AN IMPROVED ARCHITECTURE FOR WORD VERIFICATION

JunLan Feng LiMin Du
Institute of Acoustics, Chinese Academy Of Sciences,
17 Zhongguancun Rd, Beijing 100080, China
fengjl@iis.ac.cn dulm@iis.ac.cn

Abstract: Word verification is important for both LVCSR and other domain-limited tasks. The first contribution of this paper is to provide a novel strategy to train a set of anti subword models by combining maximum likelihood estimation (MLE) with minimum verification error (MVE) training. Second, this paper proposes a mechanism by which decision thresholds can be set differently for different words depending on their statistically aggressive capabilities obtained from training data.

1. INTRODUCTION

Over the past years, algorithmic advances in automatic speech recognition have demonstrated that highly accurate recognition can be achieved when the input speech is spoken with a rigid and expected speaking format, and when the testing environment resembles that of the training data used to estimate the recognition models. These two deficiencies impede the deployment of a successful service with speech technologies. This paper will focus on the first issue.

According to current speech recognition technologies, actually we are making efforts to utilize finite model sets to describe infinite speech events. However, in real speech recognition applications, it is inevitable that the speech input includes speech segments, which can not be aligned to any pre-trained model. Therefore, the ability to reject incorrectly decoded vocabulary words and out-of-vocabulary words that may appear in an utterance has been identified as an important component of any ASR based human machine interface. Word verification is such a technique to deal with this problem.

Word verification is important for both LVCSR and other domain-limited tasks. In recent years, verification has been formulated as a problem of testing statistical hypothesis. During verification, a hypothesis test is conducted for the targeted word, which results in a confidence score. The word is rejected if its confidence score lies below a test threshold τ which is predetermined. In this process, constructing reasonable alternative hypothesis models is an important part. Focusing on this point, this paper present a novel strategy to train a set of anti subword models by combining maximum likelihood estimation (MLE) with minimum verification error (MVE) training. Also, this paper proposes a mechanism by which decision thresholds τ can be set differently for different words depending on their statistically aggressive capabilities obtained from training data. These techniques were evaluated using 863 speech corpus successfully by experiments which are designed to verify the recognized results given by a speech recognizer which is based on a loopback network of 408 mandarin atonal syllables.

2. WORD VERIFICATION

Word verification can be applied as a procedure for verifying whether the observation vectors in an utterance associated with individual word hypotheses generated by a speech recognition decoder correspond to the hypothesized word label. For a continuous utterance, maximum likelihood decoding relies on a set of models to produce a sequence of hypothesized word labels and hypothesized word boundaries. In recent years, both word verification and utterance verification have been formulated as a problem of testing statistical hypothesis where the task is to test the null hypothesis (H_0) that a given word exists in a segment of speech against the alternative hypothesis (H_1), which assumes such word does not exist within the speech segment. Under Neyman-Pearson hypothesis framework, when testing for a word hypothesis, k, a segment of speech O is rejected if its likelihood ratio:

H_0: null hypothesis, O was generated by the target model λ^c

H_0: alternative hypothesis, O was generated by the alternative model λ^a

$$LR(O,H_0,H_1)=\frac{p_k(O|H_0)}{p_k(O|H_1)}=\frac{p_k(O|\lambda^c)}{p_k(O|\lambda^a)} \qquad (1)$$

falls below a verification threshold τ. Probability density functions $p_k(O|H_0)$ and $p_k(O|H_1)$ are assumed to be known. However, in real operational systems, neither $p_k(O|H_0)$ nor $p_k(O|H_1)$ are known exactly. During the past few years, discriminative training methods based on minimum classification error training (MCE)[1] and minimum verification error (MVE) [3] have been deployed for constructing the null hypothesis and the alternative hypothesis models. These techniques have shown their successfulness.

3. COMBINING MLE AND MVE

3.1 MVE

MVE is a discriminate training technology. Different from MCE which objective is to minimize the classification error, MVE aims to minimize the verification error rate. According to the equation (1) , MVE is used to adapt the parameters of the correct models λ_k^c and the anti models λ_k^a. For HMM-based word verification, λ_k^c and λ_k^a include state observation probability, the state transition matrix, the initial state probability and mixture weights. Generalize probabilistic descent (GPD) as a popular method for discriminative training, was applied to implement MVE. A set of anti subword models and correct subword models will be estimated with this framework.

A frame based distance defined for each state transition in the state sequence $\{s(t_{ui}),\ldots s(t),\ldots s(t_{uf})\}$

$$d(O_t)=\log(a^c_{s(t-1)s(t)}b^c_{s(t)}(O_t))-\log(a^a_{s(t-1)s(t)}b^a_{s(t)}(O_t)) \qquad (2)$$

The segment-based distance is obtained by averaging the frame-based distances as:

$$D_u(O^u)=\frac{1}{t_{uf}-t_{ui}+1}\sum_{t=t_{ui}}^{t=t_{uf}} d_{s(t-1)s(t)}(O_t) \qquad (3)$$

Where and t_{uf} and t_{ui} are the final and initial frame of the speech segment decoded as unit u over the segment $O^u=\{O_{tui},\ldots.O_t\ldots O_{tuf}\}$.

In GPD algorithm framework, a gradient update is performed on the expected cost function:

$$\lambda_{n+1}=\lambda_n-\varepsilon\nabla E\{F(D_u(O^u))\} \qquad (4)$$

Where λ_n is the nth update of the gradient descent procedure, and ε is the learning rate constant; $F(D_u(O^u))$ is a sigmoid function, which is smooth and differentiable with respect to all model parameters as GDP required. The cost function F for unit u can be defined as:

$$F(D_u(O^u)) = \frac{1}{1 + \exp(\sigma(u)(\alpha D_u(O^u) - \beta))} \tag{5}$$

Where the indicator function $\delta(u)$ is defined as

$$\sigma(u) = \begin{cases} 1 & u \in \text{Correct units} \\ -1 & u \in \text{Imposters} \end{cases}$$

The average cost function is a soft count of the number of Type I and Type II errors assuming that the decision threshold is τ. Imposters with scores greater than τ (type II error) and correct units with scores lower than τ (type I) tend to increase the average cost function. Therefore, if we minimize this function we can reduce the verification error rate.

3.2 Combine MLE and MVE

Unlike MLE, gradient descent needs to go through iterative hill climbing procedure to converge to the local minimum (estimate). Gradient descent usually requires many iterations (could easily be more than 100) to converge. Based on the above equation (4), the learning rate coefficient ε must be small enough in order for gradient descent to converge. However, if ε is too small, convergence is needlessly slow. On the other hand, by this procedure only the local minimum instead of the global minimum can be approximated by iterations, Therefore, a set of proper initial models will be very favorable for reducing the computation cost and finding the optimum model parameters.

In this paper, we construct anti models first by Baum-Welch algorithm, in which only speech segments classified wrongly by viterbi decoding make contribution. Assumed a segment of speech was erroneously recognized as the model sequence "mc1, mc2, mc3 ", the corresponding anti model sequences "ma1, ma2, ma3 " will be used as transcription of this segment speech. In this way, a set of anti subword models can be generated by several iterations. In the second step, MVE training is applied, which directly aim at reducing verification errors. The advantage of this framework over GPD alone is that the MLE training at least provides more reasonable initialized parameters for anti models. So, it can be expected that these anti models result in higher scores for those erroneously decoded segments than for others. By the second step, MVE training continues to expand the distance between the correct models and the anti models. In our experiments, we select several topological structures for anti models by considering the balance between the limit of available training data and models' capability to absorb variants. It will be introduced in section 5.

4. THRESHOLDS SELECTING

The second contribution is about the decision threshold τ. Setting different thresholds for different words is more reasonable than a general threshold. The reason is that for HMM-based speech decoder, we often use identical topology structures among most models, but in fact some models have more acoustical variants than others. Moreover, it is impossible that a training corpus can be assured absolutely balanced acoustically. So, the capability of each model is different. If a speech segment which was transcribed as "A", is recognized as "B", we often describe this phenomenon as, "B aggress

A". In our experiments, it is evident some Mandarin syllables are more aggressive than others. Hence, we tune the syllable decision threshold by adding a small term, which varies in the same direction as the syllable's capability to aggress others and in inverse direction with the syllable's probability to be aggressed. This method can be easily extended to word and utterance verification based on subword models' aggressive ability.

5. EXPERIMENTS

An experimental study was performed to evaluate the effectiveness of both the training procedure combining MLE and MVE described in Section 3 and the method to select variable thresholds described in section 4.

863 Mandarin reading speech corpus was used in this study. The training set consisted of 160X600 utterances; the testing set consisted of 8X600 utterances. For all the experiments, the speech recognizer is configured using 7000 context dependent HMM sub-word models. Each model contains three states with continuous observation densities and 6 Gaussian mixtures per state. All model parameters were trained using the maximum likelihood training procedure. These recognition models correspond to the null hypothesis models. Separate alternate hypothesis HMM models were trained for each sub-word unit. The feature extractor of the recognition/verification system computed 12 mel-ceptral coefficients, along with a normalized energy feature. The combined feature vector was augmented with its first- and second-order time derivatives, the so-called delta mel-cepstrum/delta energy and delta-delta cepstrum/delta-delta energy, resulting in a vector of 39 features per frame. In order to investigate the acoustic property of all models, a loopback network of 408 mandarin atonal syllables, that is often called pinyin loopback, was used for the speech recognizer instead of using statistical language models. In all experiments, each syllable was processed as a word. This is reasonable, because the pronunciation of each Chinese character just includes one of these 408 syllables and there is no evident definition for Chinese words, each Chinese character has its own separate meaning and function in the context of language.

5.1 Comparison of training methods

Word level confidence is formed through combining sub-word confidence scores. It is often the case that local mis-matches between the speech segment and HMM models can have a catastrophic effect on the accumulated score used for making a global decision. To mitigate these effects, we used a geometric means proposed in [2] :

$$LR(w_n) = \exp(\frac{1}{N_n}\sum_{i=1}^{N_n}\log LR(U_{n,i})) \qquad (6)$$

Applying the above measure to compare the verification performance of MVE and MLE-MVE training methods, Figure 1 shows the receiver operating characteristic (ROC) curves which display the false alarm rate vs. the detection rate.

It is shown that the MLE-MVE procedure can consistently out-perform the only MVE procedure by several iterations. The results show a drop in the equal error rate from 30% to 21.5%, and a reduction in the minimum error rate from 61% to 41.5%.

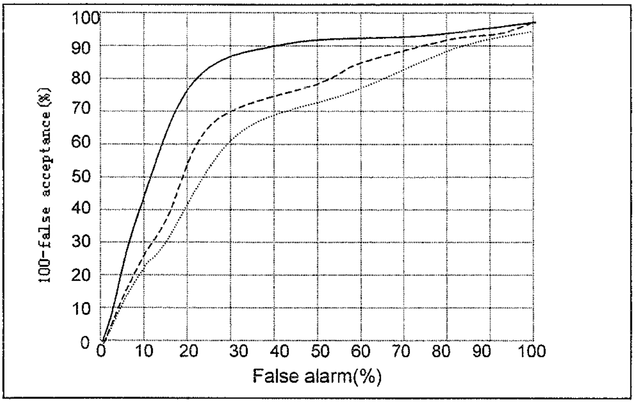

Figure 1. Rejection rate vs. word detection when using MLE, MVE, MLE-MVE to training alternative Models for word verification; ············:MLE; ------ :MVE; ——— :MLE and MVE

5.2 Threshold Choosing:

ROC is the most familiar measure to evaluate a verification procedure. Each operating point on the ROC corresponds to a different trade-off between type I and type II errors. The optimum threshold τ_{min} corresponds to a point on which type I + type II error obtains the minimum value.

In section 4, it has been proposed that HMM models with the same topological structure have different abilities to cover the corresponding acoustic variants. By testing training utterances, each model's ability can be described against its probability to be a aggressive model. In this paper, we investigated syllable-level capability. Assumed a speech segment was decoded as a syllable label: "A", we can roughly estimate its reliability according to pre-obtained this syllable's capability, even without any hypothesis test computing. In our work, we will apply this knowledge source to moderate the threshold τ.

For a given recognized word w, it was determined to be rejected or accepted by a threshold τ_w,

$$\tau_w = \tau_{\min} - a_w \tag{7}$$

where, a_w define the predicted ability of the word w, A measure to define a_w was given in the following:

$$a_w = (\frac{N_{w,w}}{N_w} - A) \times \frac{1}{\beta} \tag{8}$$

where N_w refers to the count of speech segments recognized as "w", $N_{w,w}$ means the count of speech segments correctly recognized as "w", A is the average word accuracy for all training utterances, β is a positive constant which was used to assure τ_w ranges in a small interval $[\tau_{min}-\beta, \tau_{min}+\beta]$..

Table I presents the comparison experiments, it can be found this measure consistently our-performs a general threshold for all words for several different training procedures.

Type I + Type II error(%)	MLE	MVE	MLE-MVE
A general threshold τ_{min}	67%	61%	41.5%
Variable thresholds τ_{w}	62.8%	52%	35%

Table I: Comparing the performance for two strategies to choose thresholds

6. SUMMARY

In this paper, two new ideas were contributed for word verification. First a MLE-MVE training procedure was described, then a new strategy to guide thresholds' choosing was proposed. Also, based on 863 speech corpus, we designed a series of experiments to evaluate these two measures. Their efficiency has been proved.

7. REFERENCE

[1] B.H Juang and S. Katagiri (1992), "Discrimitive learning for minimum error classification", IEEE Trans. On Signal Proc., pp.3043-3054, December 1992

[2] Eduardo Lleda and Richard C. Rose, (1999) "Utterance Verification In Continuous Speech Recognition: Decoding and Training Procedures", IEEE Transactions on Acoustics, Speech and Signal Processing

[3] Mazin G.Rahim and Chin-Hui Lee, (1997) "String-based minimum verification error (SB-MVE) training for speech recognition", Computer Speech and language, 11, pp147-160, 1997