase of just 1% relative. Given that the WERs of the CPE systems hardly change between $\alpha = 0.1$ and $\alpha = 0.3$ the question arises how effective the CPE criterion is in generating alternative decoding hypotheses. To probe this question, the CPE hypotheses were compared to the MPE hypotheses (one taken as a reference and the other as a hypothesis in normal WER scoring). These cross-system scoring results are given by the *cross-WER* column of Table 1.

For the $\alpha$-values 0.1, 0.2 and 0.3 the cross-scoring results range between 3.5–7.7% WER whereas the WER increases by only 0.2% compared to the MPE results. Thus it appears that the CPE criterion does in fact introduce complementarity into the CPE systems while maintaining the overall WER close to that of the MPE system. However, the key test is to find how well the different hypotheses combine. The hypotheses annotated with their confidence scores were combined with ROVER. Table 2 gives these results.

Table 2 shows small but consistent reductions in WER for $\alpha$ in the range from 0.1–0.3%. The largest reduction in WER is obtained for $\alpha = 0.3$ and is about 1.5% relative. For larger $\alpha$ values the ROVER performance degrades. These findings from Table 2 are in line with the results from Table 1 which already

[1] dev09 denotes in fact the dev09sub test set defined within GALE.

| System | $\alpha$ | dev08 |
|---|---|---|
| **MPE** | - | 21.6 |
| **MPE⊕CPE** | 0.1 | 21.5 |
| **MPE⊕CPE** | 0.2 | 21.4 |
| **MPE⊕CPE** | 0.3 | **21.3** |
| **MPE⊕CPE** | 0.4 | 21.6 |
| **MPE⊕CPE** | 0.5 | 22.9 |

Table 2: ROVER results combining the first iteration CPE systems with the MPE system, WER in %.

indicated that CPE training can produce complementary systems to an existing base system. The small combination gains observed in Table 2 compared with the large differences in the cross-scoring results of Table 1 seems to indicate that either the introduced complementarity mostly affects the interchange of a wrong hypothesis by a different but also incorrect hypothesis, or that with just two systems the confidence scores used are unable to select the lowest error rate output.

Although CPE training is able to produce a complementary model $\mathcal{M}_{\mathcal{C}}$ with respect to the reference system $\mathcal{M}$, the reductions in WER after ROVER are small. However, the CPE framework can be applied in an iterative manner which offers a way to increase the difference between the target system $\mathcal{M}_{\mathcal{C}}$ and the base system $\mathcal{M}$ further. Thus, a second round of CPE training was performed with the aim to push the target system further away from the MPE base system. The cross-system to generate the cross-references $\boldsymbol{w}_{\text{xref}}^{(r)}$ used the CPE model with $\alpha = 0.2$ from the first CPE training round denoted $\text{CPE}_{0.2}$. The same $\text{CPE}_{0.2}$ model was also used to initialise the CPE training. Thus, this second CPE training round can also be seen as a continuation of the first CPE training round though replacing the cross-references with a version which represents the statistics of the current complementary model $\mathcal{M}_{\mathcal{C}}$.

Table 3 and Table 4 give the corresponding CN decoding results, cross-scoring results and final ROVER WERs for the second CPE training round.

| System | $\alpha$ | CN-WER dev08 | cross-WER dev08 |
|---|---|---|---|
| **MPE** | - | 21.6 | 9.9 |
| **CPE** | 0.1 | 22.0 | 9.9 |
| **CPE** | 0.2 | 21.9 | 9.9 |
| **CPE** | 0.3 | **21.4** | 9.8 |
| **CPE** | 0.4 | 21.5 | 11.3 |
| **CPE** | 0.5 | 22.9 | 14.2 |

Table 3: CN decoding results (CN-WER) and cross-scoring results (cross-WER) for the MPE system and various second iteration CPE systems, WER in %. Cross-scoring scores the CPE systems against the MPE system.

When inspecting Table 3 for small and large $\alpha$, an increase in CN-WER is observed, whereas for $\alpha = 0.3$ and $\alpha = 0.4$ the CN-WER decreases as desired. In contrast to the expected increase in CN-WER for $\alpha = 0.5$, the increase of the CN-WER in case of small $\alpha$ may be explained by the small amount of acoustic training data and, as a consequence, by an over-training of the model. Each CPE training run uses eight iterations of parameter re-estimation. However, as a CPE trained model from the first CPE training round is used to initialise the second CPE training round, in total 16 CPE training iterations are used. For

small $\alpha$ values, when the influence of the cross-accuracy term is still small, this may lead to over-trained models.

As expected, the cross-scoring results shown by Table 3 give increased cross-WERs compared to those from the first CPE training iteration given in Table 1. Next, Table 4 presents ROVER results, combining the baseline MPE system with the CPE systems obtained by iterating the CPE framework twice.

| System | $\alpha$ | dev08 |
|---|---|---|
| **MPE** | - | 21.6 |
| **MPE⊕CPE** | 0.1 | 21.3 |
| **MPE⊕CPE** | 0.2 | 21.3 |
| **MPE⊕CPE** | 0.3 | 21.0 |
| **MPE⊕CPE** | 0.4 | **20.9** |
| **MPE⊕CPE** | 0.5 | 21.4 |

Table 4: ROVER results combining the second iteration CPE systems with the MPE system, WER in %.

Inspecting Table 4 and comparing it to the ROVER results of the first CPE training round in Table 2, additional reductions in WER are observed. The best performing combination system is obtained for $\alpha = 0.4$. When comparing the MPE baseline with the best ROVER result a reduction in absolute WER of 0.7% is obtained. A further CPE training round using the $\text{CPE}_{0.4}$ model from the second CPE training round as the cross-system did not show any further reductions in WER.

### 3.2. The 1831 hour system

In a second set of experiments we investigated if the above results scale to larger systems. Thus, as described in section 3, a 1831 hour training set with a more sophisticated model configuration was used. To address the possible problem of under-training or over-training, for each model the optimal number of training iterations was also investigated. This was typically found to be nine.

For both the dev08 and the dev09 test sets, Table 5 and Table 6 present the MPE, CPE and ROVER results for selected $\alpha$ values of the first and second CPE training round. The $\text{CPE}_{0.2}$ model of the first CPE training iteration was used as cross-system for the second CPE training round.

| System | $\alpha$ | dev08 | dev09 |
|---|---|---|---|
| **MPE** | - | 14.7 | 17.2 |
| **CPE** | 0.2 | 14.8 | 17.1 |
| **CPE** | 0.3 | 15.0 | 17.4 |
| **ROVER: MPE⊕CPE** | 0.2 | 14.6 | 17.1 |
| **ROVER: MPE⊕CPE** | 0.3 | 14.6 | 17.2 |

Table 5: MPE baseline results, first iteration CPE and ROVER results for the 1831 hours setting, WER in %.

Table 5 and Table 6 show similar trends as observed in case of the 172 hour setup. However, the absolute gains in WER are reduced and obtained for smaller values of $\alpha$s. In case of the smaller WER reductions it was found that this goes along with reduced cross-scoring WERs. In particular for the second CPE training round it was found that the cross-scoring WERs were reduced by approximately a factor of two when compared to the 172 hour set-up.

Comparing the final ROVER system from Table 6 with the MPE baseline system, reductions in absolute WER of 0.1–0.2% are found. Although these WER reductions are small, they are in line with results reported in [8]. Within a boosting framework the authors report for the same test set and a similar system training setup, a reduction of up to 0.3% in absolute WER.

| System | $\alpha$ | dev08 | dev09 |
|---|---|---|---|
| **MPE** | - | 14.7 | 17.2 |
| **CPE** | 0.3 | 14.9 | 17.2 |
| **ROVER: MPE⊕CPE** | 0.3 | 14.6 | 17.0 |

Table 6: MPE baseline results, second iteration CPE and ROVER results for the 1831 hours setting, WER in %.

## 4. Conclusions

In this work we have presented a novel criterion for training a complementary acoustic model with respect to one or more pre-existing acoustic models. The CPE criterion is an extension of the MPE criterion and is based on a linear combination of expected phone accuracies taken with respect to the ground truth transcription and the transcriptions obtained from the systems for which a complementary system is desired. The results presented in the experimental section show that the proposed criterion behaves as desired in producing complementary outputs. Furthermore absolute reductions in WER of 0.7% were obtained for a 172 hours training setup and 0.1–0.2% for a more sophisticated training configuration with 1831 hours of training data.

## 5. References


[1] M. Tomalin, M., F. Diehl, M.J.F. Gales, J. Park, & P.C. Woodland "Recent improvements to the Cambridge Arabic Speech-to-Text Systems", Proc. ICASSP, 2010.

[2] O. Siohan, O., B. Ramabhadran, & B. Kingsbury "Constructing Ensembles of ASR systems using Randomized Decision Trees", Proc. ICASSP, 2005.

[3] C. Breslin, & M.J.F. Gales "Building Multiple Complementary Systems using Directed Decision Trees", Proc. Interspeech, 2007.

[4] R. Zhang & A.I. Rudnicky "A frame level boosting training scheme for acoustic modeling", Proc. ICSLP, 2004.

[5] H. Tang, M. Hasegawa-Johnson & T.S. Huang "Towards Robust Learning of Gaussian Mixture State Emission Densities for Hidden Markov Models", Proc. ICASSP, 2010.

[6] J. Du, Y. Hu & H. Jiang "Boosted Mixture Learning of Gaussian Mixture HMMs for Speech Recognition", Proc. Interspeech, 2010.

[7] R. Tachibana, T. Fukuda, U. Chaudhari, B. Ramabhadran & P. Zhan "Frame-level AnyBoost for LVCSR with the MMI Criterion", Proc. ASRU, 2011.

[8] G. Saon & H. Soltau "Boosting systems for large vocabulary continuous speech recognition", Speech Communication, 54:212–218, 2012.

[9] D.Povey & P.C. Woodland "Minimum Phone Error and I-Smoothing for Improved Discriminative Training", Proc. ICASSP, 2002.

[10] F. Diehl, M.J.F. Gales, X. Liu M. Tomalin & P.C.Woodland "Word Boundary Modelling and Full Covariance Gaussians for Arabic Speech-to-Text Systems", Proc. Interspeech, 2011.

[11] J.G. Fiscus "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proc. ASRU, 1997.