CLASSIFIER DESIGN FOR VERIFICATION OF MULTI-CLASS RECOGNITION DECISION

Tomoko Matsui[1], *Frank K. Soong*[2] *and Biing-Hwang Juang*[2*]

[1]ATR Spoken Language Translation Research Labs, Kyoto, 619-0288 JAPAN
[2]Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

This paper investigates a 2-class classifier approach with the aim of improving the word verification performance. The classifier operates on a discriminant function which is a linear combination of the smoothed likelihood ratios for the N-best candidates and the background (BG) and out-of-vocabulary (OOV) filler models, and is optimized using discriminative training to minimize the classification error. This paper discusses several strategies involving the likelihood ratio based formulation and the use of N-best candidates and the BG and OOV models in the classifier. In word verification experiments using a connected-digit database containing utterances recorded in a moving car with a hands-free microphone, the likelihood ratio based formulation achieved a relative error reduction of 35% in comparison with a likelihood based formulation. In addition, we observed that the use of N-best candidates and the BG and OOV models improved the performance with a relative error reduction of roughly 10%.

1. INTRODUCTION

In human-machine dialogue systems, a high word accuracy achieved by the automatic speech recognition (ASR) system is essential. However, the ASR performance can be seriously degraded, especially in noisy and hands-free environments. To enhance the ASR performance and to design a friendlier voice user interface, a procedure is often integrated into ASR systems to verify or reaffirm the recognized word sequence [1]-[10].

In our previous work [11], we proposed a word verification procedure using a classification approach. The procedure is a post-classification measure, and is used before the recognition result is declared final. Although recognition errors are inevitable, the verification procedure can potentially reduce the negative impact of an incorrect recognition decision or false triggering due to background interference, as found in hands-free environments.

Moreover, the procedure uses a 2-class classifier. Although the traditional framework of hypothesis testing based on a likelihood ratio is often used in word verification, in reality, the legitimacy of the likelihood ratio test may not be upheld due to limited training data and incorrect speech segmentation. Cast in a 2-class classifier scheme, a more sophisticated test function becomes possible and several additional refinements can be made: for example, tests at different levels (e.g., the phone and word levels can be accommodated [5]-[7]); use of different features (e.g., the state duration can be incorporated [5]-[7]); attempted use of non-linear classifiers; and integration of supplementary tests using N-best hypotheses. Therefore, a 2-class classifier formulation can be a good alternative to the traditional framework.

In [11], several classifier design strategies were investigated: a likelihood ratio based formulation, the use of N-best candidate scores, modification of segmentation boundaries, application of BG and OOV filler models, and minimization of verification errors via discriminative training. This paper further validates some of these strategies and searches for a principle of classifier design for multi-class recognition decision.

2. CLASSIFIER DESIGN

Figure 1 shows a block diagram of our classifier designed to verify a word w_u in a recognized word sequence $W = \{w_1, w_2, .., w_U\}$ with the corresponding segmentation $\{X_u\}_{u=1}^{U}$ of the acoustic observation. The

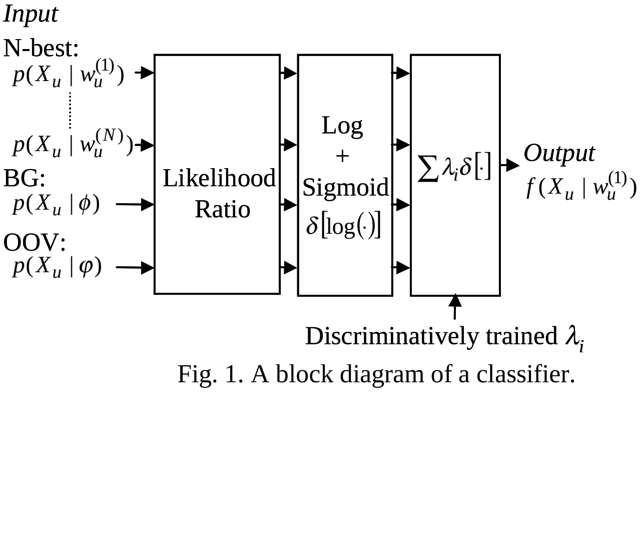

Fig. 1. A block diagram of a classifier.

[*] B.-H. Juang has joined AVAYA Labs Research.

input parameters are the likelihood values of the N-best candidates and the BG and OOV filler models. In this figure, $w_u^{(i)}$ indicates the ith best word candidate, ϕ the BG model, and φ the OOV model. δ is a sigmoid function. To supplement the traditional hypothesis testing and make it robust to noise, the likelihood ratios for the N-best candidates are used. To alleviate insertions, the likelihood ratios for the BG and OOV models are used. To reduce the adverse impact of doubtful likelihood values due to outliers, the likelihood ratios are smoothed, compressed, and regulated by logarithm and sigmoid functions. The classifier then operates on a discriminant function which is a linear combination of the smoothed likelihood ratios and is optimized using discriminative training with the GPD method [12], to minimize the classification error. The discriminant function is defined as follows.

$$\begin{aligned} f(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \delta\left[\log\left(\frac{p(X_u \mid w_u^{(i)})}{p(X_u \mid \overline{w}_u^{(i)})}\right)\right] + \\ \lambda_\phi \delta\left[\log\left(\frac{p(X_u \mid \phi)}{p(X_u \mid \overline{\phi})}\right)\right] + \lambda_\varphi \delta\left[\log\left(\frac{p(X_u \mid \varphi)}{p(X_u \mid \overline{\varphi})}\right)\right] \end{aligned} \quad (1)$$

The likelihood values of anti-models $\overline{w}_u^{(i)}, \overline{\phi}, \text{and}\, \overline{\varphi}$ are approximated by the geometric mean of all of the likelihood values except for the values of $w_u^{(i)}, \phi, \text{and}\, \varphi$, respectively. We do not use any extra models for this approximation. It should be noted that a geometric mean in the likelihood space implicitly suggests an arithmetic mean in the log-likelihood space and has the nature of suppressing statistically doubtful values.

3. VALIDATION OF STRATEGY

This section validates strategies of the likelihood ratio based formulation and the use of N-best candidates, and the BG and OOV models, through word verification experiments.

3.1. Database and System Description

All of the experiments were carried out using the car voice user interface (CARVUI) database, containing utterances recorded in a running car through a 16-channel microphone array located on a sun-visor. Fifty six speakers including some non-native English speakers uttered a number of phonetically-balanced TIMIT sentences, several digit strings with 1 to 7 digits in length, and about 85 short commands for car applications. The data was originally sampled at 24 kHz. In our experiments, hands-free speech data recorded through a channel of the microphone array was used, and all data was down-sampled to 8 kHz after proper low-pass filtering.

For the baseline recognizer, a set of speaker-independent monophone acoustic models were built for 41 phones and three short/long/noisy silences using 3,984 utterances of digit strings and TIMIT sentences uttered by 45 speakers. The total number of mixture components is 2,055 and the averaged number of mixture components per state is 15.8. A feature vector of 39 components, consisting of 12th-order mel-frequency cepstral coefficients plus a normalized log energy term and both of their first and second derivatives, was derived once every 10 ms over a 20 ms Hamming windowed segment. The number of filters used for cepstral computation was 18. Cepstral mean subtraction was applied for each utterance both in training and testing. A finite state grammar with digit strings of an unknown length was used as the language model. The lexicon size was 11 including /0/ to /9/ and /oh/.

The BG filler model was composed of a "silence-loop" model consisting of 3-state long, 1-state short, and 3-state noisy "silence" models in the above speaker-independent monophone models. The OOV filler model was composed of a phone-loop model consisting of 41 phone models in speaker-independent monophone models having lower acoustic resolution, with 483 mixture components in total and 3.9 per state.

In the GPD training, 7,481 correct segments in digit strings uttered by the same 45 speakers as those for the acoustic models were used. The number of digit per speaker was 50.

In the testing, 965 correctly recognized tokens, 125 substitution errors, and 72 insertion segments uttered by seven speakers, who were different from the training speakers, were used. The word-correct rate with the speaker-independent model set was 87.5% on the average. We used the equal error rate (EER) of word verification as the performance measure in our evaluation. The verification threshold was set a posteriori and was digit-dependent. The classifier coefficients were estimated for each digit.

3.2. Likelihood Ratio Based Formulation

Our classifier was formulated based on the likelihood ratio. Here, we confirm the appropriateness of this formulation through comparative experiments with classifiers in (2), (3), and (4).

$$\begin{aligned} f^{2CC(L)}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \delta\left[\log\left(p(X_u \mid w_u^{(i)})\right)\right] + \\ \lambda_\varphi \delta\left[\log\left(p(X_u \mid \varphi)\right)\right] \end{aligned} \quad (2)$$

$$\begin{aligned} f^{2CC(LR)}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \delta\left[\log\left(\frac{p(X_u \mid w_u^{(i)})}{p(X_u \mid \overline{w}_u^{(i)})}\right)\right] + \\ \lambda_\varphi \delta\left[\log\left(\frac{p(X_u \mid \varphi)}{p(X_u \mid \overline{\varphi})}\right)\right] \end{aligned} \quad (3)$$

$$f^{2CC(LR/OOV)}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \delta \left[\log\left(\frac{p(X_u \mid w_u^{(i)})}{p(X_u \mid \varphi)} \right) \right] \tag{4}$$

The input parameters (likelihood values) of the N-best candidates and the OOV model are common for all classifiers. The "2CC(LR)" in (3) is our classifier of the likelihood ratio based form, while "2CC(L)" in (2) is based on the likelihood. The "2CC(LR/OOV)" in (4) is a variation of our classifier and uses the likelihood value of the OOV model as that of an anti-model for each N-best candidate. Table 1 lists the equal error rates of word verification. The likelihood ratio based form in (3) and (4) performs better than the likelihood base form in (2). The relative error reduction from "2CC(L)" to "2CC(LR)" is 35.0%.

If we ignore the sigmoid function, (3) and (4) are transformed as follows.

$$\begin{aligned} & f^{2CC(LR)}(X_u \mid w_u^{(1)}) \\ & \overset{\text{ignore } \delta}{\Rightarrow} \sum_{i=1}^{N} \left[\lambda_i - \frac{1}{N} \left(\sum_{j=1, j \neq i}^{N} \lambda_j + \lambda_\varphi \right) \right] \cdot \log\left(p(X_u \mid w_u^{(i)})\right) \\ & \qquad + \left[\lambda_\varphi - \frac{1}{N} \sum_{j=1}^{N} \lambda_j \right] \cdot \log\left(p(X_u \mid \varphi)\right) \end{aligned} \tag{5}$$

$$\begin{aligned} & f^{2CC(LR/OOV)}(X_u \mid w_u^{(1)}) \\ & \overset{\text{ignore } \delta}{\Rightarrow} \sum_{i=1}^{N} \lambda_i \cdot \log\left(p(X_u \mid w_u^{(i)})\right) - \left(\sum_{j=1}^{N} \lambda_i \right) \cdot \log\left(p(X_u \mid \varphi)\right) \end{aligned} \tag{6}$$

Note that the forms in (2), (5), and (6) are equivalent without the sigmoid function. In order to confirm the influence of the sigmoid function, we conducted verification experiments in which the classifier coefficients λ_i of "2CC(LR)" and "2CC(LR/OOV)" are plugged into "2CC(L)" to initialize λ_i according to (5) and (6). Table 2 lists the equal error rates. In this table, "initial(LR)" and "initial(LR/OOV)" indicate the performance of the plugged-in initial coefficients for "2CC (L)". The "GPD(LR)" and "GPD(LR/OOV)" indicate respectively the performance of applying the GPD training subsequently. Although the GPD training had no significant effect, the plugged-in coefficients for "2CC(L)" performed as well as the original "2CC(LR)" and "2CC(LR/OOV)". Since the sigmoid function preserves the ranking of the candidates when it is applied to the corresponding log-likelihood (ratio) values, this substitution can be considered to work well. This indicates that the sigmoid function itself does not evidently produce the difference in performance. Hence, it follows that there can be many sub-optima for the classifier coefficients.

2CC (L)	2CC (LR)	2CC (LR/OOV)
37.4	24.3	25.9

Table 1. Equal error rates (%) of word verification with different forms.

2CC(L)				
Original	Initial (LR)	GPD (LR)	Initial (LR/OOV)	GPD (LR/OOV)
37.4	25.4	24.6	26.9	26.3

Table 2. Equal error rates (%) of word verification when using coefficients of "2CC(LR)" and "2CC(LR/OOV)" to initialize "2CC(L)".

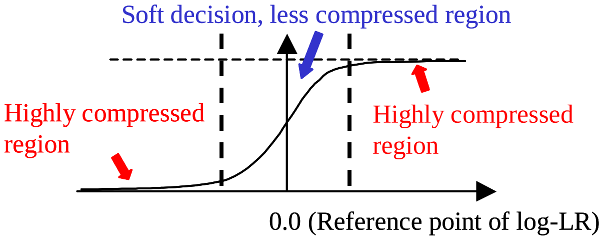
Soft decision, less compressed region
Highly compressed region
Highly compressed region
0.0 (Reference point of log-LR)

Fig. 2. Robust statistical nature of the sigmoid function.

Figure 2 shows a sigmoid function, which substantially compresses values considered outlier or untrustworthy, while keeping the center region somewhat intact. Although the log-likelihood ratio has a meaningful and anchor value of 0.0, the likelihood value does not. Therefore, even when the sigmoid function compresses the range into [0, 1], estimation with the likelihood value can be unstable and prone to undesirable sub-optima. However, the log-likelihood ratio implicitly has an effect of normalizing the log-likelihood value with a reference point of 0.0, and numerically can lead to a better estimate.

When designing a classifier, it is important to formulate the classifier based on a component with a reference point like the likelihood ratio.

3.3. Effects of N-best Scores

We investigated the effect of using N-best likelihood ratios through comparative experiments with 1-, 2- and 11-best scores. Table 3 lists the equal error rates of word verification with and without the BG and OOV models. When these models are not used, use of the 2-best likelihood ratios gave the best performance and the relative error reduction from 1-best to 2-best was 11.4%. When these models were used, in contrast, use of the 11-best likelihood ratios gave the best performance and the relative error reduction from 1-best to 11-best was 10.1%. These results indicate that although the optimal *N* can be

BG and OOV	1-best	2-best	11-best
Without	29.0	25.7	27.2
With	26.8	26.0	24.1

Table 3. Equal error rates (%) of word verification for 1-, 2- and 11-best likelihood ratios with/without the BG and OOV models.

application dependent, utilizing the N-best likelihood ratios is effective.

3.4. Effects of the BG and OOV models

We confirm the effects of the BG and OOV models for insertions. Table 4 lists the equal error rates of word verification using test data excluding/including insertions in our classifier with/without the BG and OOV models. When these models were not used, the error rate for the test data including insertions was relatively higher (4.6%) than that excluding insertions. When these models were used, in contrast, the error rate for the test data including insertions was relatively lower (5.5%) than that excluding insertions. The relative error reduction from the error rate including insertions without the BG and OOV models to that with the models was 11.4%. It can therefore be considered that the incorporation of the BG and OOV models as supplementary knowledge is effective especially when the training data does not have sufficient occurrences of insertions like our training data.

4. CONCLUSIONS AND FUTURE WORK

This paper validates several strategies of a likelihood ratio based formulation and the use of the N-best candidates, and the BG and OOV models, in a 2-class classifier approach to improving the word verification performance. Through connected digit recognition experiments using distant-talking, hands-free CARVUI speech data, we show that it is important to formulate a classifier based on a component with a reference point like the likelihood ratio. The difference in performance between these formulations can be as much as 35% in the relative error rate. Although the optimal *N* can be application dependent, utilizing N-best likelihood ratio is effective and improves the performance by roughly 10%. In addition, the BG and OOV models as supplementary knowledge are effective and lead to improvement in performance by roughly 10% of the relative error rate, especially when the training data does not have sufficient occurrences of insertions.

Insertions	11-best	11-best+BG+OOV
Excluding	26.0	25.5
Including	27.2	24.1

Table 4. Equal error rates (%) of word verification excluding/including insertions with/without the BG and OOV models.

Future work includes investigating the task independency issue and performance of these strategies for large vocabulary recognition. We also plan to optimize *N* in a best "cohort" sense and to further study effects of context on word verification.

REFERENCES

[1] C. V. Neti, S. Roukos and E. Eide, "Word-based confidence measures as a guide for stack search in speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Seattle, Munich, Germany, pp. 883-886,1997.

[2] E. Lleida and R. C. Rose, "Utterance verification in continuous speech: decoding and training procedures," *IEEE Trans. Speech Audio Processing*, Vol. 8, pp. 126-139, 2000.

[3] M. –W. Koo, C. -H. Lee and B. -H. Juang, "A new decoder based on a generalized confidence score," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Seattle, WA, pp. 213-216, 1998.

[4] R. A. Sukkar and C. -H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Processing*, Vol. 4, pp. 420-429, 1996.

[5] R. A. Sukkar, A. R. Setlur, C.-H. Lee and J. Jacob, "Verifying and correcting recognition string hypotheses using discriminative utterance verification," *Speech Commun.*, Vol. 22, pp. 333-342, 1997.

[6] C. Garcia-Mateo, W. Reichl and S. Ortmanns, " On combining confidence measures in HMM-based speech recognizers," in *Workshop Automatic Speech Recognition Understanding (ASRU)*, 1999.

[7] C. Ma, M. A. Randolph and J. Drish, "A support vector machines-based rejection technique for speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Salt Lake City, UT, 2001.

[8] S. O. Kamppari and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Istanbul, Turkey, pp. 1799-1802, 2000.

[9] T. Kawahara, C. -H. Lee and B. -H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. Speech Audio Processing*, Vol. 6, pp. 558-568, 1998.

[10] P. Ramesh, C.-H. Lee and B.-H. Juang, "Context dependent anti subword modeling for utterance verification," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998.

[11] T. Matsui, F. K. Soong and B. –H. Juang, "Verification of multi-class recognition decision using classification approach," *In Proc. ASRU workshop*, 2001.

[12] S. Katagiri, C. -H Lee and B. -H. Juang, "New discriminative training algorithm based on the generalized probabilistic descent method," In *Proc. IEEE workshop, Neural Networks for Signal Processing*, pp. 299-300, 1991.