International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Margin Based Learning Framework with Geometric Margin Minimum Classification Error for Robust Speech Recognition

Syed Abbas Ali[a*], Najmi Ghani Haider[b]

[a]Department of Computer & Information Systems, N.E.D University, Karachi-75270, Pakistan

[b]Department of Computer Science & Technology, N.E.D University, Karachi-75270, Pakistan

[a]Email: saaj.scholar@hotmail.com

[b]Email: nghaider@gmail.com

Abstract

Statistical learning theory combines empirical risk and generalization function in single optimized objective function of margin based learning for optimization. Margin concept incorporating in Hidden Markov Model (HMM) for speech recognition, Margin based learning frame work based on minimum classification error (MCE) training criteria show higher capability over any other conventional DT methods in improving classification robustness (generalization capability) of the acoustic model by increasing the functional margin of the acoustic model. This paper introduces Geometric Margin based separation measure in the loss function definition of margin based learning frame work instead of functional margin separation measure to develop a mathematical framework of new optimize objective function of soft margin estimation (SME) for ASR. Derived SME objective function based on Geometric Margin based separation (misclassification) measure would be capable for representing the strength of margin based learning framework in term of classification robustness by minimizing the classification error probability as well as maximizing the geometric margin.

Key Words: Separation measure, minimum classification error, geometric margin, classification robustness, soft margin estimation (SME)

1. Introduction

One of the most successful statistical pattern recognition approaches to model the speech signal as stochastic patterns called Hidden Markov Model (HMM). One of the significant objectives of statistical pattern classifier design is to minimize the classification error probability for all input training samples [1]. A group of discriminative training (DT) criterion of pattern classifiers has been widely studied to reduce the classification error probability [2,3,4,5,6], and discriminative training (DT) methods has become the main research focus in the field of speech recognition. Minimum error classification (MCE) training criterion [2,3,19] among other DT criterion is used as training criteria in speech recognition to show high recognition/classification accuracy. In speech recognition, Minimum classification error (MCE) criterion is based on direct minimization of loss function and total error counts

[a]*Corresponding author:
E-mail address: saaj.scholar@hotmail.com

in input training samples for stochastic patterns of speech signal to increase the classification robustness or generalization capability of the acoustic model. MCE approximates the empirical risk on input training data as differentiable and smoothed objective function and due to lack of direct optimization method, MCE depend on indirect method as control of MCE loss function smoothness [7]. MCE framework make use of three step definition of each input training sample includes; 1) Mapping of input training sample and acoustic model parameter to 0-1 loss function representing classification error 2) Discriminant function and 3) Separation (Misclassification) measure uses to compare the match between the input training samples to correct category with incorrect categories. Separation (Misclassification) measure is one the definition in MCE framework that is of great interest in this paper. Conventional separation (misclassification) measure of MCE frame work is equivalent to functional margin, in which changing in monotone increasing values of separation (misclassification) measure depends upon the adjustment of Λ using MCE training for misclassified and correctly classified data samples, from positive value to negative value and changing in larger absolute values in negative domain respectively[8]. Larger values of separation (misclassification) measure in negative domain (Ŝ in Fig 1) indicate the higher certainty of decision correctness, which reflect the property of separation(misclassification) measures common to functional margin and has been commonly used on pattern recognition/ classification from the earliest research periods for attaining high robustness to unseen data samples [9]. Recent studies [10,11,12] shows inefficiency in separation (misclassification) measure as well as functional margin of MCE training framework due to effect of discriminant function on class boundary. To overcome this problem, Large Geometric Margin Minimum Classification Error (LGM-MCE) [13,14] has been introduced by substituting the functional Margin based separation (misclassification) measure, which represents the geometric distance between class boundary and its closest input training samples as a measure for directly signify the strength of robustness [15]. This paper incorporates separation (misclassification) measure formulated using Large Geometric Margin based MCE (LGM-MCE) training criterion in loss function definition of margin based learning frame work proposed in [17] based on soft margin estimation used in SVM [9,16] to represent the strength of margin based learning framework in term of classification robustness by maximizing the geometric margin [1] as well as minimizing the loss function. Rest of the paper is organized as follows. The subsequent section provides the formulation of conventional Minimum Classification Error (MCE) based separation (misclassification) measures for ASR. This is followed by discussion related to inefficiency of separation (misclassification) measure. Section 3 describe the mathematical formulation of Large geometric margin (LGM-MCE) training using Geometric Margin MCE based separation (misclassification) measure that increases robustness to unseen data sample by maximizing the geometric margin. Soft Margin Estimation (SME) framework for speech recognition is summarized in section 4. In section 5, we present new derived optimized optimize objective function of soft margin estimation (SME) corresponding to Large Geometric Margin MCE (LGM-MCE) training criterion to find strength of robustness (generalization capability) for robust speech recognition. Finally, the conclusion is drawn in section 6.

2. Conventional MCE based Separation Measures (Functional Margin)

Classical Bayesian decision theory [18] is a fundamental approach to handle the pattern recognition problem and qualifies the transaction between decision function based on probabilistic approach and the cost related with this decision function. Automatic speech recognition adopts statistical pattern recognition approach which has its roots in Bayesian decision theory to model the speech signal as stochastic patterns [27]. Consider a speech signal represented as a sequence of an observation vector (input pattern) O= (o_1, o_2, o_3, ……… , o_T) and the pattern recognition task represents patterns with Y classes. One of an unknown pattern is observed from the sequence of an observation vector O and recognized as belonging to one of the Y classes (C_y; y= 1,…….,Y) A speech recognizer with function C maps the observation vector O to a class identity represented by C_k, where K Ɛ I_Y = {K, K = 1, 2, 3, ------Y} called as a decision function $C(O)$.The classifier of the MCE framework adopts decision rule for classification based on linear discriminant functions [2]:

$$C(O) = C_k \;\; if \;\; K = arg \, {}^{max}_{y} g_y(O_t, \Lambda) \tag{1}$$

g$_y(O_t, \Lambda)$ is the linear discriminant function of C$_y$ that specifies the degree to which O belongs to C$_y$, whereas as Λ represents the set of acoustic model parameter for classifier training g$_y(O_t, \Lambda)$(y $=$ 1, ……., Y) is supposed to be differentiable in Λ. In MCE formulation, smooth misclassification measure is use to distinct the competing class from the true class [19] is defined in (2),

$$d(O_t, \Lambda) = -g_t(O_t, \Lambda) + log[\ \frac{1}{N-1} \sum_{y, y\neq t} exp\ g_y\ (O_t, \Lambda)\eta]^{1/\eta} \tag{2}$$

when η approaches infinity, the negative and positive values of (2) represents true and misclassification respectively and $d(O_t, \Lambda)$ becomes

$$d(O_t, \Lambda) = -g_t(O_t, \Lambda) + \underset{y, y\neq t}{max}\ g_y(O_t, \Lambda) \tag{3}$$

In conventional MCE criteria for Automatic Speech Recognition (ASR), the separation (misclassification) measure formulated in [5],

$$d(O_t, \Lambda) = -g_t(O_t, \Lambda) + log[\sum_{\hat{W}_t \neq W_t} P(O_t | \hat{W}_t). P(\hat{W}_t)] \tag{4}$$

where as

$$g_t(O_t, \Lambda) = log\ P(O_t|W_t) = log[P\ (O_t|W_t).\ P(W_t)] \tag{5}$$

By plugging the value of (5) into (4), we can get the equation of MCE based separation (misclassification) measures for ASR,

$$d(O_t, \Lambda) = -\ log[\ P\ (O_t|W_t). P(W_t)] + log[\sum_{\hat{W}_t \neq W_t} P(O_t | \hat{W}_t). P(\hat{W}_t)] \tag{6}$$

where W_t and $\hat{W}_t$ represent the true transcription for all utterances O_t and all possible training data samples in a hypothesis space. Misclassification measure in (6) is a continuous function of acoustic model parameter Λ and tries to emulate decision rule for the observation vectors O, if $d(O_t,\Lambda) \leq 0$ implies correct decision while $d(O_t,\Lambda) > 0$ means wrong decision or misclassification. To obtain the smoothed error count for $O_{t,}$ the misclassification measure is introduced in to sigmoid function as,

$$\ell_t(O, \Lambda) = \ell[d(O_t, \Lambda)] \tag{7}$$

ℓ (.) is a logistic sigmoid function (which is an example of smoothed classification error count loss) can be defined as

$$\ell[d(O_t, \Lambda)] = \frac{1}{1+exp(-\gamma d(O_t, \Lambda))} (\gamma > 0) \tag{8}$$

γ is a positive value number which is belonging to smoothness of loss function. Functional smoothness of MCE training in Eqs. (2) and (8) depend on the adjustment of γ and η and lead one of the standard gradient search procedures which adjust Λ every time, one data sample randomly taken from finite set of training samples. The adjusting mechanism of Λ can be written as

$$\Lambda \leftarrow \Lambda - \mathcal{E}\nabla_{\Lambda}\,\ell(d(O_t,\Lambda))\ (\mathcal{E} > 0) \qquad (9)$$

∇_{Λ} represents the gradient operator with respect to Λ and $\mathcal{E}$ is a learning coefficient of monotonically decreasing function.

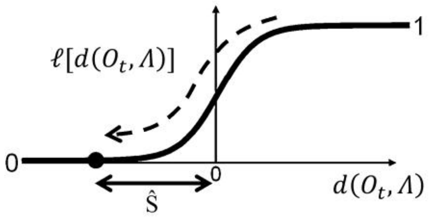

Fig 1 Smooth Classification error count (separation measure)

The adjustment mechanism of Λ using Minimum error classification training changes the values of separation (misclassification) measure for misclassified and correctly classified data samples, from positive value to negative value and changing in absolute large values in negative domain respectively. Larger amount of negative absolute values of separation (misclassification) measure (shown in Fig 1) indicate the decision correctness. This property of separation (misclassification) measures common to functional margin [9] and has been commonly used on pattern recognition/ classification from the earliest research periods for attaining high robustness to unseen data patterns. In negative region of Fig 6.1, the separation (misclassification) measure performs as a margin between incorrect and correct decisions. Most of the recently proposed research approaches based on boosting [20] and large margin [21,22,23,28] primarily exploit this concept of functional margin as shown by horizontal line in the negative direction of Fig 1. From the above analysis, it can be suggested that the MCE training not only focus on the minimization of classification error but also improve the robustness to unseen data sample by increasing the margin. The effect of MCE training has been found inefficient by recent studies [10,11,12] and the main issue of this inefficiency was produced by conventional separation (misclassification) measure, which is equivalent to the functional margin. The reason behind this inefficiency of functional margin as well as separation (misclassification) measure can be understand from the fact that classification boundary does not change by the multiplication of constant positive value common to all discriminant function $\{g_y(O_t,\Lambda)\}_{y=1}^{Y} whereas,$ it does change the absolute negative value of the separation (misclassification) measure. This example clearly evident that by increasing the absolute negative value of separation (misclassification) measure so called functional margin, does not ensure the enhancement of training robustness. This issue leads us to reformulate the separation (misclassification) measure which can directly represent the strength of classification robustness.

3. Geometric Margin Minimum Classification Error Framework

The new version of the minimum error classification (MCE) has been formulated by replacing the conventional separation (misclassification) measure with geometric margin, which represents the geometric distance between class boundary and its closest input training samples and directly reveal the classification robustness [9]. Geometric margin derived in SVM for limited class to two class linear discriminant function whereas, for nonlinearly separable case SVM [9,16] represents the strength of learning framework in term of classification robustness by geometric margin maximization[1] as well as minimizing the hinge loss function (as shown in Fig 2). Incorporated the idea of geometric margin into minimum error classification(MCE) training method, newly formulated geometric margin for general class discriminant function was presented in [13,14].

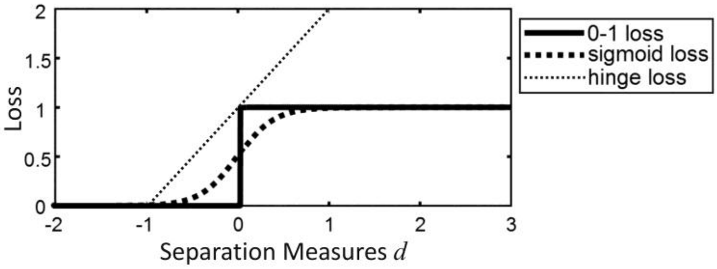

Fig 2 Loss functions

The formulation of geometric margin in Minimum error classification framework for general class of discriminant functions is well established in [13,14,15]. For our discussion, consider a fixed dimensional vector of input data (pattern) sample O_t and for simplicity $\{g_y(O_t, \Lambda)\}_{y=1}^{Y} are$ differentiable in both O_t and Λ. $\mathcal{F}(\Lambda)$ is the set of points at which the separation (misclassification) measure value becomes zero as shown in Fig 3:

$$\mathcal{F}(\Lambda) = \{ O_t | \, d(O_t, \Lambda) = 0\}, \tag{10}$$

define as a boundary which signifies that input data samples are classified as C_y or not. ŕ is define as the Euclidean (geometric) distance between $\mathcal{F}(\Lambda)$ and O_t^ (correctly classified input training sample) belongs to C_y and can be achieved by the solution of constrained minimization problem,

$$\min_{O_t} || \, O_t\bullet \; - \; O_t\hat{}\,|| \qquad \textit{Subject to} \quad d(O_t, \Lambda) = 0 \, , \tag{11}$$

where ||.|| indicates the Euclidean norm and ŕ equals to || O_t•- O_t^|| where O_t•solves the minimization problem in (11). Lagrange multiplier λ is used to solve (11) and define the cost function where O_t•must satisfy the equations:

$$2(O_t\bullet - O_t\hat{}\,)+ \lambda \nabla_{O_t} \; d(O_t\bullet \, , \Lambda) = 0, \tag{12}$$

$$d(O_t\bullet \, , \Lambda) = 0 \tag{13}$$

By considering ($\nabla_{O_t} \; d(O_t\bullet \, , \Lambda) \neq 0$) from (12), ŕ will become

$$ŕ = \; |\lambda|/2 \, \left\| \lambda \, \nabla_{O_t} \; d(O_t\bullet, \Lambda) \right\|, \tag{14}$$

by expanding $d(O_t, \Lambda)$ at point O_t•as follows:

$$d(O_t, \Lambda) = d(O_t\bullet \, , \Lambda) + \nabla_{O_t} \; d(O_t\bullet, \Lambda)^T (O_t - O_t\bullet \,) + O \; (||O_t - \; O_t\bullet \, ||) \tag{15}$$

equating $d(O_t, \Lambda)$ is zero and by approximating O_t=O_t^ in (15), we can get

$$\nabla_{O_t} \; d(O_t\bullet \, , \Lambda)^T (O_t\hat{} - O_t\bullet \,) = d(O_t\hat{}, \Lambda) + o \; (ŕ) \tag{16}$$

from (12), we can get

$$\nabla_{O_t}\ d(O_t,\Lambda)^T(O_t{}^\wedge - O_t\bullet) = {}^{\lambda}/_{2}\, \| \nabla_{O_t}\ d(O_t\bullet,\Lambda)\| \qquad (17)$$

After solving last two equations for λ and substituting results in (14), we obtain equation of Euclidean distance in term of separation (misclassification) measure,

$$\acute{\mathrm{r}} = \frac{|d(O_t{}^\wedge,\Lambda)+o\ (\acute{\mathrm{r}})|}{\|\nabla_{O_t}\ d(O_t\bullet,\Lambda)\|}. \qquad (18)$$

Eq. (18) shows that $|d(O_t{}^\wedge,\Lambda)|$ is equal to functional margin for $O_t{}^\wedge$ and when $O_t{}^\wedge$ is appropriately close to class decision boundary, the functional margin for $O_t{}^\wedge$ is corresponding to geometric margin that is normalized by norm of the gradient of functional margin (or separation measure) at $O_t\bullet$, which is the nearest point to $O_t{}^\wedge$ among all points on class decision boundary.

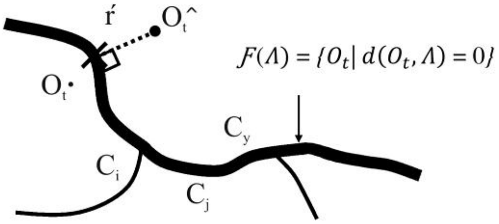

Fig 3 General class of discriminant functions based geometric margin ŕ

As per assumption for general methods of discriminant functions that $O_t{}^\wedge$ is appropriately near to class decision boundary, $o(\acute{\mathrm{r}})$ can be ignore and substitute $O_t\bullet$ by $O_t{}^\wedge$ in (18). Another form of geometric margin can be obtained as follows (even if $O_t{}^\wedge$ is not near to class decision boundary):

$$\acute{\mathrm{r}} = \frac{-\ d(O_t{}^\wedge,\Lambda)}{\|\nabla_{O_t}\ d(O_t{}^\wedge,\Lambda)\|}\ . \qquad (19)$$

Eq. (19) concludes that geometric margin can be increase, by reducing the norm of gradient of the separation (misclassification) measure and/or increasing the functional margin in the region of class decision boundary. Variation in input data sample reflects the variation in the result of classification decision which is represented by the value of denominator in (19). Variation of the classification decision can be suppressed by reducing the norm of gradient of the separation (misclassification) measure, which may result in high robustness [13]. Considering multiple samples rather than one sample $O_t{}^\wedge$ around the class decision boundary, replace $O_t{}^\wedge$ by O_t (one of the input training data samples close to boundary). The new LGM-MCE method [13,14] based separation (misclassification) measure defined as,

$$\text{Ɗ}(O_t,\Lambda) = \frac{d(O_t,\Lambda)}{\|\nabla_{O_t}\ d(O_t,\Lambda)\|}\ , \qquad (20)$$

Eq. (20) corresponds to sign-reversed geometric margin, in addition with correct and incorrect classification decision based on negative and positive values respectively. Now, this time Fig 1 illustrate the geometric margin based separation (misclassification) measure Ɗ $((O_t,\Lambda)$ instead of functional margin $d(O_t,\Lambda)$ and negative direction increases the geometric margin (Ŝ in Fig 1) with decreasing in classification error counts [20].

4. Soft Margin Estimation (SME) Framework for Speech Recognition

To improve the generalization capability of automatic speech recognition, margin based discriminative training criteria called soft margin estimation (SME) [22,24] was proposed to make direct use of an idea of margin in SVM [29] and based on the concept of statistical learning theory (SLT) [16] which is bounded by summation of two target optimization function: an empirical risk function and generalization function. In this section, focus will be on defining the separation (misclassification) measure and formulation of hinge loss function used in SVM for soft margin estimation. Log likelihood ratio (LLR) [19] is used to define separation (misclassification) measure for Soft margin estimation (SME) and can be represented as

$$d(O_t, \Lambda) = log\left[\frac{P(O_t|W_t)}{P(O_t|\hat{W}_t)}\right], \qquad (21)$$

separation measure based on log likelihood (LLR) in (21) provide the correct classification if $\mathrm{d}(O_t, \Lambda) > 0$, otherwise in correct classification would be acquired by classifier. $P(O_t|\hat{W}_t)$ and $P(O_t|W_t)$ represent the likelihood values for the competing and true transcription respectively. Precise separation model can be obtained for each utterance by selecting the frames with different acoustic model labels in competing and true transcription. The Log likelihood frame average value of separation measure for each utterance and discriminative information can be achieved by the selected frame. The equation of the precise model can be defined as:

$$d(O_t, \Lambda) = \frac{1}{n_t}\sum_j log\left[\frac{P(O_{tj}|W_t)}{P(O_{tj}|\hat{W}_t)}\right] \Gamma(O_{tj} \in F_t)\,. \qquad (22)$$

Separation measure in (22) having frame with different labels in competing transcription represented by $\mathrm{F_t}$ whereas, O_{tj} and n_t are the j^th frame for utterance O_t and the number of frame in F_t respectively. The main objective of incorporating margin concept in Soft margin estimation (SME) is to improve the generalization capability of learning classifier. In margin based learning classifier the correct and incorrect decision depends on the value of the soft marginρ, if the value of soft margin is less than the separation measure $d(O_t, \Lambda)$ a correct true decision can be acquired by classifier whereas, loss will be occur when the soft margin is greater than the separation measures and loss function used in SVM is defined as hinge loss functions:

$$l((O_t\,, \Lambda) = \begin{cases} \rho - d(O_t, \Lambda) if\ \rho > d(O_t, \Lambda) \\ 0, \qquad\qquad otherwise \end{cases}. \qquad (23)$$

A test risk bound $R_{true}(z)$ in (24) comprises of VC dimension ‘ɦ’ (capacity measure of the set of function), empirical risk $R_{emp}(f)$ and m is the number of training sample. Equation (24) shows two optimization functions: empirical risk and generalization function, and having probability “1-*τ*” which is bound as:

$$R_{true}(z) \leq R_{emp}(f) + \frac{\sqrt{\text{ɦ}\,(log\left(\frac{2m}{\text{ɦ}}\right)+1-log\left(\frac{\tau}{4}\right))}}{m}. \qquad (24)$$

One possible effort to minimize the test risk bound is to directly minimize the right hand side of (24), but due to computation difficulty and monotonic increasing function of VC dimension ‘ɦ’, generalization function cannot be minimized directly. Vapnik [16] show that, the property of VC dimension “ɦ” which is bounded by decreasing function of margin and can be reduced by increasing the margin. The test risk bound can be estimated by combining the two optimization function in single optimized object function of soft margin estimation (SME),

$$\Lambda_{SME} = \frac{\lambda}{\rho} + \frac{1}{N}\sum_{t=1}^{N} l(O_t, \Lambda) \tag{25}$$

plugging (23) in (25), optimized objective function of soft margin estimation (SME) can be defines as:

$$\Lambda_{SME} = \frac{\lambda}{\rho} + \frac{1}{N}\sum_{t=1}^{N}\big(\rho - d(O_t, \Lambda)\big)\, \text{f}\big(O_t \epsilon \hat{\text{U}}\big) \tag{26}$$

f in (26) denote indicator function whereas, set of utterances Û represent separation measure $d(O_t, \Lambda)$ less than soft margin ρ.

5. Geometric Margin Separation Measure Incorporating in Soft Margin based MCE Criteria

Margin based learning framework proposed in [17] incorporated the separation (misclassification) measure corresponding to conventional discriminative training criterion such as Minimum Classification error (MCE) [19], Maximum Mutual Information (MMI) Estimation [25] and Minimum Word/Phone Error (MWE/MPE) [4], in the loss function definition of soft margin estimation (SME) [22] used in SVM [9] to enhance the performance of automatic speech recognition. In speech recognition applications, minimum error classification (MCE) among all other Discriminative training (DT) criterion show significant progress to increase the classification robustness or generalization capability of the acoustic model. In MCE formulation, separation (misclassification) measure is defined for each training utterances O_t in (6) and substituting this separation measure in the logistic sigmoid function in (8), we can get the equivalent form of MCE criterion [26] as follows:

$$arg\, max \sum_{t=1}^{T} \frac{P(O_t|W_t)\,.P(W_t)}{\sum_{\hat{W}_t} P(O_t|\hat{W}_t).P(\hat{W}_t)} \;. \tag{27}$$

Traditional separation (misclassification) measure of Minimum error classification framework, which is corresponding to functional margin, $\text{d}(O_t, \Lambda)$ can be obtained from (27) and represented as [17]:

$$\frac{P(O_t|W_t)\,.P(W_t)}{\sum_{\hat{W}_t} P(O_t|\hat{W}_t).P(\hat{W}_t)}\,. \tag{28}$$

The formulation of geometric margin in the Minimum error classification framework has been constructed by introducing the idea of geometric margin into minimum error classification (MCE) training criterion and newly proposed LGM-MCE training criterion [13,14] directly increases the geometric margin. Geometric Margin based separation (misclassification) measure refers as $\mathcal{D}(O_t, \Lambda)$ and define in Eq. (20). Based on the above discussion in section 6.2, geometric margin based separation (misclassification) measure $\mathcal{D}(O_t, \Lambda)$ can be replace by functional margin MCE $d(O_t, \Lambda)$. By placing Geometric margin based separation (misclassification) measure instead of functional margin separation (misclassification) measure in the loss function definition of Soft margin (SME) based MCE framework in (26), we can get equation of new optimize objective function of soft margin estimation (SME) corresponding to Large Geometric Margin MCE (LGM-MCE) training criterion as follows:

$$\Lambda_{SME} = \frac{\lambda}{\rho} + \frac{1}{N}\sum_{t=1}^{N}\big(\rho - \mathcal{D}(O_t, \Lambda)\big)\, \text{f}\big(O_t \epsilon \hat{\text{U}}\big)\,. \tag{29}$$

By substituting the MCE framework based traditional separation (misclassification) measure of (28) in (20), we can get equation of Geometric Margin based separation measure in term of Functional Margin MCE for ASR. Eq. (29)

provide mathematical framework based on soft margin estimation (SME) for automatic speech recognition (ASR) with Large Geometric Margin based MCE (LGM-MCE) criterion to represent the strength of margin based learning framework in term of classification robustness by maximizing and minimizing the geometric margin and classification error probability respectively.

6. Conclusion

In this paper, motivated by the Geometric Margin MCE (LGM-MCE) training criterion [13,14] , we revisited margin based learning framework proposed in [17] and derived soft margin based new optimized objective function for ASR by substituting the Functional Margin MCE with Geometric Margin based (LGM-MCE) separation (misclassification) measure to signify the strength of classification robustness through increasing the geometric margin. The objective of introducing Geometric Margin (LGM-MCE) training concept in soft margin based MCE framework used in SVM is to minimize the classification error probability and maximize the classification robustness (generalization capability) to unseen data samples by directly increasing the geometric margin of the acoustic model.

References

[1] C.M.Bishop. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[2] B. -H. Juang, and S. Katagiri. “Discriminative learning for Minimum Error Classification.” IEEE Trans. Signal Processing, vol.40, pp.3043-3054, 1992.

[3] R. Schlueter, W.Macherey, B.Muller, and H.Ney. “Comparison of discriminative training criteria and optimization methods for speech recognition.” Speech Communication, vol.34, pp.287-310, 2001.

[4] D. Povey and P. Woodland. “Minimum Phone error and I-smoothing for improved discriminative training,” in Proc. ICCASP, vol.1, pp. 105-108, 2002.

[5] J. Hui. “Discriminative training of HMMs for automatic speech recognition: A survey.” Computer speech and language, Elsevier Ltd. 2010.

[6] T. Jebara. *Machine Learning: Discriminative and Generative*. Kluwer, 2004.

[7] E.McDermott and S. Katagiri. “A derivation of minimum classification error from the theoretical classification risk using parzen estimation.” Computer Speech and Language, Vol.18, pp. 107-122, 2004.

[8] H.Watanabe, S. Katagiri, K. Yamada, E. McDermott, A. Nakamura,S. Watanabe, and M. Ohsaki. “Minimum error classification with geometric margin control,” in Proc. IEEE ICASSP, pp. 2170-2173, Mar. 2010.

[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[10] T. He, Y. Hu, and Q. Huo. “An approach to large margin design of prototype-based pattern classifiers,” in Proc. ICASSP, pp.II-625-628, April 2007.

[11] T. He and Q. Huo. “A study of a new misclassification measure for minimum classification error training of prototype-based pattern classifiers,” in Proc. ICPR, Dec. 2008.

[12] Y. Wang and Q. Huo. “Sample-separation-margin based minimum classification error training of pattern classifiers with quadratic discriminant functions,” in Proc. ICASSP, pp.1866-1869, Mar. 2010.

[13] H. Watanabe, S. Katagiri, K. Yamada, E. McDermott, A. Nakamura, S. Watanabe, M. Ohsaki. “ Minimum error classification with geometric margin control, “ in Proc. IEEE, pp. 2170–2173. 2010.

[14] H. Watanabe and S. Katagiri. “Minimum classification error training with geometric margin enhancement for robust pattern recognition,” in Proc. IEEE MLSP (CD version), 1–6. 2011.

[15] H. Watanabe, S. Katagiri, S. Matsuda, H. Kashioka, M. Ohsaki, T. Ohashi. “ Robust and Efficient Pattern Classification Using Large Geometric Margin Minimum Classification Error Training.” Journal of signal Processing System. Springer Science NY, June 2013.

[16] V. Vapnik. *The nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[17] S.A. Ali, N.G.Haider and M.K.Pathan. “Margin Based Learning: A Framework for Acoustic Model Parameter Estimation.” I.J. Intelligent Systems and Applications, Vol. 2, No.12, pp. 26-31, November 2012.

[18] R. Duda, P. Hart and D. Stork. *Pattern Classification*, 2nd edition, John Wiley, 2001.

[19] B. -H. Juang, W. Chou, and C.-H. Lee. “Minimum Classification Error rate methods for speech recognition.” IEEE Trans. on Speech and Audio Proc., vol.5, no.3, pp.257-265, 1997.

[20] H. Tanaka, H. Watanabe, S. Katagiri, M. Ohsaki. “Experimental evaluation of kernel minimum classification error training,” in Proc. IEEE TENCON, 1–6. 2012.

[21] H. Jiang, X. Li and C. Liu. “Large Margin Hidden Markov models for speech recognition.” IEEE Trans. Audio, Speech, and Language Processing, vol.14, no.5, pp.1584-1595, 2006.

[22] J. Li, M.Yuan and C.-H.Lee. “Approximate test risk bound minimization through soft margin estimation.” IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.8, pp.2393-2404, 2007.

[23] C. Liu, H.Jiang and L.Rigazio. "Recent improvement of minimum relative margin estimation of HMMs for speech recognition," in Proc. ICASSP, vol.1, pp.269-272, 2006.

[24] J. Li, M.Yuan and C.-H.Lee. "Soft margin estimation of Hidden Markov Model parameters," in Proc. Interspeech, pp.2422-2425, 2006.

[25] Y. Normandin. "Maximum Mutual Information Estimation of Hidden Markov Models," In Automatic Speech and Speaker Recognition. Kluwer Academics Publishers, 1996.

[26] X.He, L.Deng, and W.Chou. "Discriminative Learning in sequential pattern recognition: A unified view for optimization-based speech recognition." IEEE Signal Processing Magazine, pp. 14-36, 2008.

[27] J.F.Gales, S. Young. "The Application of Hidden Markov Models in speech Recognition." Foundation and trends in Signal Processing, Vol.1, No.3 (2007) 195-304, 2008.

[28] D.Yu, L.Deng, X.He and A.Acero. "Large Margin minimum classification error training: A theoretical risk minimization perspective." Computer. Speech Language, vol.22, pp.415-429, 2008.

[29] C. Burges. "A tutorial on support Vector machine for pattern recognition." Data Mining and Knowledge Discovery, vol.2, no. 2, pp.121-167, 1998.