UTTERANCE VERIFICATION UNDER DISTRIBUTED DETECTION AND FUSION FRAMEWORK

Taeyoon Kim *Hanseok Ko*

Dept. of Electronics and Computer Engineering, Korea University
5Ka-1 Anam-dong, Sungbuk-ku, Seoul, 136-701, KOREA
tykim@ispl.korea.ac.kr hsko@korea.ac.kr

Abstract

In this paper, we consider an application of distributed detection and fusion framework to utterance verification (UV) and confidence measure (CM) objectives. We formulate the UV as a distributed detection and Bayesian fusion problem by combining various individual UV methods. We essentially design an optimal fusion rule that achieves minimum error rate. In the relevant isolated word OOV rejection experiments, the proposed method consistently outperforms over the individual UV methods.

1. Introduction

There are many spoken language applications that require accurate confidence measure (CM) for recognized hypothesis or detecting unreliable and/or undesirable speech input. Utterance verification (UV) is a process of verifying whether a recognized hypothesis is true or false. Thus, it can be considered as a detection problem. Confidence measure is a more general concept which measures quantitatively how well a model match to given observations. In contrast to the utterance verification approach which gives a value of '0' (reject) or '1' (accept), CM gives a continuous value laid between 0 and 1. In the decision making perspective, UV is a hard decision making process and confidence measuring is a soft decision making process. UV and CM are mainly applied to keyword spotting, phrase detection, OOV rejection applications. Also, it can provide a capability of dialog system to manage feedback response according to the reliability of recognized sentences, phrases or words.

The key problem in UV and CM is the judicial selection of an effective method that results in low error rate under acceptable processing speed. Traditionally, there have been several different kinds of approaches in UV and CM evaluation. One of those is the likelihood-based statistical hypothesis testing method that uses normalized likelihood as test statistics. Either anti-model or background filler model are used for alternate hypothesis modeling. The anti-model represents incorrectly decoded hypotheses that are frequently confused with null hypothesis model. Discriminative techniques are widely used for training these models. The background filler model has a role of canceling the source of variability (like channel distortion) on the confidence measure. Thus, modeling of alternate hypothesis should reflect these two conditions. To implement vocabulary independent UV method, sub-word unit CM is used. Sub-word level CM is calculated from the likelihood ratio of correct and alternative hypothesis models of sub-word unit. Then, the word level CM is obtained by weighting and averaging these sub-word level likelihood ratios. While there are many different ways of combining sub-word level confidence scores, most of them fall into one of the following 2 categories: arithmetic or geometric average of sub-word level confidence scores. Compared with the arithmetic average, geometric average tends to give high weight to poorly matched sub-word level units. Thus, it is sensitive to locally mismatched scores. In addition, some CM's use the side information obtained from decoding process, such as the number of active words during the hypothesis decoding, word duration, N-best list, etc…

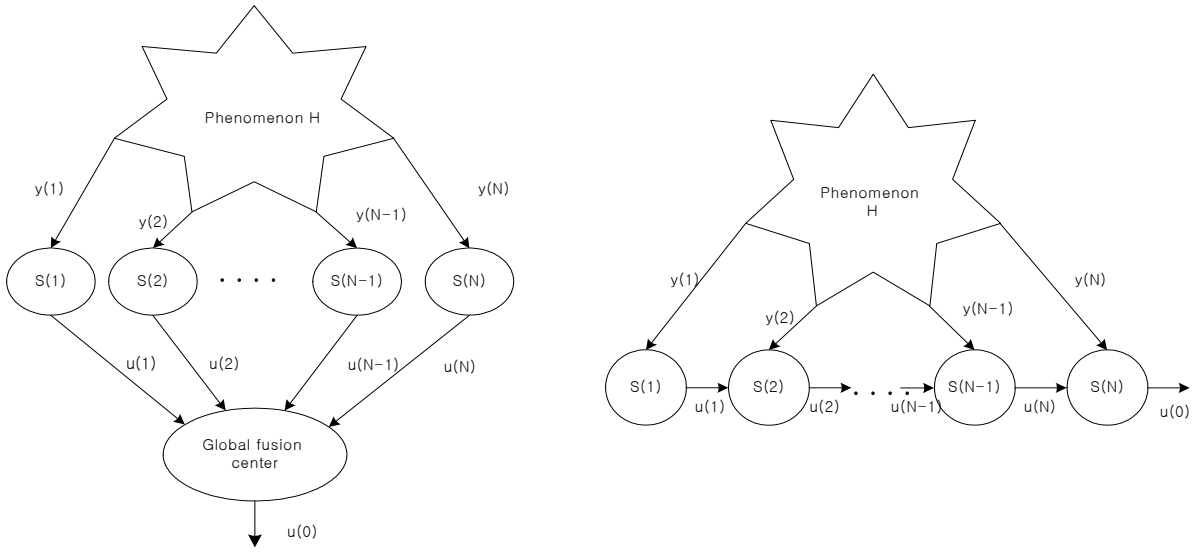
As we examine these scoring techniques, it turns out that many choices are available in CM evaluation and UV. This is because the concept of confidence or reliability is so broad and its general mathematical modeling is almost impossible. To implement this broad concept, combining as much as information sources seems reasonable. Recently, many researchers have proposed techniques of combining various CM and showed that these methods give more reliable CM. These techniques include neural network based CM combining [1], using SVM classifier [2], and LDA analysis [3], etc [4]. However, most of those combining techniques lack the crucial mathematical analysis for conducting the relevant performance evaluations.

In this paper, we address the combining problem under distributed detection and Bayesian fusion framework and present the mathematical analysis of the combining process and its relevant results.

2. Distributed detection and fusion for UV

Distributed sensor problem was originally motivated by their applications in surveillance system. In distributed sensor detection, it is assumed that each local sensor communicates their data to a global sensor that perform optimal detection based on statistical technique. This scheme is referred to a "centralized" processing. In "decentralized" processing, some pre-processing or decision making is carried out in each local sensor and the extracted information or local decision is sent to a global center where fusion of received information and global decision making is made [5]. There are two areas of distributed signal processing, namely distributed detection and distributed estimation. Here we concern about only the distributed detection for application to UV. In a similar manner, CM evaluation problem can be formulated under distributed estimation framework.

Applying to UV problem, aforementioned various CM modeling and evaluation methods can be treated as local sensors which detect or verify acoustic events concerned.



(a) Parallel configuration (b) Serial configuration

Figure 1: Sensor network configurations

Final decision is made in global fusion center which combine the local decisions sent by each local utterance verifier. As shown in Figure 1, sensor network can be organized in a number of ways (parallel, serial, tree,…). In this paper, we only consider parallel configuration for UV where each sensor observes common phenomenon and makes local decision regarding it. From Figure 1.(a), two LRT equations can be written as follows

$$\Lambda(y_i) = \frac{P(y_i \mid H_1)}{P(y_i \mid H_0)} \geq t_i, \tag{1}$$

$$\Lambda(\mathbf{u}) = \frac{P(\mathbf{u} \mid H_1)}{P(\mathbf{u} \mid H_0)} = \prod_{i=1}^{N} \frac{P(u_i \mid H_1)}{P(u_i \mid H_0)} \geq \eta_0 \tag{2}$$

,where y_i is observation of local sensor *i*, t_i is local threshold, u_i is local decision and η_0 is global decision threshold value. Bayesian formulation of the detection problem is possible, in which the objective is to minimize the Bayesian risk [6]. Note that we need to find optimal decision rules for each local sensor as well as a global one. Further more, local decision makings can be coupled each others. In this situation, system-wide optimization is required which leads to complex optimization process. For simplicity, we assume that local decision makings are independent of each other and the local decision rule is already determined. Then we can limit our attention to global fusion rule of local decisions.

Let each u_i, $i = 1,...,N$ has associated probability of miss P_{M_i} and probability of false alarm P_{F_i} defined as,

$$P_{M_i} = P(u_i = 0 \mid H_1), \tag{4}$$

$$P_{F_i} = P(u_i = 1 \mid H_0), \tag{5}$$

As mentioned in the above, these values are assumed to be already determined. Now, the problem can be viewed as a hypothesis testing problem with local decisions as an observation set:

$$\frac{P(u_1, u_2, ...u_N \mid H_1)}{P(u_1, u_2, ...u_N \mid H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} = \eta_0 \tag{5}$$

,where C_{ij} is cost associated with decision making for H_j when H_i is present. Due to the independence of each local decisions, LHS of (5) can be re-written as

$$\begin{aligned} \frac{P(u_1, u_2, ..u_N \mid H_1)}{P(u_1, u_2, ..u_N \mid H_0)} &= \prod_{i=1}^{N} \frac{P(u_i \mid H_1)}{P(u_i \mid H_0)} \\ &= \prod_{s1} \frac{P(u_i = 1 \mid H_1)}{P(u_i = 1 \mid H_0)} \prod_{s0} \frac{P(u_i = 0 \mid H_1)}{P(u_i = 0 \mid H_0)} \\ &= \prod_{s1} \frac{1 - P_{M_i}}{P_{F_i}} \prod_{s0} \frac{1 - P_{F_i}}{P_{M_i}} \end{aligned} \tag{6}$$

Substituting (6) into (5) and taking logarithm,

$$\sum_{i=1} \left[u_i \log \frac{1 - P_{M_i}}{P_{F_i}} + (1 - u_i) \log \frac{P_{M_i}}{1 - P_{F_i}} \right] \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \log \eta_0 \tag{7}$$

This form is a weighted sum of local sensor decisions, and the weight reflects the reliability of each local sensor. To choose the global threshold η_0, we need to know the costs and priori probabilities which are application specific. However, in

UV method	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6
Subword LRT	0.1882	* 0.5392	0.1882	0.1882	0.1882	* 0.5392
Phonetic filler	0.1448	0.1448	* 0.2161	0.1448	0.1448	* 0.2038
5-Best	0.2432	0.2432	0.2432	* 0.3191	0.2432	* 0.3191
256 mix. GMM	0.1862	0.1862	0.1862	0.1862	* 0.2251	* 0.2251
Bayesian fusion	0.1383	0.1377	0.1493	0.1337	0.1333	0.1516

(*): operating point doesn't have minimum total error

Table 1: Total error rate at various operating points.

most UV applications, instead of computing η_0, this is dealt as another control variable to obtain an operating point of appropriate performance.

3. Experiments

3.1. Experimental setup

This is a preliminary experiment to see the effectiveness and limitations of distributed sensor detection and fusion framework for UV problem, especially combining of various UV methods. For this purpose, isolated word OOV rejection task in office environment was designed.

Training data set consists of about 120,000 utterance of 6,000 isolated words recorded in office environment. We used a different speech corpus for testing data set. It consists of about 15,000 utterances of 452 isolated word recorded in silent environment. The microphones used in training data collection and testing data collection are different. Among the 15,000 utterances in test data, only 6000 utterances of 200 words are used as in-vocabulary(IV) data and the utterances of the other 252 words are used as out-of-vocabulary (OOV) data.

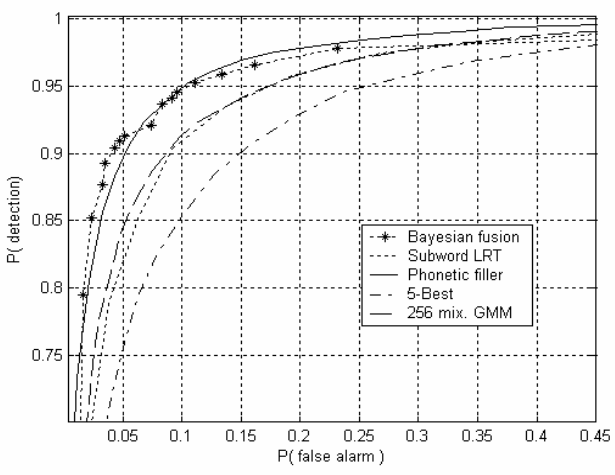


Figure 2: ROC curves of individual UV methods and Bayesian fusion.

3.2. Performance Evaluation

Among the various UV methods, sub-word based LRT, N-best method, phonetic filler method and word level GMM method were tested as a baseline.

In sub-word based LRT, anti-model was constructed by 41 context independent (CI) sub-word HMMs except one HMM of itself. Then the log likelihood of anti-model is computed by arithmetic mean of segmental log likelihood scores of 41 CI sub-word HMMs. Word level CM is obtained by geometric average of sub-word unit CMs. In a similar manner, the phonetic filler was constructed by a network connecting 42 CI sub-word HMMs in parallel. Obtained associated word level log likelihood score was used for normalizing the log likelihood of given hypothesis word. 256 Gaussian mixtures model were used in the GMM method for the word level hypothesis normalization.

To evaluate the performance of UV method for OOV rejection task, hypothesized words are compared against the true one with each hypothesis being classified as correct or incorrect. The confidence scores are then compared with threshold and the hypothesized words are either accepted or rejected. The threshold value can be varied to control the trade-off between false rejection and false alarm. Finally, the detection results obtained from individual methods were fused under Bayesian formulation and evaluated the performance. In the Bayesian fusion, each UV method's operating point and corresponding P_{M_i} and P_{F_i} are selected such that each local UV has minimum total error : $P_{M_i} + P_{F_i}$.

Receiver operating characteristic (ROC) curves are plotted in figure (2). The curve indicates that among the individual UV method, phonetic filler model outperforms all the other methods in every operating point. Compared with fusion method, phonetic filler shows better performance in some operating points. But, we should be more careful in this analysis. Since, in the fusion method, we assumed that each local sensor operates in its minimum total error condition and minimized the total error in a fusion center, we should compare the performance of Bayesian fusion with the individual ones in its minimum total error operating condition. Table 1. shows the total error of the individual methods that operate in conditions including minimum and non-minimum total error points and achievable minimum total errors of the Bayesian fusion. It can be seen that when all the UV methods operate in its minimum total error points (condition 1), Bayesian fusion gives the smallest minimum total error. We can also find out that the Bayesian fusion gives the smallest total error consistently in other 5 conditions. It means that even if the local UV methods do not operate in its minimum total error condition, the Bayesian fusion gives the smallest total error compared with each local UV method.

Among the 6 conditions, in table (1), the smallest minimum total error of Bayesian fusion was obtained not in

condition 1 but in condition 5 where 256 mix. GMM method does not operate in its minimum total error point. In addition, the total error of Bayesian fusion in conditions 2, 4 and 5 is smaller than that in condition 1, where all the UV methods operate in its minimum total error point. This puzzling result suggests that, to obtain the minimum total error, we should optimize the local UV rule as well as the global fusion rule, simultaneously. In this experiment, we assumed that the local decision rules were pre-determined and did not optimize the local UV rules. We think that this assumption leads to those sub-optimal results.

4. Conclusions

We proposed a distributed detection and Bayesian fusion framework for UV, especially judicially combining individual UV methods. This approach enables mathematical analysis of UV fusion and its performance. An assumption that each local UV decision rules are independent and their decision costs are uncoupled leads to simple decision rule, but it gave sub-optimal results. We expect that simultaneous optimization skill like person-by-person optimization (PBPO)[6] could give better results. For future work, we plan to apply a distributed Neyman-Pearson(NP) formulation which enables maximization of probability of detection under false alarm rate constraints. In most speech recognition application, this approach could be more useful than making application operate under minimum total error condition.

5. Acknowledgment

This work was supported by Speech Information Technology Research Center.

6. References

[1] Mitch Weintraub, Francoise Beaufays, et al., "Neural-Network Based Measures of Confidence For Word Recognition," *in proc. ICASSP*, pp. 887-890, 1997.

[2] Rong Zhang, Alexander Rudnicky, "Word Level Confidence Annotation using Combinations of Features," *in proc. EUROSPEECH* , pp. 2105–2108, 2001.

[3] Simo Kamppari, Timothy Hazen, "Word and Phone Level Acoustic Confidence Scoring," *in proc. ICASSP*, pp. 1799-1802, 2000.

[4] Delphine Charlet, Guy Mercer, Denis Jouver, "On Combining Confidence Measures for Improved Rejection of Incorrect Data," *in proc. EUROSPEECH,* pp. 2113-2116, 2001.

[5] Ramanarayanan Viswanathan, Pramod Varshney, "Distributed Detection With Multiple Sensors: Part 1 – Fundamentals," Proceedings of the IEEE, vol. 85, No. 1, pp. 54-63, 1997.

[6] Pramod Varshney, *Distributed Detection and Data Fusion*, Springer-Verlag, New York, 1997.