

A New Approach to Minimize Utterance Verification Error Rate for a Specific Operating Point

Wing-Hei AU, Man-Hung SIU

Department of Electrical and Electronic Engineering,
Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong
eehei@ust.hk, eemsiu@ust.hk

Abstract

In many telephony applications that use speech recognition, it is important to identify and reject out-of-vocabulary words or utterances without keywords by means of utterance verification (UV). Typically, UV is performed based on the likelihood ratio of the target model versus an alternative model. The “goodness” of the models and the particular criteria used for estimating these models can have significant impact on its performance. Because the UV problem can be considered as a two-class classification problem, minimum classification error (MCE) training is a natural choice. Earlier work has focused on MCE training to reduce total classification errors. In this paper, we extend the MCE approach to minimize the error rates. In particular, we focus on the error rates at certain operating points and show how this can result in a significant EER reduction for phone verification on the TIMIT and a non-native kids corpus. While the particular technique is developed on utterance verification, it can also be generalized for other verification tasks such as speaker verification.

1. Introduction

Automatic speech recognition has been applied in different telephony applications ranging from weather information retrieval [9] to call routing applications [10]. While the interface progressively allows more flexible linguistic structures, it is nevertheless necessary to identify and reject out-of-vocabulary words or utterances without keywords. This technique is known as utterance verification (UV). UV can also be used, for example, to verify the spoken speech against a set of words that the user is expected to utter, such as in reading aid applications [11].

Unlike in speech recognition, the goal of UV is not to transcribe speech but to decide whether the target words are being spoken. Because of this, UV is often treated as a hypothesis testing problem [2, 4, 5]. The null hypothesis states that the target words are spoken. Typically, the likelihood ratio between the null hypothesis and the alternative hypothesis is used as a score measuring the certainty that the null hypothesis is correct. A decision is made by comparing this score to a pre-set threshold.

Under this framework, one key issue in a verification task, including utterance verification and speaker verification, is the estimation of the models that represent the null hypothesis and the alternative hypothesis. Recent research has shown that discriminative training approaches such as MCE training [8] can outperform the traditional Maximum Likelihood (ML) approaches in some speech recognition [8] and verification [5, 6, 7] tasks. One key question to be addressed in all dis-

criminative training approaches is the formulation of the cost function. Ideally, the cost function should reflect the actual performance relevant to the application. An estimation of the verification error has been proposed in [7]. However, for many applications, the performance is often dependent on a fixed false rejection (FR) rate or false acceptance (FA) rate instead of the total verification rate.

In this paper, we propose a discriminative training approach in which the verification is specifically optimized at a particular operating point, such as at the 5% FR or the equal error rate (EER). This algorithm was evaluated in two different corpora under two different verification scenarios. In the TIMIT corpus, the goal was to verify whether the result obtained by the phoneme recognizer is correct. In a non-native kid’s corpus, the objective was to verify the words pronounced by the children. Both tasks perform phoneme verification, which is often used as the building block for word or sentence verification tasks [6].

The rest of this paper is organized as follows. The general framework of verification via likelihood ratio test is discussed in the next section. The general discriminative training framework and our particular approach for operating point specific training is described in Section 3. Experimental results are given in Section 4 and this is followed by the Conclusion in Section 5.

2. Verification

2.1. Verification Based on Likelihood Ratio

A verification problem, one that verifies whether a claim should be accepted, is often treated as a statistical hypothesis testing problem in which the null hypothesis represents that the claim should be accepted. In speech related verifications, such as speaker verification and utterance verification, a system is often presented a speech segment, O_k , that is claimed to come from a certain model set λ_0 . To verify its claim, a model for the alternative hypothesis, λ_a is formed and the log likelihood ratio,

$$\text{LLR}(O_k, \lambda_0, \lambda_a) = \log \left[\frac{p(O_k | \lambda_0)}{p(O_k | \lambda_a)} \right], \quad (1)$$

between the two models is computed. A decision can be made by comparing $\text{LLR}(O_k, \lambda_0, \lambda_a)$ with a threshold τ . That is,

$$\text{LLR}(O_k, \lambda_0, \lambda_a) \quad \begin{cases} > \tau & \text{accepted} \\ \leq \tau & \text{rejected.} \end{cases} \quad (2)$$

The formation of λ_0 and λ_a is one key issue in the verification problem. For example, in word verification [1, 2, 3, 4], λ_0 for a word can be formed using the concatenation of models in its phonemic transcription. The identity of the word can be

obtained via speech recognition, as in [1, 3]. Typically, it is often easier to formulate the null hypothesis because the data can be thought of as being generated from a known model. However, the alternative hypothesis represents all the data not coming from the known model. The possibilities, in fact, can be unbounded. In speech related verifications, it is often modeled using a general model. For example, a universal background model is used in speaker verification for λ_a . Others have used a Gaussian mixture model (GMM). For utterance verification, the alternative model tends to model the general speech and models such as GMM or a parallel loop of phonemes [3] can be used.

2.2. Verification Evaluation

When the log likelihood ratio $LLR(O_k, \lambda_0, \lambda_a)$ is compared to the threshold, a decision is made and two types of error can occur. If the claim is actually false but is accepted, this is called a false acceptance. If the claim is actually true but is rejected, this is called false rejection. The FA and the FR rates can be traded off by setting a different threshold. For example, if the threshold set is too high, most claims are rejected and this results in high FR rate but very low FA rate. This trade-off is often described by an ROC (Receiver Operating Characteristic) curve. Often, the actual FR or FA rate of interest is application dependent. In order to get a uniform comparison, it is common to compare the equal error rate, the point at which FA rate equals FR rate.

3. Training Procedure

3.1. Discriminative Training for Verification

Minimum Classification Error (MCE) training has recently been used in speech recognition and verification tasks [5, 6, 7, 8]. Instead of maximizing the model likelihood, the goal find a set of model parameters that minimizes a cost function which is an estimate of the number of classification errors. This is achieved by formulating a distance d between the true label for the segments and the hypothesized label so that it is positive when there is an error, and negative if there is not. Then, a sigmoid between 0 and 1 is used as an error counting function to count the number of errors. The cost function is the sum of all the errors from all the segments.

Mathematically, denote the true label of the k -th segment, s_k as δ_k where

$$\delta_k = \begin{cases} +1 & \text{if } s_k \text{ is correct} \\ -1 & \text{if } s_k \text{ is incorrect} \end{cases} \quad (3)$$

Combining this with the decision rule in Equation 2, an error counting function $e(O_k, \lambda_0, \lambda_a, \tau)$ is denoted as

$$e(O_k, \lambda_0, \lambda_a, \tau) = \frac{1}{1 + \exp(-\gamma d(O_k, \lambda_0, \lambda_a, \tau))} \quad (4)$$

where

$$d(O_k, \lambda_0, \lambda_a, \tau) = -\delta_k \left(\frac{1}{M_k} LLR(O_k, \lambda_0, \lambda_a) - \tau \right), \quad (5)$$

γ is the slope of the sigmoid function and M_k is the duration of s_k . Notice that a d larger than zero means that there is a verification error. This can either be a false rejection or a false acceptance error, and a $d(O_k, \lambda_0, \lambda_a)$ smaller or equal to zero implies a correct decision. The final cost function for K segments is,

$$C_l(O_1^K, \lambda_0, \lambda_a, \tau) = \sum_{k=1}^K e(O_k, \lambda_0, \lambda_a, \tau). \quad (6)$$

$C_l(O_1^K, \lambda_0, \lambda_a, \tau)$ can be optimized by adjusting the value of the model parameters via the GPD algorithm [8]. This approach was first proposed in [7] with τ set to zero to simplify the formulation.

3.2. Optimizing Performance at a Selected Operating Point

The focus of Equation 4 is to reduce the verification error. However, one is often more concerned about balancing the FA and FR rates. In this case, a more general form of error counting function is needed. Denote $\hat{e}_l(O_k, \lambda_0, \lambda_a, \tau)$ as

$$\hat{e}_l(O_k, \lambda_0, \lambda_a, \tau) = v(k) e_l(O_k, \lambda_0, \lambda_a, \tau), \quad (7)$$

where $v(k)$ can be thought of a general weighting term on the error count from k with $v(k) = 1$ for all k when minimizing total error count. We also added the subscript l to indicate that this is the error at the l -th iteration. The $v(k)$ weighting term can be thought of as the importance of the segment k . For example, if we would like to re-sample the class prior, one can re-weight $v(k)$.

To minimize the error rate instead of the count, each error is normalized to its contribution of the FR and FA rates. Then,

$$v(k) = \begin{cases} \frac{2}{K + \sum_{i=1}^K \delta_i} & \text{if } \delta_k = 1 \\ \frac{2}{K - \sum_{i=1}^K \delta_i} & \text{if } \delta_k = -1. \end{cases} \quad (8)$$

The new cost function, $\hat{C}_l(O_1^K, \lambda_0, \lambda_a, \tau)$ based on $\hat{e}_l(O_k, \lambda_0, \lambda_a, \tau)$ again is the sum of the error counts, expressed as

$$\hat{C}_l(O_1^K, \lambda_0, \lambda_a, \tau) = \sum_{k=1}^K \hat{e}_l(O_k, \lambda_0, \lambda_a, \tau). \quad (9)$$

While the above formulation allows the optimization of the total error rate, it is not clear that a zero threshold used would be at the operating point of interest, say at FR, f . Define the mapping from a threshold to an FR rate g at iteration l as

$$g = ROC_l(\tau). \quad (10)$$

Instead of using $\tau = 0$, an initial threshold, $\tau_0 = ROC_0^{-1}(f)$ that corresponds to an FR of f can be used. However, because the re-estimation of the parameters is done iteratively, this threshold may need to be adjusted so that it is consistently optimizing the point $FR = g$. So,

$$\tau_l = ROC_l^{-1}(f). \quad (11)$$

4. Experiments

Two sets of phoneme verification experiments were performed to evaluate the operating point specific MCE training on two different corpora, TIMIT and a non-native kids corpus with different approaches of obtaining the null hypothesis. In the TIMIT experiments, the null hypothesis consists of a phoneme sequence obtained using a monophone recognizer. In the non-native kids corpus, the null hypothesis is the dictionary pronunciation of the word which the child is reading out. In both sets of experiments, the feature vectors that were used consisted of 39 parameters including 12 MFCC and one energy term with delta and delta-delta. The initial model was trained using the minimum classification error training framework [8] to minimize recognition error.

A sequence of experiments was performed:

1. baseline: using initial model,
2. minimum verification error (MVE): minimizing the error count using Equation 4 and $\tau = 0$,
3. minimum verification rate (MVR): using Equation 9 with $\tau = 0$,
4. minimum verification rate with a fixed non-zero τ (MVRft): using Equation 9 with $\tau = \tau_0$ for EER and FR = 5%,
5. minimum verification rate with a variable τ (MVRvt): using Equation 9 and Equation 11 for EER and FR = 5%.

4.1. Verification of Phoneme Recognition

The corpus TIMIT was used in this set of experiments. A phone set containing 40 monophones, a silence and a short pause was used. The standard training and testing set was used [13]. The maximum number of Gaussian mixtures was set at 16. This model was also used as the baseline model. The phone recognition accuracy of the baseline model is 59.13%. Except for those non-speech events, silences and short pauses, each phoneme in the output hypotheses of the recognizer was verified. The GMM of 32 mixture components was used as the alternative model.

	baseline	MVE	MVR ($\tau = 0$),
EER	39.2%	37.7%	37.0%

Table 1: EER of the baseline model, MVE model and MVR model in TIMIT

In Table 1, minimizing error count, error rate and the baseline are compared. It is clear that minimizing verification error directly is significantly better than the baseline. Furthermore, minimizing the rate is slightly better than minimizing the count. In a further analysis of the training and test statistics we found that minimizing the count is actually biased to reduce false rejections simply because there are more phonemes that are labeled as correctly recognized in the training set. Introducing the normalizing term reduces this biasing effect.

We then compared the performances of setting different thresholds in the optimization, either fixed or adjusted between iterations. The results are summarized in Table 2.

	FA at FR = 5%	EER
1. Baseline	87.2%	39.2%
2. fixed init. $\tau = 0$	85.9%	37.0%
3. fixed init. τ at EER	85.5%	36.3%
4. with re-adjustment τ at EER	85.5%	36.3%
5. fixed init. τ at 5%FR	85.0%	36.1%
6. with re-adjustment τ at 5%FR	84.6%	36.1%

Table 2: Comparison of different model at different operating points in TIMIT

The first two rows are the result of the baseline model and the result of minimizing the error rate at $\tau = 0$. The third row shows that fixing τ at the threshold of the EER at first iteration gave a small gain both for the 5% FR and EER as compared to starting at $\tau = 0$. However, the re-adjustment of the threshold to EER did not result in any improvement, as shown in the fourth row. In the fifth, as in the EER experiment illustrated in row 3, we fixed τ at the 5% FR point in the first iteration. This

gives a better 5% FR result. Re-adjusting the threshold between iterations resulted in a small improvement.

4.2. Verification of Children’s Pronunciation

4.2.1. Data

This corpus contains English speech data of over 15000 utterances from children ranging from grade 3-5. The speakers, who are gender balanced, are all Hong Kong youngsters. Some speakers have a heavy accent and the speech of others is close to that of a native speaker. The data was recorded in a quiet environment. The data includes sentences and words. The sentences were manually annotated according to the pronunciation variations. The words are phonetically transcribed. 8000 utterances, containing both isolated words and sentences, were used to train the initial MCE recognition model. 3341 isolated words were used to evaluate the verification performance.

4.2.2. Experiment

The initial model is a word-position dependent model, meaning that the different sets of phonetic models were trained for word-begin phonemes, mid-word phonemes and word-end phonemes. The maximum number of mixtures was set to 20 and a 3-state straightly left-to-right model was used. The phoneme recognition accuracy on the evaluation set was 56.42%. This is worse than TIMIT.

A phone level verifying task was done to evaluate the performance of the training procedure. Instead of verifying the recognized phoneme sequence, the target phoneme sequence was obtained from the dictionary and compared with the spoken word so as to verify that the speaker pronounced the word correctly. Each phoneme in the given word was verified independently. Instead of using a GMM, as in the verification of phoneme recognition using TIMIT, an unconstrained mid-word phoneme loop was used to represent the alternative hypothesis.

	baseline	MVE	with MVR($\tau = 0$)
EER	31.3%	28.9%	22.7%

Table 3: EER of the baseline model, MVE model and MVR model in the kids’ corpus

In Table 3 the performance of minimizing the number of errors, minimizing error rate as compared to the baseline, is tabulated. Again, as can be observed from Table 1, both discriminative approaches resulted in a significant EER reduction. However, it is interesting to note that the rate minimization in this case gave a much more significant improvement compared to that presented in Table 1.

	FA at FR = 5%	EER
1. Baseline	69.4%	31.3%
2. fixed init $\tau = 0$	61.6%	22.7%
3. fixed init. τ at EER	60.5%	21.7%
4. with re-adjustment τ at EER	60.0%	21.6%
5. fixed init. τ at 5%FR	59.3%	21.4%
6. with re-adjustment τ at 5%FR	59.0%	21.4%

Table 4: Comparison between different threshold setting strategies for pronunciation verification

In Table 4 the results of using different approaches to set thresholds during the MCE model training are tabulated. As

can be observed in the verification of TIMIT phoneme recognition, the result of using a zero threshold is not much worse than that of using a good initial threshold. Irrespective of the operating point of interest, a good starting threshold can significantly improve the performance for both operating points. In fact, it appears that it is not important to have the exact threshold of the operating point so long as it is close to the right operating point. A comparison of the re-adjustment with a good fixed initial operating point, shows that the re-adjustment does give a slight advantage in 5% FR. One issue worth looking at in regard to the operating points of the tests is that the 5% FR threshold in training and the 5% FR threshold in the test are not the same. So, while we tried to optimize the operating point in the training, the test result at the same FR rate may need a different threshold. For the EER, this is even more difficult because the EER point is not defined as a fixed FR or FA rate but as the point at which the two rates converge. Because the EER in training is not the same as that in the test, their corresponding thresholds are also different. Because of this, the result of 5% FR and EER uses thresholds that are close to the optimized one. This may explain why the EER is actually better if the 5% FR optimized threshold is used.

In Figure 1, the DET [12] curves of the different experiments on the kids' corpus are shown. It can be seen that for most of the range, the use of a non-zero threshold and minimizing the error rate gives good improvements across a wide range of operating points. It is also interesting to see that at the extreme low FR, there is no performance gain for the MVR model that was trained with $\tau = 0$ while there is still some improvement for model that was trained for 5% FR. This means that employing an appropriate threshold in training can improve the performance at the extreme operating points.

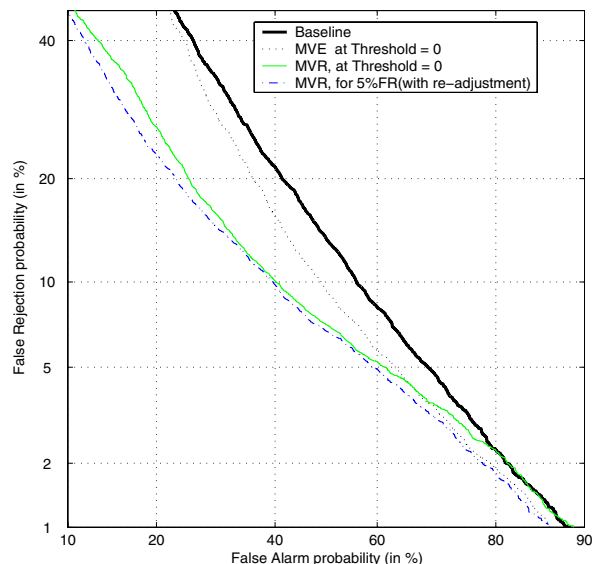


Figure 1: Performance comparison of different strategies in setting threshold τ

5. Conclusion

In this paper, we discussed how we explored the effect of minimizing error rate under an MCE discriminative training framework for verification. We further explored the effect of including the threshold and adjusting the threshold during the MCE

training. We found that minimizing the error rate instead of the count gave significant gains in both the verification of recognized phonemes and for the verification of pronunciation over a wide range of operating points, except in extremely low FR rates. The use of a good initial threshold is also important but re-adjustment of the threshold gave only a marginal improvement. While the approach was developed for phoneme verification, it can, potentially, be applied to other verification tasks such as speaker verification.

6. References

- [1] J. Caminero, C. de la Torre, L. Vilarrubia, C. Martin and L. Hernandez, "On-line garbage modeling with discriminant analysis for utterance verification," *Proc. ICSLP'96*, vol. 4, pp. 2111-2114, Oct. 1996.
- [2] M. G. Rahim, C. H. Lee and B. H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 266-277, May 1997.
- [3] R. C. Rose and E. Lleida, "Speech recognition using automatically derived acoustic baseforms," *Proc. ICASSP'97*, vol. 2, pp. 1271-1274, Apr. 1997.
- [4] A. R. Setlur, R. A. Sukkar and J. Jacob, "Correcting recognition errors via discriminative utterance verification," *Proc. ICSLP'96*.
- [5] R. C. Rose, B. H. Juang, and C. H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," *Proc. ICASSP'95*, vol. 1, pp. 281-284, May 1995.
- [6] R. A. Sukkar, A. R. Setlur, M. G. Rahim and C. H. Lee, "Utterance verification of keyword string using word-based minimum verification error (WB-MVE) training," *Proc. ICASSP'96*, vol. 1, pp. 518-521, May 1996.
- [7] R. A. Sukkar, "Subword-based minimum verification error (SB-MVE) training for task independent utterance verification," *Proc. ICASSP'98*, vol. 1, pp. 229-232, May 1998.
- [8] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proc. IEEE*, vol. 88, pp. 1201-1223, Aug. 2000.
- [9] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen and L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 1, pp. 85-96, Jan. 2000.
- [10] J. Golden, O. Kimball, Man-Hung Siu and H. Gish, "Automatic topic identification for two-level call routing," *Proc. ICASSP'99*, vol. 1, pp. 509-512, Mar. 1999.
- [11] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95-108, 2000.
- [12] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, *Proc. EuroSpeech*, vol. 4, pp. 1895-1898, 1997.
- [13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet and N. L. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, Dec. 1990.