Word Level Confidence Measures Using *N*-Best Sub-Hypotheses Likelihood Ratio

Beng T. Tan, Yong Gu, and Trevor Thomas

Vocalis Ltd., UK
beng@vocalis.com

Abstract

This paper proposes an efficient confidence measure applied at the word level by combining various likelihood ratio tests. The estimates are derived from the local *N*-best sub-hypotheses. This approach allows the confidence measures to take into account the effect of neighboring words and still provides the estimate localized around the word to be verified. It produces an effective confidence measure that is usable for various tasks. We compared the results with other likelihood ratio based confidence measures including garbage model, *N*-best homogeneity and online garbage models. The proposed method gave more than 30% relative false accept rate reduction over other methods and the rejection performance was less task-dependent.

1. Introduction

It is inevitable that a speech recognition system will make some error. Therefore, it is desirable to improve the system performance through utterance verification (UV). The objectives of UV are to reject out-of-vocabulary (OOV) keyword events, to detect erroneous recognition event, and to determine which part of an input utterance is reliably detected.

It has been shown that powerful UV features can be derived from statistical likelihood ratio test using alternative hypotheses given by either the *N*-best algorithm or alternative hypothesis models. The *N*-best homogeneity (NBH) measure has been applied widely for spontaneous speech recognition. It measures the ratio between the sum of likelihoods for all hypotheses in which a keyword appears and the sum of all likelihoods in the *N*-best list [1][3][4].

Instead of using all *N*-Best hypotheses, we can simply compute the second-best likelihood ratio (SBLR) which has been shown to have better rejection rate than the alternative recognition models for rejecting mis-recognized events [2][5]. This feature has been widely applied at the string level because the alternative hypothesis at the word level is not directly available from the *N*-best list.

Alternatively, we can define a set of models to evaluate the score of alternative hypothesis. A simple solution is to have garbage models (GM), such as free grammar monophone models, running parallel with the recognition task [7]. More specific anti-word models (AWM) have also been used successfully to generate alternative hypothesis score [2]. Another very attractive method is to apply on-line garbage modeling (OGM) [8][9]. Local active states can be considered as a set of single-state garbage models without transition probability or linguistic penalty. The frame-by-frame state probabilities of the entire set of the active HMM states are accumulated in time to generate the online garbage model scores.

In this paper, we are looking for an efficient UV procedure at the lowest cost. Therefore, we choose not to apply the GM and AWM approaches. In addition, we are interested in word confidence measures for grammar network with different degree of constraints. Although the word confidence measures can be computed using the string likelihood as in the NBH measures, additional improvement can be achieved if alternative word likelihood is provided. On the other hand, the likelihood ratio derived from single word segment would inherently assume that the confidence measure of individual word decoded in a continuous utterance is independent of the rest of the utterance.

We relax these restrictions and propose an efficient UV procedure applied at the word level by deriving confidence measure from *N*-best sub-hypotheses localized around the word to be verified. This approach allows the confidence measures to take into account the effect of neighboring words and still provides the estimate localized around the word to be verified. We provide a unified framework that can take advantage of combining the NBH, SBLR and OGM based UV features that have more consistent rejection performance for various tasks.

2. *N*-Best Sub-Hypotheses

The local *N*-best sub-hypotheses are constructed from the *N*-best list. The top hypothesis is used as a reference and dynamic alignment is applied to align the hypothesized word boundaries. The dissimilarity between two word boundaries is the difference in number of frames. An example of optimal alignment is shown in Figure 1. The numbers labeled inside the circles along the horizontal and vertical axes are the word boundaries. The local dissimilarities are labeled along the optimal path.

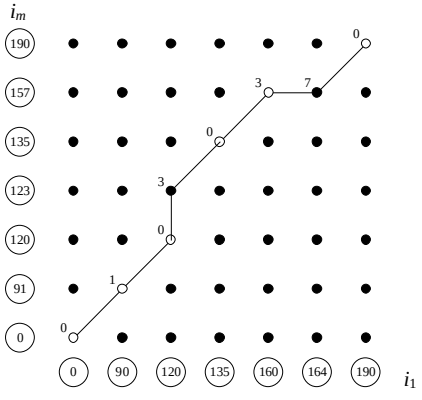

Figure 1 Sub-hypotheses alignment

Suppose we have a node (i_1, i_m) which belongs to the best path through the gird by aligning the top and the m^{th} best hypotheses. The partial minimum accumulated distortion of this node is

$$D(i_1, i_m) = d(i_1, i_m) + \min\{D(i_1 - 1, i_m), D(i_1 - 1, i_m - 1), D(i_1, i_m - 1)\} \quad (1)$$

where D is the partial minimum accumulated distortion and d is the local dissimilarity. We then re-combine the alignments to generate the N-best sub-hypotheses as follows. Each node in the top hypothesis is visited to decide whether it is a valid node for sub-hypothesis boundaries. The local best path must come from the diagonal path and the local dissimilarity is less than or equal to a threshold τ. Those nodes that satisfied these conditions are shown as white circles in Figure 1. These conditions must be meet for all alignments and are re-expressed as follows:

$$\sum_{\substack{m=2 \text{ s.t.} \\ D(i_1,i_m)=D(i_1-1,i_m-1)+d(i_1,i_m) \\ \text{and } d(i_1,i_m)\le\tau}}^{N} 1 = N - 1 \quad (2)$$

If we set the threshold to infinite, we will simply select all aligned boundaries to form local N-best sub-hypotheses. This approach produces most localized sub-hypotheses (MLSH). Figure 2 shows a word graph derived from the N-best results. The number labeled inside each node is the index of frame indicating the boundary of each word segment. The boundaries of the most localized N-best sub-hypotheses are indicated by dashed lines.

We can reduce the overall mismatch in time alignment of sub-hypotheses by selecting only those nodes that have a local dissimilarity less than a threshold value. A strongest condition would be to set the threshold of local dissimilarity to zero i.e. we will keep only those nodes that have zero local distance. The boundaries of the corresponding sub-hypotheses are shown as shaded nodes in Figure 2. Since the overall mismatch in time alignments of sub-hypotheses are zero, this approach produces the most stable sub-hypotheses (MSSH). A threshold between these two extreme cases can be chosen to generate sub-hypotheses of various lengths.

3. Confidence Measures

The proposed UV procedure is shown in Figure 3. The N-best list generated by the recognizer is converted to local N-best sub-hypotheses through dynamic alignment. Each sub-hypothesis may contain more than one word and all words in a single sub-hypothesis may share the same confidence measure. The NBH based UV features using the sub-hypothesis likelihood are converted into two confidence levels – high and low. In this paper, if the NBH based UV feature has a value of one, the first level confidence measure is set to high. This implies that we can not find any alternative sub-hypothesis that does not consist of the word to be verified and the OGM based UV feature would be used to further enhance the confidence measures. When the NBH based UV feature generate a low confidence measure, the UV feature based on SBLR will be computed. These UV features are then converted to posterior probability using a simple parametric solution by assuming the posterior takes the form of a sigmoid [10]. Each of these modules will be described in the following sections.

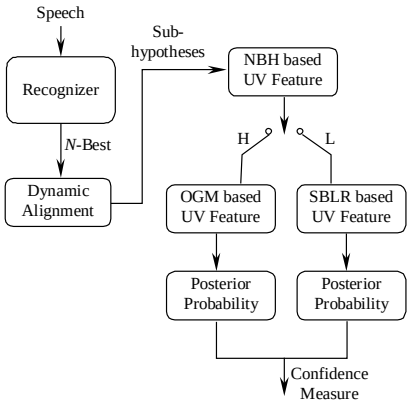

Figure 3 A block diagram of the combined UV system

Let $S_n = w_{1,n}\ w_{2,n}\ \ldots\ w_{M,n}$ be the n^{th} best hypothesis string of length M produced by the recognizer. Let O be the entire observation sequence and $O_{m,n}$ be the observation sequence corresponding to speech segment for word $w_{m,n}$ in S_n as determined by the decoding algorithm. The starting time and ending time of a word w are denoted as $t_{s,w}$ and $t_{e,w}$, respectively. For simplicity, we will represent the duration of a string, S, or word, w, by T_S and Tw, respectively.

3.1. NBH based UV Feature

Since the measurement is based on the occurrence of a keyword in the N-best hypotheses, the grammar must allow the same keyword to appear freely in the N-best hypotheses. Therefore, this feature is particularly useful for recognition task with free grammar using language models and is almost powerless for a simple isolated word recognition task because the grammar network is highly constraint. The NBH

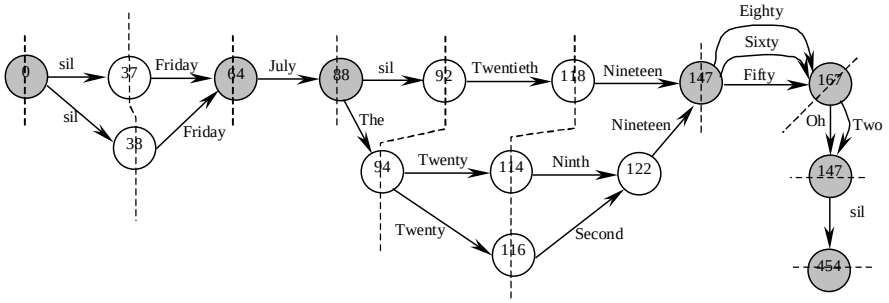

Figure 2 A word graph derived from N-best result

based UV feature of word $w_{m,n}$ is defined by Weintraub [6] as:

$$\chi_{NBH}(m,n) = \frac{\sum_{r:w_{m,n}\in S_r} P(S_r)P(O \mid S_r)}{\sum_{l=1}^{N} P(S_l)P(O \mid S_l)} \quad (3)$$

where $P(O|S_r)$ is the acoustic HMM probability, $P(S_r)$ is the language model probability, and S_r is the hypothesis that contain the word, $w_{m,n}$, to be verified. The typical value of N can go up to 100 for a reliable measurement. Since we are looking for a fast UV procedure, we select N to be around 20, which is much smaller than the typical number used. The NBH is computed directly from the local N-best sub-hypotheses.

The UV procedure proposed here is intended for general tasks including isolated word recognition. Therefore, we avoid using the NBH feature directly. It was observed that for small N, when the NBH measure is very high, the confidence that the hypothesized word is correct is also very high. However, when the measure is less than certain value, it is less reliable and tends to make more error because the statistics collected from small N-best list is less statistically significant. Therefore, the NBH based UV feature is only used to assign a simple high or low confidence measure. For an isolated word recognition task, all the words will have a low NBH measure and thus the algorithm will automatically switch to the SBLR based UV feature.

3.2. OGM based UV Feature

For a hypothesized word with high NBH based UV measures, this feature can be converted directly into a confidence measure. However, the rejection power can be further enhanced with the help of the OGM based UV feature at very low computational cost. It is obtained by accumulating the L-best local scores of active states in the recognition network at a frame-by-frame basis. The online garbage score is combined with the N-best likelihoods through the linear disriminant analysis (LDA) to produce a useful UV feature. The N and L were respectively set to 15 and 20 by Caminero et al. [9] to generate string confidence measures. Since the OGM feature is used only after a high confidence is generated from the NBH based UV features, the information that can be extracted from the N-best sub-hypotheses become less interesting and N is simply set to one. The OGM based UV feature at word level is defined as:

$$\chi_{OGM}(m,n) = \log P(O_{m,n} \mid \lambda_{m,n}) + \sum_{l=1}^{L} a_l \log \varphi_{OGM}(t_{s,w_{m,n}}, t_{e,w_{m,n}}, l) \quad (4)$$

where $\varphi_{OGM}(t_{s,w_{m,n}}, t_{e,w_{m,n}}, l)$ is the l^{th} best local score accumulated from within the word segment.

3.3. SBLR based UV feature

The SBLR based UV features are usually applied at the string level. With the help of N-best sub-hypotheses, this UV feature can be easily used to produce word confidence measures. To verify a word, we must find an alternative sub-hypothesis that do not consist of the word to be verified. Suppose the word to be verified, $w_{m,n}$, is assigned to a local N-best list consists of sub-hypotheses R_1, R_2, … R_N and the word $w_{m,n}$ is found in the sub-hypotheses R_n. The modified SBLR feature is computed by finding the best alternative sub-hypothesis that do not contain the word $w_{m,n}$. The SBLR measure is defined as

$$\chi_{SBLR}(m,n) = \frac{\log(P(O_{R_n} \mid R_n))}{T_{R_n}} - \max_{\substack{i \\ \text{s.t. } w_{m,n}\notin R_i}} \frac{\log(P(O_{R_i} \mid R_i))}{T_{R_i}} \quad (5)$$

4. Experiments

The UV algorithms were evaluated on both isolated word and connected word recognition. Speech data for connected speech recognition include date, number, spell and postcode recognition. The isolated word recognition task was evaluated on a field database that consisted of about 9000 utterances and the vocabulary set consisted of 32 names. The database was selected and divided into training and testing set for UV procedure. There is no out-of-vocabulary utterance.

The acoustic models were trained from a separate database and were modeled by continuous density multiple Gaussian distributions hidden Markov model (HMM). Each model had 3 emitting states with a left-to-right topology and 16 mixture components. The acoustic features used were 12-cepstra and their first and second derivative.

We designed experiments to compare the standard NBH, OGM, GM, and our proposed UV procedure. The recognizer generated about 20 best hypotheses. The standard NBH was computed as in Equation (3) using string score with N set to 20. The standard OGM based UV feature was computed at the word level with L being set to 5. No additional improvement was observed when L is set to a larger number. For GM based UV feature, free grammar monophone models running parallel with the recognition task were used as garbage models. The GM based UV feature was computed by time aligning the word to be verified with the garbage models. We also compared the performance of the most localized sub-hypothesis (MLSH) and the most stable sub-hypothesis (MSSH) by setting the threshold τ to infinitive and zero, respectively.

5. Results and Discussions

Table 1 shows the correlation coefficients with the correct/incorrect tag for various UV features under investigation. The correlation coefficients give us a guideline to choose and compare the useful UV features. If the absolute value of this correlation coefficient is high, the corresponding feature can be regarded as a good UV feature. The proposed UV methods using MLSH and MSSH were significantly better than the rest. The GM based UV feature had the lowest correlation coefficients.

Figure 4 shows the averaged performance of various UV features. The worst performance was given by the GM based UV feature. We expect better performance of GM method for rejecting the out-of-vocabulary events. The best performance was given by using the most stable sub-hypothesis (UV-

MSSH). By rejecting 10% of the correctly recognized words, the UV-MSSH system can reject, on average, 64% of the incorrectly recognized words. The proposed method gave more than 30% relative false acceptance rate reduction over other methods. It is interesting to note that the MSSH, which contained more information from neighboring word, performed better than the MLSH. This suggests that it is useful to encode this information into the UV features.

UV Features	Correlation
OGM	0.3089
NBH	0.2046
GM	0.0515
UV-MLSH	0.4062
UV-MSSH	0.4705

Table 1 Absolute value of correlation coefficients to correct/incorrect tags

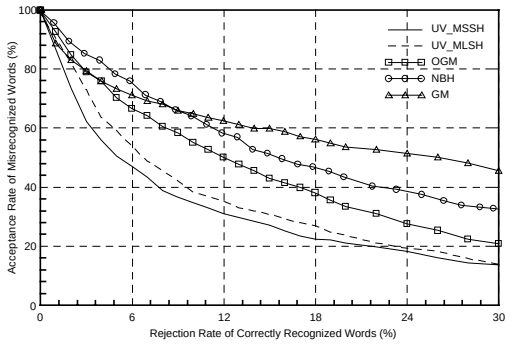

Figure 4 Performances of various UV features

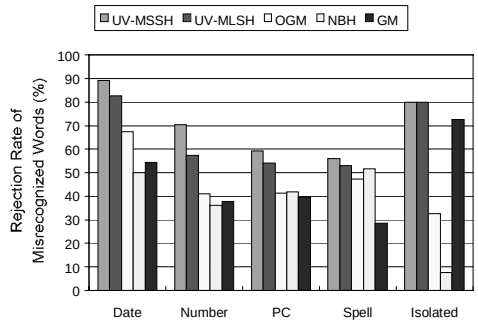

Figure 5 Performances of various UV features at 10% rejection rate of correctly recognized words.

Figure 5 compares the performance of various UV features at 10% rejection rate of correctly recognized words. The most difficult task is the recognition of alphanumeric string, which includes the postcode (PC) and spell recognition. The isolated word and date recognition consists of highly constrained grammar network and has the best rejection performance. Comparing the performance of OGM, NBH and GM based UV features, we found that the OGM is best for date recognition, NBH for spell recognition and GM for isolated word recognition. They have similar performance for number and postcode recognition. The NBH based UV feature is very useful for connected word recognition but is almost powerless for isolated word recognition. The GM based UV feature performs significantly better than OGM and NBH for isolated word recognition task. The UV-MSSH and UV-MLSH out-perform other methods for all tasks and the rejection performances are less task-dependent.

6. Conclusions

This paper proposes an efficient confidence measure applied at word level based on likelihood ratio test. In particular we combine the *N*-best homogeneity, online garbage models and second best likelihood ratio using local *N*-best sub-hypotheses. Various confidence measures were evaluated on isolated and connected word recognition. The proposed method out-performed others by more than 30%.

Local *N*-best sub-hypotheses have been shown a very useful framework for extracting UV feature. In our proposed UV procedure, the most stable sub-hypotheses (MSSH) had a better performance than the most localized sub-hypotheses (MLSH). This suggests that it is important to include neighboring information to decide the correctness of a recognized word. The derivation of *N*-best sub-hypotheses also inherently takes into account the time alignment information in the *N*-best results. Further research is required to explore useful features that can be derived from local sub-hypotheses.

7. References

[1] Cox, S. and Rose, R. C. "Confidence Measures for the Switchboard Database", *Proc. ICASSP, Vol. I*, pp. 511-514, 1996.

[2] Sukkar, R. A., Setlur, A. R., Lee, C. H., & Jacob, J. "Verifying and correcting recognition string hypotheses using discriminative utterance verification", *Speech Communication*, 22:333-342, 1997.

[3] Weintraub, M. "LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting", In *Proc. ICASSP*, vol. 1, pp. 297-300, 1995.

[4] Rueber, B. "Obtaining Confidence Measures from Sentence Probabilities", *Proc. Eurospeech* pp. 739-742, 1997.

[5] Tan, B. T., Gu, Y., and Thomas, T. "Evaluation and Implementation of a Voice-Activated Dialing System with Utterance Verification", *Proc. ICSLP*, Sydney, Australia, 1998.

[6] Wendemuth, A., Rose, G., and Dolfing, J. G. "Advances in Confidence Measures for Large Vocabulary", *Proc. ICASSP*, pp. 705-708, 1999.

[7] Young, S. R., and Ward, W. "Recognition Confidence Measures for Spontaneous Spoken Dialog", *Proc. Eurospeech*, pp. 1177-1179, 1993.

[8] Bourlard, H., D'hoore, B., and Boite, J. M. "Optimizing Recognition and Rejection Performance in Wordspotting Systems", *Proc. ICASSP, Vol. 1*, pp. 373-376, 1994.

[9] Caminero, J., de la Torre, C., Villarrubia, L., Martine, C., and Hernandez, L. "On-Line Garbage Modeling with Discriminant Analysis for Utterance Verification", *Proc. ICSLP, Vol. 4*, pp. 2111-2114, 1996.

[10] Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000.