International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

DISCRIMINATIVE TRANSFORM FOR CONFIDENCE ESTIMATION IN MANDARIN SPEECH RECOGNITION

Gang Guo, Ren-Hua Wang

iFly Speech Lab
Department of Electronic Engineering
University of Science and Technology of China
paulguo@ustc.edu rhw@ustc.edu.cn

ABSTRACT

In automatic speech recognition (ASR) application, log likelihood ratio testing (LRT) is one of the most popular techniques to obtain confidence measure (CM). Unlike traditional (log likelihood ratio) LLR related method, we apply non-linear transformations towards LLRs before computing string-level CMs. Different phonemes may have different transformation functions. Through suitable LLR transformations, the verification performances of those string-level CMs may increase. Transformation functions are implemented by Multi Layer Perceptron (MLP). Two algorithms are used to optimize the parameters of MLPs: One is the Minimum Verification Error (MVE) algorithm [2]; another is the Figure-of-Merit (FOM) training algorithm [3]. In our mandarin command recognition system, the two methods remarkably improve the performances of confidence measures for out-of-vocabulary words rejection compared with the performances of standard LRT related CMs, and we obtain a best 45.5% relative reduction in equal error rate (EER). In addition, in our mandarin command recognition experiments, the FOM training algorithm outperforms the MVE algorithm even they share an approximately same best performance, while due to limited experimental setups in our experiments, which algorithm is the better still needs to be explored.

1. INTRODUCTION

In an ASR application system, confidence measure (CM) is critical to reject out-of-vocabulary (OOV) words, and to assist dialogue management, etc. There have lots of CMs proposed before, such as posterior probability, LLR releated CM, etc. LLR releated CMs are calculated based on LRT (log likelihood ratio testing). The calculating of LLR depends on the likelihood derived from the acoustic model and the likelihood derived from the alternative model. Filler model [1] is widely used as alternative model.

In command recognition application, after the string level CM obtained, we can compare it with a threshold to judge the hypothesized phrase/word is out-of- vocabulary or not. The judging is based on hypothesis testing, therefore two types of error may occur: FR (false rejection) and FA (false alarm). We call the detection rate against FA curve as ROC (Receiver Operating Characteristic) curve. There are many ways to evaluate the performance of CM, such as FOM (Figure of merit) and EER (equal error rate). FOM [4] [8] is defined as the average detection rate over a FA range, namely the average area below the ROC curve over a FA range. EER is related to the value when FA equals to FR. Generally, the larger the FOM or the smaller the EER is, the better the performance of the CM is.

The Figure-of-merit (FOM) training algorithm is proposed in [3] to adjust the model parameters in keyword spotting application. FOM training algorithm is a back propagation algorithm aiming at maximizing FOM.

Minimum Verification Error (MVE) algorithm [2][6] is a discriminative algorithm for utterance verification. In [2], the author applied the algorithm to adjust parameters in non-keyword spotting application. The aim of the algorithm is to minimize the verification error, and parameters are trained by a generalized probabilistic descent (GPD) method.

It is clear that different phonemes have different H0 (null hypothesis) probabilistic density distributions and H1 (alternative hypothesis) probabilistic density distributions, but the judging threshold is fixed for all the CMs, therefore the CM computing formula via LLRs can not ensure the maximizing of classification performance of H0 utterance and H1 utterance. For string-level CMs application, we can assign each phoneme a transformation function (functions can also be shared by phonemes belong to a same class), based on some discriminative algorithms, such as FOM training algorithm or MVE training algorithm, etc, we can adjust transformation functions to adjust H0 probabilistic density distributions and H1 probabilistic density distributions of different phonemes, and then improve the performances of CMs.

The organization of this paper is as follows: In section 2, we present the baseline system and the idea of LLR transformation. In Section3, the FOM training algorithm

0-7803-8678-7/04/$20.00 ©2004 IEEE

and the MVE algorithm for MLP parameters training are proposed. Experimental setup and corresponding experimental results are proposed in section 4. Conclusion will be drawn in section 5.

2. FRAMEWORK

2.1. The Baseline System

Phone-level log likelihood ratio (LLR) is defined as:

$$LLR_u = \ln P(o_{ts}^{te} | \lambda_u) - \ln P(o_{ts}^{te} | \lambda_{alter\_u}) \quad (1)$$

where o_{ts}^{te} is a observation which starts at frame *ts* and ends at frame *te*, and λ_u and λ_{alter_u} are parameters of the target unit model and the alternative model respectively.

The calculating of CM is a two-pass recognition process. During the first recognition process, speech boundaries are obtained, and then we can obtain the LLRs. After LLRs obtained, the CM can be calculated according to the CM definition.

Given the phone-level LLR, We can obtain the string-level CM via the following formula

$$CM1 = \frac{1}{T} \sum_{i=1}^{N} PLLR_i \quad (2)$$

where $PLLR_i$ is the LLR of the i-th phoneme, and *T* is the number of frames, and *N* is the number of phonemes in the string. For OOV rejection application, once CM1 obtained, CM1 can be compared with a threshold to judge the phrase or the word be out-of-vocabulary or not?

Filler model is a universal model trained by all the data, which is used as the alternative model. In our paper.

2.2. CM After Transformation of LLR

Given phone-level LLRs and transformation functions, a new CM can be computed as follows:

$$CM2 = \frac{1}{T} \sum_{i=1}^{N} f_{class(Ph_i)}(PLLR_i) \quad (3)$$

where phonemes belong to the same class share a transformation function.

In our command recognition application, we find that if all the transformation functions are defined as:

$$f_{class(Ph_i)}(PLLR_i) = \min(PLLR_i, 0) \quad (4)$$

the performance of CM2 is obviously better than that of CM1, therefore we define CM3 as

$$CM3 = \frac{1}{T} \sum_{i=1}^{N} \min(PLLR_i, 0) \quad (5)$$

and in section 4 we will compare it with CMs via LLR transformation functions trained using the two training algorithms.

Since we do not know the form of transformation functions, we apply Multi Layer Perceptron (MLP) for transformation, and all MLPs include an input unit, multiple hidden units and an output unit.

3. TRAINING OF TRANSFORMATION FUNCTIONS

3.1. FOM (Figure of Merit) Training

As mentioned before, FOM [8] is the average detection rate over a FA range, and it equals the average area below the ROC curve over a FA range. Generally a large FOM is preferred.

Parameters of MLPs can be trained using a Maximum FOM Training algorithm. The details of FOM training algorithm can be found in [3][7].

Similarly, if defined

$$Ph_i Out = f_{class(Ph_i)}(PLLR_i) \quad (6)$$

$$CM = CM\_F(Ph_i Out) \quad (7)$$

where *CM_F( )* is the CM definition, i.e. formula (2).

Parameters of MLPs can be optimized by the following iterative formula:

$$para(t+1) = para(t) + \eta \frac{\partial FOM}{\partial CM} \frac{\partial CM}{\partial Ph_i Out} \frac{\partial Ph_i Out}{\partial para(t)} \quad (8)$$

where *para(t)* are parameters of MLPs at iteration t.

$\partial CM / \partial Ph_i Out$ can be calculated via the definition of current CM.

Ph_iOut is the output of a MLP, so $\partial Ph_i Out / \partial para(t)$ can be obtained by a Back Propagation (BP) algorithm.

After parameters of MLPs obtained, we can calculate CM2. For future comparing, we define CM2 derived from MLPs which trained using the algorithm as CM4.

3.2. MVE Training

Parameters of MLPs can also be trained using the MVE algorithm proposed in [2].

Given $Ph_i Out$ and *CM* defined in (6) and (7), and define

$$R = sigmoid\ (b * CM) \quad (9)$$

$$b = \begin{cases} -1 & if\ \ H0 \\ 1 & if\ \ H1 \end{cases} \quad (10)$$

the aim of this discriminative training algorithm is to minimize the sum of all R. To fulfill it, parameters of MLPs can be optimized by the following iterative formula:

$$para(t+1) = para(t) + \eta \frac{\partial R}{\partial CM} \frac{\partial CM}{\partial Ph_i Out} \frac{\partial Ph_i Out}{\partial para(t)} \qquad (11)$$

Similarly, after parameters of MLPs obtained, we could obtain CM2, and we define current CM2 as CM5.

4. EXPERIMENTS

4.1. Experimental Setup

Experiments were carried out on a mandarin command recognition system that processes 16 kHz PCM speech data. All Chinese syllables are formed by some meaningful combinations of 21 Initials (on the left of syllables) and 38 Finals (on the right of syllables), and we adopt 93 right-context Initials and 38 Finals as acoustic models. Right-context Initials are 3-state left-to-right HMMs, and Finals are 5-state left-to-right HMMs. All of the 131 models are 4 mixture and without state skipping. The feature of the recognition system consists of 12 MFCCs, C0 and their 1-order and 2-order delta coefficients. The filler model is with 3 states and 32 mixtures.

Since the MLPs are for function approximation, we set the activation function of input layers as sigmoid function, and set the activation function of hidden layers as pure linear function. The number of input layer units and output layer units of each MLP are 1, and the number of hidden layer units of each MLP is 20.

In our experiments, CMs are used for OOV rejection. The evaluation data contains Chinese website name utterances by 14 individuals (7 male and 7 female). Each speaker speaks 193 words. 135 of the 193 words are defined as in-vocabulary, and 58 of the 193 words are considered as OOV words.

The data used for MLPs training contains Chinese name, Tang poem, and stock name utterances by other 14 individuals (7 male and 7 male), which means that the training corpus is not correlated with the evaluation corpus. Each speaker speaks 399 words.

MLPs can be shared by phonemes. Phonemes belong to the same class share a MLP, and we test three kinds of class definition in our experiments:

1 Class: all phonemes share a MLP.

10 Classes: we classify all the 131 phonemes into ten classes by k-mean clustering, and phonemes belong to the same class share a MLP.

113 Classes: each phoneme with enough training data owns a MLP, and phonemes with insufficient training data are merged into corresponding similar phonemes.

We will present both FOM and EER in experimental results. (Here the FOM is over a FA range [0, 1.0].)

4.2. Training strategy

For each training data, we firstly obtain LLRs related to the correct word from the transcription then we can obtain LLRs related to 4 competitive words (namely incorrect word) by a NBest recognition procedure. LLRs from the 4 competitive words are considered as H1, and LLRs from the word are considered as H0, then we train those MLPs according to the training algorithms mentioned in section 3.

4.3. Experimental Results

Performances of the two algorithms is described in the following tables.

CM	FOM	Improvement
CM1	0.9743	0%
CM3	0.9878	52.5%
CM4 (1 Class)	0.9900	61.1%
CM5 (1 Class)	0.9897	59.7%
CM4(10 Classes)	0.9932	73.6%
CM5(10 Classes)	0.9928	72.1%
CM4 (113 Classes)	0.9908	64.2%
CM5 (113 Classes)	0.9889	56.9%

Table 1: Performance (FOM)

Where the “improvement” means relative improvement of (1-FOM) compared with that of the baseline system (CM1). (Because the limit of FOM is 1, to demonstrate the improvement more clearly, we present the relative improvement of 1-FOM other than FOM). 1 Class, 10 Classes and 113 Classes mean class definition of MLPs mentioned in section 4.1.

CM	EER	reduction
CM1	7.6%	0%
CM3	5.7%	24.3%
CM4 (1 Class)	4.9%	35.6%
CM5 (1 Class)	5.5%	27.3%
CM4(10 Classes)	4.2%	44.8%
CM5(10 Classes)	4.1%	45.5%
CM4 (113 Classes)	4.7%	37.2%
CM5 (113 Classes)	5.2%	31.8%

Table 2: Performance (EER)

Where the “reduction” means relative reduction of EER compared with that of the baseline system (CM1).

As described in section 3, CM3 is derived from CM1, For CM3 all phonemes share a fixed transformation function:

$$f(x) = \min(x, 0) \qquad (12)$$

CM4 and CM5 are all derived from CM2, and CM4 is obtained via MLPs trained using the FOM training algorithm while CM5 is related to the MVE algorithm.

From table1 and table2, we can conclude that

1. In our OOV word rejection application, CM3 outperforms CM1 significantly, that is to say, the all-phone-shared fixed transformation function (12) improved the performance of confidence measure remarkably for our OOV word rejection application.
2. LLR transformations which trained using the two algorithms mentioned in Section 3 enhance the performance of CMs compared with CMs via fixed LLR transformation function (12), with whichever kind of MLP class definition.
3. If we compare FOM or EER of the same CM with 1-class MLP, 10-class MLPs and 113-class MLPs, we find that CMs with 10-class MLPs outperform CMs with 1-class MLP and CMs with 113-class MLPs, and on the whole CMs with 113-class MLP outperform CMs with 1-class MLPs slightly. It is reasonable that CMs with 10-class MLPs or 113-class MLPs outperform CMs with 1-class, whereas it is theoretically unreasonable that CMs with 10-class MLPs outperform 113-class MLPs. The cause may be that 113-class MLPs are delicately adapted to enhance the performance of training data, hence undermine the generality of those MLPs.
4. From table 1 and table 2, we may find that in our experiments, it seems that the FOM training algorithm outperforms the MVE algorithm on the whole, but when considering the best verification performance, the performance of FOM training algorithm is very similar to that of MVE algorithm (i.e. CM4 vs CM5, both with 10-class definition). Since we just try three MLP class definitions, more experiments may be required to find which algorithm is the better

5. CONCLUSION

In this paper, we explore a MLP transformation method towards LLRs to boost the performance of CMs. each transformation function is implemented by a MLP, and we test 1 MLP, 10 MLPs, and 113 MLPs for 113 mandarin phonemes. To obtain the parameters of MLPs, we try two MLP training algorithms(FOM Training algorithm and MVE algorithm). To show the effect of our method, we present a basic CM (CM1) and an extended CM with a fixed all-phone shared transformation function (CM3) for comparing. The experimental results show that after LLR transformation (via MLP) the performances of CMs all remarkably improve, compared with CM1 and CM3. Furthermore, when adopt 10-class MLP definition other than 1-class MLP definition and 113-class MLP definition, CMs perform the best, namely it is not a fact that CMs with more MLPs will perform better, maybe because the increasing of MLP numbers undermines the generality of MLPs, Hence an appropriate MLP class definition may need to be explored. Finally, we find that in our mandarin command recognition system the FOM training algorithm shares a nearly same best performance with the MVE training method (with 10-class definition), although with other MLP class definition we tried the FOM training algorithm outperforms the MVE algorithm on the whole. The comparison between the two algorithms needs to be explored additionally.

5. REFERENCE

[1] E. Lleida, J.B. Mariño, J. Salavedra, A. Bonafonte, E. Monte, and A. Martínez. Out-of-vocabulary word modelling and rejection for keyword spotting," In *Proc of EuroSpeech*, pp. 1265–1268, 1993.

[2] R.A. Sukkar and C.H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Proc.*, Vol.4, No.6, pp.420-429, 1996.

[3] E. Chang and R. Lippmann, "Figure of merit training for detection and spotting," in *Proc. Conf. Neural Information Processing Systems*, vol.6, pp.1019-1026, Denver, 1993.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Pryzybocki, "The DET curve in assessment of detection task performance," In *Proc of EuroSpeech*, pp. 1895–1898, 1997.

[5] R.C. Rose, B.H. Juang and C.H. Lee, "A training procedure for verifying string hypothesis in continuous speech recognition." In *proc. ICASSP*, pp.281-284, 1995.

[6] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," in *Proceedings of IEEE*, Vol. 88, No.8, pp.1201-1223, 2000.

[7] X.H. Li, E. Chang and B.Q. Dai, "Improving speaker verification with figure of merit training" In *proc. ICASSP*, pp.693-696, 2002.

[8] David Arthur Gethin Williams, "Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition," PhD Thesis, University of Sheffield, Sheffield, England, 1999.