ERROR TYPE CLASSIFICATION AND WORD ACCURACY ESTIMATION USING ALIGNMENT FEATURES FROM WORD CONFUSION NETWORK

Atsunori Ogawa, Takaaki Hori and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation
{ogawa.atsunori,hori.t,nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

This paper addresses error type classification in continuous speech recognition (CSR). In CSR, errors are classified into three types, namely, the substitution, insertion and deletion errors, by making an alignment between a recognized word sequence and its reference transcription with a dynamic programming (DP) procedure. We propose a method for deriving such alignment features from a word confusion network (WCN) without using the reference transcription. We show experimentally that the WCN-based alignment features steadily improve the performance of error type classification. They also improve the performance of out-of-vocabulary (OOV) word detection, since OOV word utterances are highly correlated with a particular alignment pattern. In addition, we show that the word accuracy can be estimated from the WCN-based alignment features and more accurately from the error type classification result without using the reference transcription.

Index Terms— Speech recognition, error type classification, word accuracy estimation, alignment features, word confusion network

1. INTRODUCTION

Errors are essentially unavoidable, and therefore, based on this basic premise, we have to tackle the problems in speech recognition. Namely, auxiliary processing which takes into account the existence of errors in a recognition result plays a critical role in a practical use of speech recognition. For example, *confidence estimation*, which scores the reliability of a recognition result, is one of such auxiliary processing. And *OOV word detection* is another example, since it means the detection of parts in a recognition result which could never be correctly recognized. Many approaches have been proposed for achieving accurate confidence estimation and OOV word detection, e.g. [1, 2, 3, 4, 5].

Confidence estimation and OOV word detection are useful for the detection of unreliable parts in a recognition result. As an extension of the auxiliary processing that deal with recognition errors, this paper addresses the *classification of error types* in continuous speech recognition (CSR), namely, the *substitution, insertion and deletion*. These three types of errors are counted based on an alignment between a recognized word sequence and its reference transcription with a dynamic programming (DP) procedure. If the types of errors are automatically detected and classified without using the reference transcriptions, it can help the development of practical speech application systems. For example, in spoken document retrieval systems, many insertion errors in spoken document transcriptions degrade the precision of the search performance [6]. If we are to improve the precision in this situation, we can put the documents that have less frequently insertion errors at higher ranks in the retrieved document list. As another example, we can estimate the word accuracy from the classification result of the three types of errors without using the reference transcriptions. This is not made possible only with the erroneous part detection. It should be noted that the error type classification includes confidence estimation, since it also detects correctly recognized words along with the three types of errors. To the best of our knowledge, little effort has been made as regards the error type classification compared with confidence estimation and OOV word detection.

We treat the error type classification as the problem of discriminative classification using many kinds of features, as with the recent trends in confidence estimation and OOV word detection. In particular, we employ the features that are designed to directly improve the performance of error type classification. They are *alignment features* derived from a *word confusion network (WCN)* , by summing the posterior probabilities attached to the words, for every type of error (Section 2). Namely, they are the *substitution, insertion and deletion error probabilities* and are used in a discriminative classifier (Section 3.1). We show experimentally that the WCN-based alignment features steadily improves the performance of error type classification (Section 4.1). They also improve the performance of OOV word detection, since OOV words are highly correlated with a particular alignment pattern (Section 3.2). In addition, we show that the word accuracy can be estimated from the WCN-based alignment features and more accurately from the error type classification result without using the reference transcription (Section 4.2).

2. ALIGNMENT FEATURES FROM WORD CONFUSION NETWORK

A word confusion network (WCN) is a compact representation of multiple recognition results (recognized word sequences). It can be efficiently obtained by converting a recognition lattice with consensus decoding [7]. An example of a WCN is shown at the top of Fig. 1. A recognized word (or a null word) is represented as an arc, and all competing recognized words in a segment are represented as arcs that share the same start and end nodes (e.g. the words w_i^c, w_i^d, w_i^e and null word ε_i are competing in the segment i). Each competing word in a segment has a posterior probability ($p(w_i^c)$, $p(w_i^d)$, $p(w_i^e)$ and $p(\varepsilon_i)$) and these posterior probabilities are summed to one ($\sum_v p(w_i^v) + p(\varepsilon_i) = 1$). The word that has the highest posterior probability in a segment (the best word w_i^c) is selected as a word in the 1-best recognition result. And this posterior probability ($p(w_i^c)$) can be used as a confidence measure [4, 6] that scores the reliability of the best word, i.e. the *Correct probability* of the segment ($p_i(\mathrm{C}) = p(w_i^c)$).

In addition to the confidence score (correct probability), we propose a method for deriving the substitution, insertion and deletion error probabilities from a WCN. The *Substitution error probability* of a segment is obtained by summing the posterior probabilities in the segment excluding the highest posterior probability (confidence score, i.e. correct probability, of the best word) and the posterior probability of the null word ($p_i(\mathrm{S}) = p(w_i^d) + p(w_i^e)$). If this substitution error probability is high, the WCN estimates that the best word in the segment (w_i^c) is incorrect and a substitution error (S) occurs.

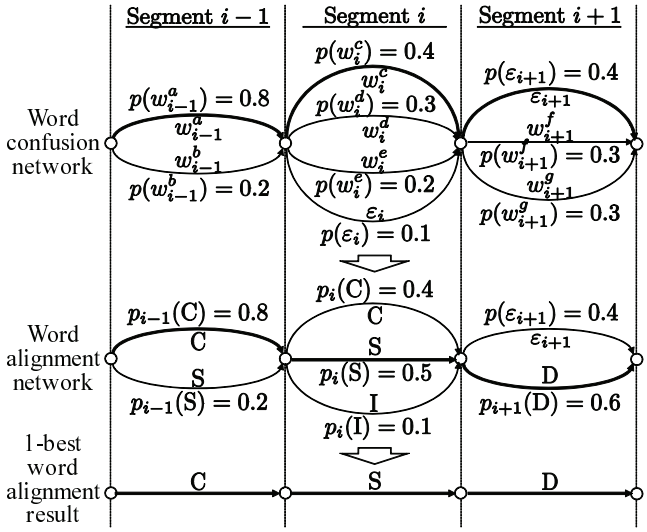

Fig. 1. Derivation of an alignment network and 1-best alignment result from a word confusion network. Best paths are drawn with bold curve lines.

When the posterior probability of a null word is not the highest in a segment (e.g. segment i), it is equal to the *Insertion error probability* of the segment ($p_i(\mathrm{I}) = p(\varepsilon_i)$). If this insertion error probability is high, the WCN estimates that the best word in the segment (w_i^c) may be incorrect and an insertion error (I) occurs. If the posterior probability of a null word is the highest in a segment (e.g. segment $i+1$), no recognized word is selected from this segment as a word in the 1-best recognition result, that is, the null word is selected. And the *Deletion error probability* of this segment is obtained by summing the posterior probabilities of all competing words against the null word in the segment ($p_{i+1}(\mathrm{D}) = p(w_{i+1}^f) + p(w_{i+1}^g)$). If this deletion error probability is high, the WCN estimates that a deletion error occurs in the segment.

As a result of the above procedure, we can obtain an *alignment network* with the correct, substitution error, insertion error and deletion error probabilities as shown at the middle of Fig. 1. Furthermore, by selecting the best path in the alignment network, we can obtain the *1-best alignment result* as shown at the bottom of Fig. 1. This 1-best alignment result is equivalent to that usually obtained with a DP procedure between a recognized word sequence and its reference transcription (correct word sequence) when we calculate the word accuracy of a recognition result [8]. We can expect the *alignment features* obtained from a WCN (the probabilities in an alignment network and the 1-best alignment result) can improve the performance of the error type classification. In addition, in a preliminary experiment, we found that the OOV word utterances are highly correlated with a particular alignment pattern obtained with DP (Section 3.2), and thus, we can also expect the WCN-based alignment features to improve the performance of OOV word detection. Moreover, we expect to be able to estimate the word accuracy of a recognition result directly from the WCN-based alignment features without using the reference transcription.

3. ERROR TYPE CLASSIFICATION METHOD

As with [5], we used conditional random fields (CRF) [9] as the discriminative classifiers for the estimation of error types, confidence and OOV words. In the following, we describe the features and labels used in the CRFs.

Table 1. Features of a recognized word. WCN-based alignment features are written with bold characters.

ID	Feature
1	Recognized word itself
2	Part-of-speech (POS)
3	Language model back-off behavior
4	Number of frames
5	Number of phonemes
6	Average number of frames per phoneme
7	**Confidence score (correct probability)**
8	**Substitution error probability**
9	**Insertion error probability**

3.1. Features

A feature vector of a recognized word is obtained by a main speech recognition process and certain additional processes, e.g. the conversion of a recognition lattice to a WCN. The features used in the experiments are listed in Table 1. The first seven features are basic ones [5]. The confidence score (correct probability), a WCN-based feature, is included in the basic features (7th).

In addition to the above seven basic features, we added the substitution and insertion error probabilities obtained from an alignment network shown at the middle of Fig. 1 as the WCN-based features (8th and 9th). We do not (cannot) use the deletion error probability, since we estimate the label of the *actual* recognized word in a recognition result not the null word. We do not use the 1-best alignment result shown at the bottom of Fig. 1, since such a *hard* decision result will often be harmful to the estimation. Moreover, we do not use the other WCN-based features, e.g. those proposed in [4], since we want to clarify the pure effectiveness of the substitution and insertion error probabilities. With reference to [1, 5], all features excluding the 1st–3rd features are quantized into seven bins with a uniform-occupancy binning function.

It has been reported that the contextual information of the features improves the performance of confidence estimation [5] and OOV word detection [4]. We also employ the contextual features in our experiments. With the pattern "pxsxny", we represent the preceding and succeeding x feature contexts in addition to the current feature with maximum y feature n-grams at each feature dimension. On the basis of a preliminary experiment, we changed x from 0 (i.e. only current feature) to 2 and changed y from 1 to 3. As a result, there are seven feature patterns as follows: p0s0n1, p1s1n1, p1s1n2, p1s1n3, p2s2n1, p2s2n2 and p2s2n3.

3.2. Labels

In error type classification, the label for a recognized word takes three values, i.e. the recognized word is Correct (C), a Substitution error (S) or an Insertion error (I). It does not take the value D, i.e. a Deletion error, since we estimate the label of the *actual* recognized word in a recognition result not the null word. In confidence estimation, a label takes binary value that indicates the recognized word is correct (0) or incorrect (1). We make these C/S/I and correct/incorrect labels by making an alignment between the 1-best recognition result and its reference transcription by using the NIST SCLITE scoring package [8].

In OOV word detection, we have to define the segment that is *influenced* by an OOV word utterance. We define this segment based on an alignment result obtained with SCLITE as shown in Fig. 2. In this example, we consider that an OOV word utterance "dissimilar" can influence not only a directly corresponding recognized word "similar" but also one preceding and one succeeding recognized word "this" and "about". "similar" and "this" are in the influenced segment and misrecognized. Thus, we give them a label that indicates "the uttered word that corresponds to the segment of

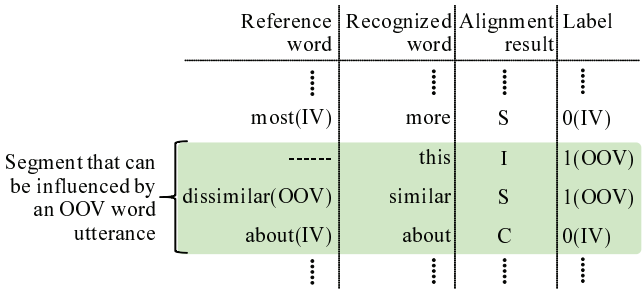

Fig. 2. Labeling in an OOV word utterance segment. The alignment pattern "I+S+C" is frequently observed in OOV word utterance segments.

the recognized word is an OOV word." "about" is also in the influenced segment, however, it is correctly recognized. Thus, we give it a label that indicates "the recognized word is an in-vocabulary (IV) word." This example well explains the correlation between the OOV word utterances and the alignment patterns. An OOV word utterance tends to be misrecognized as a sequence of short words, and thus, as shown in this example, a particular alignment pattern "I+S+C" is very frequently observed in the OOV word utterance segments. In a preliminary experiment, we confirmed that this alignment pattern occurs in more than 25% of the OOV word utterance segments.

4. EXPERIMENTS

We conducted initial experiments to assess the effectiveness of the WCN-based alignment features (the substitution and insertion error probabilities) in the estimation of error types (CSI), confidence, OOV words and word accuracy. All the experiments were performed with our speech recognition platform SOLON [10] using the MIT lecture speech corpus [11].

4.1. Error Type Classification Results

An HMM-based acoustic model was discriminatively trained by using 110 hours (104 lectures) of speech data with a differenced maximum mutual information (dMMI) criterion [12]. It had 2565 states and each state had 32-mixture Gaussian pdfs. A word trigram language model was trained by using 6.2M words of manually transcribed lecture speech. The vocabulary size of the lexicon was 16.5k and there were 48 POS classes.

The CRF training data consisted of 207 hours (228 lectures) of speech data (1.92M words, some lectures are also included in the acoustic model training data). We ran speech recognition against this CRF training data, generated the confusion networks, extracted the features described in Section 3.1 and provided labels using the procedure described in Section 3.2. Then, we trained individual CRFs for the error type (CSI) classification, confidence estimation and OOV word detection with the seven feature context patterns described in Section 3.1. In total, we obtained 21 CRFs.

The evaluation data consisted of 7 hours (8 lectures) of speech data (6.2k utterances, 72k words). We conducted feature extraction against this evaluation data using the procedure that we employed for the CRF training data. The OOV rate was 3.07%. The 1-best recognition results were obtained from WCNs and the word accuracy was 72.16%. Then, by using each of 21 CRFs, we conducted the error type (CSI) classification, confidence estimation and OOV word detection.

Table 2 shows the accuracies of the error type (CSI) classification, confidence estimation and OOV word detection *without* and *with* the WCN-based alignment features (the substitution and insertion error probabilities). It can be seen that the less frequently occurring labels are more difficult to detect. However, we can confirm that the WCN-based alignment features steadily improve the estimation accuracies of all labels, especially those of the less frequently occurring labels. The selected feature patterns indicate the importance of considering the feature contexts and n-grams as reported in [4, 5]. In particular, the error type (CSI) classification needs larger n-grams (3-grams). The CRF for the error type (CSI) classification solves the three-class classification problem. However its detection performance with respect to the correct word (C) is the same as that of the confidence estimation, i.e. correct/incorrect two-class classification.

Table 3 shows weight rankings for the WCN-based alignment features (including the confidence score (correct probability)) in each CRF. We can confirm that reasonable weights are given to the features in each CRF. This is especially obvious in the CRF for the error type (CSI) classification, i.e. each of the correct / substitution error / insertion error probabilities is given a large weight in each of the C/S/I detections. In the CRF for the OOV word detection, larger weights are given to all three features. This indicates that the WCN-based alignment features contribute to the detection of the particular alignment pattern "I+S+C" in the segments of the OOV word utterances as shown in Fig. 2.

4.2. Word Accuracy Estimation Results

There are two methods for estimating the word accuracy directly from the WCN-based alignment features *without* using the reference transcriptions. The first method uses the 1-best alignment result shown at the bottom of Fig. 1. The second method uses the probabilities associated with the alignment network shown in the middle of Fig. 1. In the second method, to calculate the word accuracy, i.e. count up the number of CSIDs, of a recognition result, we sum up the correct, substitution error, insertion error and deletion error probabilities for all the segments in a WCN. However, since only

Table 2. Accuracies of the error type (CSI) classification, confidence estimation (correct/incorrect classification) and OOV word detection (IV/OOV classification) *without* and *with* the WCN-based alignment features (the substitution and insertion error probabilities). #T: number of true labels, #R: number of retrieved labels, #C: number of correct labels in the retrived labels, Recall=#C/#T, Precision=#C/#R, F-measure=2·Recall·Precision/(Recall+Precision), Feature pattern: one that gives highest estimation accuracy.

Label	#T	#R		#C		Recall		Precision		F-measure		Feature pattern
		w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	
C	55613	58313	58031	51080	51048	**0.919**	0.918	0.876	**0.880**	0.897	**0.898**	p2s2n3
S	31231	12295	12421	6957	7072	0.526	**0.535**	0.566	**0.569**	0.545	**0.551**	
I	3450	1686	1842	686	771	0.199	**0.224**	0.407	**0.419**	0.267	**0.291**	
Correct	55613	57119	57212	50559	50703	0.909	**0.912**	0.885	**0.886**	0.897	**0.899**	p2s2n2
Incorrect	16681	15715	15082	10121	10172	0.607	**0.610**	0.667	**0.674**	0.635	**0.641**	
IV	68009	66074	66347	63984	64307	0.940	**0.946**	0.968	**0.969**	0.954	**0.957**	p2s2n2
OOV	4285	6220	5947	2195	2245	0.512	**0.524**	0.353	**0.378**	0.418	**0.439**	

Table 3. Weight rankings for the WCN-based alignment features (the correct probability (confidence score), substitution error probability and insertion error probability) in each CRF. In all CRFs, the largest weight is given to the recognized word and the second largest weight is given to the POS of the recognized word, since they are not quantized and have more classes than the other features.

Label	Cor. prob.	Sub. err. prob.	Ins. err. prob.
C	**3**	7	5
S	4	**3**	5
I	4	7	**3**
Correct	3	5	6
Incorrect	3	5	6
IV	3	5	4
OOV	3	5	4

the *actual* words in a recognition result are considered and the null words (e.g. segment $i+1$ in a WCN of Fig. 1) are ignored, we cannot count the number of insertion and deletion errors with the first 1-best estimation method and we cannot sum up the deletion error probabilities with the second probabilistic estimation method.

As with the above WCN-based method, we can estimate the word accuracy from the error type (CSI) classification result (with the WCN-based alignment features) shown in Table 2. As a result of the CSI classification, a recognized word is given the correct, substitution error and insertion error probabilities that are summed to one. And by using these probabilities, we can estimate the word accuracy with the 1-best (i.e. select the highest probability at each recognized word in a recognition result) and probabilistic methods. These "CSI-based" methods take the substitution and insertion errors into consideration, but, they neglect the deletion errors.

Table 4 shows the word accuracies calculated by SCLITE with the reference transcriptions and estimated by the WCN- and CSI-based 1-best and probabilistic estimation methods *without* the reference transcriptions. With these four estimation methods, the reference transcriptions are not used, and thus, the number of words is estimated as #N = #C + #S. The word accuracies of the CSI-based methods, especially the CSI-based probabilistic method, is closer to that of SCLITE than the WCD-based methods. However, as shown in Fig. 3, the correlation of the utterance-level word accuracies obtained with SCLITE and the CSI-based probabilistic method is not very high. The correlation coefficient is 0.71. And from this figure, it can be seen that the lower word accuracies are more difficult to estimate.

5. CONCLUSION AND FUTURE WORK

We have proposed a method for deriving alignment features from a word confusion network (WCN) and used it in error type classification and word accuracy estimation. In the initial experiments, we confirmed that the WCN-based alignment features steadily improves the performance of error type classification, especially that of the insertion error, the less frequent events. And the word accuracy estimated from the error type classification results without using the reference transcriptions was close to that calculated by the SCLITE scoring tool [8] using the reference transcriptions.

However, the total performance of error type classification and the estimation performance of lower word accuracy still remain at unsatisfactory level. To improve the total performance of error type classification, we are planning to add effective features, e.g. those proposed in [2, 3, 4], and investigate the joint estimation of the error types, confidence and OOV words [13]. We expect that by considering the other WCN-based features, e.g. those proposed in [4], we will be able to improve the estimation performance of the lower word accuracy. We are also planning to consider the deletion errors in the error type classification and word accuracy estimation.

Table 4. Word accuracy calculated by SCLITE with the reference transcriptions and those estimated by the WCN- and CSI-based 1-best and probabilistic estimation methods *without* using the reference transcriptions. #N: number of words, #C: number of correctly recognized words, #S: number of substitution error words, #I: number of insertion error words, #D: number of deletion words, %Correct=100·#C/#N, Word Accuracy=100·(#C−#I)/#N.

Method	#N	#C	#S	#I	#D	%Cor.	WACC
SCLITE	72283	55613	13231	3450	3439	76.94	72.16
WCN 1-best	72294	68183	4111	0	0	94.31	94.31
WCN prob.	69836	61876	7960	2458	0	88.60	85.08
CSI 1-best	70452	58031	12421	1842	0	82.37	79.76
CSI prob.	68662	54838	13824	3632	0	**79.87**	**74.58**

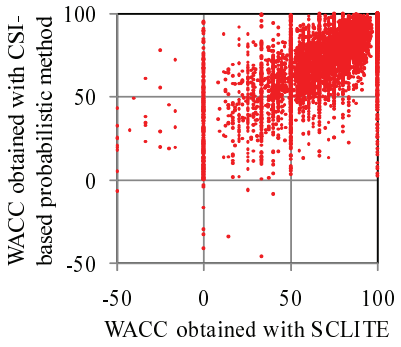

Fig. 3. Correlation of utterance-level word accuracies obtained with SCLITE and CSI-based probabilistic method.

6. REFERENCES

[1] C. White, J. Droppo, A. Acero and J. Odell, "Maximum entropy confidence estimation for speech recognition," Proc. ICASSP, pp. 809–812, 2007.

[2] C. White, G. Zweig, L. Burget, P. Schwarz and H. Hermansky, "Confidence estimation, OOV detection, and language ID using phone-to-word transcription and phone-level alignments," Proc. ICASSP, pp. 4085–4088, 2008.

[3] S. Kombrink, L. Burget, P. Matějka, M. Karafiát and H. Hermansky, "Posterior-based out of vocabulary word detection in telephone speech," Proc. Interspeech, pp. 80–83, 2009.

[4] B. Lecouteux, G. Lingarès and B. Favre, "Combined low level and high level features for out-of-vocabulary word detection," Proc. Interspeech, pp. 1187–1190, 2009.

[5] J. Fayolle, F. Moreau, C. Raymond, G. Gravier and P. Gros, "CRF-based combination of contextual features to improve a posteriori word-level confidence measures," Proc. Interspeech, pp. 1942–1945, 2010.

[6] J. Mamou, D. Carmel and R. Hoory, "Spoken document retrieval from call-center conversations," Proc. SIGIR, pp. 51–58, 2006.

[7] L. Mangu, E. Brill and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," Computer Speech and Language, vol. 14, pp. 373–400, 2000.

[8] NIST SCLITE Scoring Package Version 1.5, http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm.

[9] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," Proc. ICML, pp. 282–289, 2001.

[10] T. Hori, C. Hori, Y. Minami and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. ASLP, vol. 15, no. 4, pp. 1352–1365, May 2007.

[11] H.-A. Chang and J.R. Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," Proc. ICASSP, pp. 4481–4484, 2009.

[12] E. McDermott, S. Watanabe and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," Proc. ICASSP, pp. 4894–4897, 2010.

[13] A. Ogawa and A. Nakamura, "A novel confidence measure based on marginalization of jointly estimated error cause probabilities," Proc. Interspeech, pp. 242–245, 2010.