CONFIDENCE OF AGREEMENT AMONG MULTIPLE LVCSR MODELS AND MODEL COMBINATION BY SVM

Takehito Utsuro†, *Yasuhiro Kodama*‡, *Tomohiro Watanabe*‡, *Hiromitsu Nishizaki*‡, *Seiichi Nakagawa*‡

†Dpt. Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
‡Dpt. Information and Computer Sciences, Toyohashi University of Technology
utsuro@i.kyoto-u.ac.jp, {kodama,watanabe,nisizaki,nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

For many practical applications of speech recognition systems, it is quite desirable to have an estimate of confidence for each hypothesized word. Unlike previous works on confidence measures, we have proposed features for confidence measures that are extracted from outputs of *more than one* LVCSR models. For further analysis of the proposed confidence measure, this paper examines the correlation between each word's confidence and the word's features such as its part-of-speech and syllable length. We then apply SVM learning technique to the task of combining outputs of multiple LVCSR models, where, as features of SVM learning, information such as the pairs of the models which output the hypothesized word are useful for improving the word recognition rate. Experimental results show that the combination results achieve a relative word error reduction of up to 72 % against the best performing single model and that of up to 36 % against ROVER.

1. INTRODUCTION

Since current speech recognizers' outputs are far from perfect and always include a certain amount of recognition errors, it is quite desirable to have an estimate of confidence for each hypothesized word. This is especially true for many practical applications of speech recognition systems such as automatic weighting of additional, non-speech knowledge sources, keyword based speech understanding, and recognition error rejection – confirmation in spoken dialogue systems. Most of previous works on confidence measures (e.g., [1]) are based on features available in a single LVCSR model. However, it is well known that a voting scheme such as ROVER (*Recognizer output voting error reduction*) for combining multiple speech recognizers' outputs can achieve word error reduction [2, 3, 4, 5]. Considering the success of a simple voting scheme such as ROVER, it also seems quite possible to improve reliability of previously studied features for confidence measures by simply exploiting more than one speech recognizers' outputs. From this observation, unlike those previous works on confidence measures, we have been studying features for confidence measures that are extracted from outputs of more than one LVCSR models. More specifically, we experimentally evaluated the agreement among the outputs of multiple Japanese LVCSR models, with respect to whether it is effective as an estimate of confidence for each hypothesized word [6, 7].

Our previous study [6] reported that the agreement between the outputs with two different acoustic models can achieve quite reliable confidence, and also showed that the proposed measure of confidence outperforms previously studied features for confidence measures such as the *acoustic stability* and the *hypothesis density* [1]. We also reported evaluation results with 26 distinct acoustic models and identified the features of acoustic models most effective in achieving high confidence [7]. The most remarkable results are as follows: for the newspaper sentence utterances, nearly 99% precision is achieved by decreasing 94% word correct rate of the best performing single model by only 7%. For the broadcast news speech, nearly 95% precision is achieved by decreasing 72% word correct rate of the best performing single model by only 8%.

Based on those results of our previous studies, for further analysis of the proposed confidence measure, this paper examines the correlation between each word's confidence and the word's features such as its part-of-speech and syllable length. As the result of this analysis, to our surprise, we show that functional words such as particles and auxiliary verbs tend to have higher confidence values than content words such as nouns and verbs. We also show that the confidence of each word varies according to its syllable length. Finally, we apply the Support Vector Machine (SVM) [8] learning technique to the task of combining outputs of multiple LVCSR models. A Support Vector Machine is trained for choosing the most confident one among several hypothesized words, where, as features of SVM learning, information such as the pairs of the models which output the hypothesized word, its part-of-speech, and its syllable length are useful for improving the word recognition rate.

Model combination by high performance machine learning techniques such as SVM learning has advantages over that by voting schemes such as ROVER [2] and others [3, 4, 5], especially when the majority of participating models are not reliable. In the model combination techniques based on voting schemes, outputs of multiple LVCSR models are combined according to simple majority vote or weighted majority vote based on confidence of each hypothesized word such as its likelihood. The results of model combination by those voting techniques can be harmed when the majority of participating models have quite low performance and output word recognition errors with high confidence. On the other hand, in the model combination by high performance machine learning techniques such as SVM learning, among those participating models, reliable ones and unreliable ones are easily discriminated through the training process of machine learning framework. Furthermore, depending on the features of hypothesized words such as its part-of-speech and syllable length, outputs of multiple models are combined in an optimal fashion so as to minimize word recognition errors in the combination results.

Experimental results show that model combination by SVM achieves the followings: i.e., for the newspaper sentence utterances, a relative word error reduction of 72 % against the best performing single model and that of 36 % against ROVER; for the broadcast news speech, a relative word error reduction of 39 % against the best performing single model and that of 14 % against ROVER.

2. SPECIFICATION OF JAPANESE LVCSR SYSTEMS

2.1. Decoders

As the decoders of Japanese LVCSR systems, we use the one named Julius, which is provided by IPA Japanese dictation free software project [9], as well as the one named SPOJUS [10], which has been developed in our laboratory. Both decoders are composed of two decoding passes, where the first pass uses the word bigram, and the second pass uses the word trigram.

2.2. Acoustic Models

The acoustic models of Japanese LVCSR systems are based on Gaussian mixture HMM. We evaluate phoneme-based HMMs as well as syllable-based HMMs.

2.2.1. Acoustic Models with the Decoder JULIUS

As the acoustic models used with the decoder Julius, we evaluate phoneme-based HMMs as well as syllable-based HMMs. The following four types of HMMs are evaluated: i) triphone model, ii) phonetic tied mixture (PTM) triphone model, iii) monophone model, and iv) syllable model. Every HMM phoneme model is gender-dependent (male). For each of the four models above, we evaluate both HMMs *with* and *without* the short pause state, which amount to 8 acoustic models in total.

2.2.2. Acoustic Models with the Decoder SPOJUS

The acoustic models used with the decoder SPOJUS are based on syllable HMMs, which have been developed in our laboratory [11]. The acoustic models are gender-dependent (male) syllable unit HMMs. Among various combinations of features of acoustic models[1], we carefully choose 9 acoustic models so that they include the best performing ones as well as a sufficient number of minimal pairs which have difference in only one feature. Then, for each of the 9 models, we evaluate both HMMs *with* and *without* the short pause states, which amount to 18 acoustic models in total.

2.3. Language Models

As the language models, the following two types of word bigram / trigram language models for 20k vocabulary size are evaluated: 1) the one trained using 45 months Mainichi newspaper articles, 2) the one trained using 5 years Japanese NHK (Japan Broadcasting Corporation) broadcast news scripts (about 120,000 sentences).

2.4. Evaluation Data Sets

The evaluation data sets consist of newspaper sentence utterances, which are relatively easier for speech recognizers, and rather harder broadcast news speech: 1) 100 newspaper sentence utterances from 10 male speakers consisting of 1,565 words, selected by IPA Japanese dictation free software project [9] from the JNAS (Japanese Newspaper Article Sentences) speech data [12], 2) 175 Japanese NHK broadcast news (June 1st, 1996) speech sentences consisting of 6,813 words, uttered by 14 male speakers (six announcers and eight reporters).

2.5. Word Recognition Rates

Word correct and accuracy rates of the individual LVCSR models for the above two evaluation data sets are measured, where for the recognition of the newspaper sentence utterances, the language model used is the one trained using newspaper articles, and for the recognition of the broadcast news speech, the language model used is the one trained using broadcast news scripts. Word recognition rates for the above two evaluation data sets are summarized as below:

newspaper sentence utterances		
decoder	word correct (%)	word accuracy (%)
Julius	93.9(max) to 73.8(min)	91.3(max) to 70.3(min)
SPOJUS	91.1(max) to 79.5(min)	86.2(max) to 55.3(min)
broadcast news speech		
decoder	word correct (%)	word accuracy (%)
Julius	72.4(max) to 50.4(min)	69.2(max) to 40.8(min)
SPOJUS	71.5(max) to 55.6(min)	63.9(max) to 38.9(min)

3. A METRIC FOR EVALUATING CONFIDENCE

This section gives the definition of our metric for evaluating confidence. In this paper, we focus on estimating correctly recognized words and evaluate confidence according to recall/precision rates of estimating correctly recognized words. The following gives a procedure for evaluating the agreement among the outputs of multiple LVCSR models as an estimate of correctly recognized words. First, let us suppose that we have two outputs Hyp_1 and Hyp_2 of two LVCSR models, each of which is represented as a sequence of hypothesized words:

$$\begin{aligned} Hyp_1 &= w_{11}, \ldots, w_{1i}, \ldots, w_{1k} \\ Hyp_2 &= w_{21}, \ldots, w_{2j}, \ldots, w_{2l} \end{aligned}$$

Hyp_1 and Hyp_2 are aligned by Dynamic Time Warping. Then, a list named *agreed word list* is constructed by collecting those words w_{1i} (= w_{2j}) that satisfy the constraint: w_{1i} and w_{2j} are aligned together by Dynamic Time Warping, and w_{1i} and w_{2j} are lexically identical. Finally, the following recall/precision rates are calculated by comparing the agreed word list with the reference sentence considering both the lexical form and the position of each word.

$$\begin{aligned} Recall &= \frac{\text{\# of correct words in the agreed word list}}{\text{\# of words in the reference sentence}} \\ Precision &= \frac{\text{\# of correct words in the agreed word list}}{\text{\# of words in the agreed word list}} \end{aligned}$$

4. CORRELATION BETWEEN WORD FEATURES AND CONFIDENCE

As we reported in [7], experimenting with 26 (=8+18) distinct Japanese LVCSR models with various acoustic models, we have evaluated 325 pairs of all the 26 LVCSR models in terms of confidence of agreement between the outputs of the two constituent models. For further analysis of this confidence measure, this section examines the correlation between each word's confidence and the word's features such as its part-of-speech and syllable length.

4.1. Parts-of-Speech of Words

First, in order to examine the correlation between each word's confidence and its part-of-speech, the language models are trained with words annotated with their parts-of-speech[2]. Then, for each of the nine part-of-speech categories of CHASEN, we evaluate the 325 LVCSR model pairs in terms of confidence of agreement between the outputs of the two constituent models. More specifically, for each of the 325 LVCSR model pairs, we evaluate the precision/recall of the agreement between their outputs and plot their precision values in descending order. For the newspaper sentence

[1] Sampling frequencies, frame shift lengths, feature parameters, covariance matrices, and self loop transition / duration control.

[2] Parts-of-speech of words are annotated by the Japanese morphological analyzer CHASEN (`http://chasen.aist-nara.ac.jp/`), where the coarsest nine part-of-speech categories are used in this work.

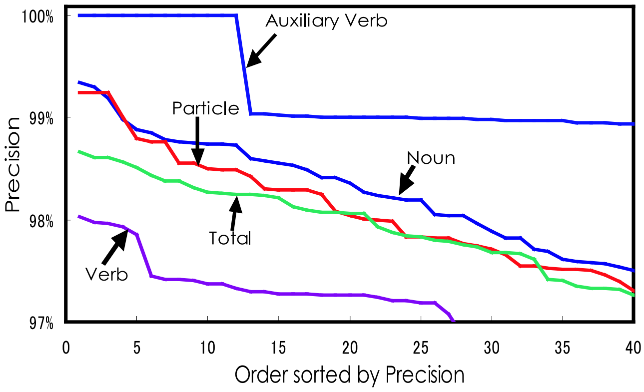

Fig. 1: Distribution of Precision per Part-of-Speech of Words (Newspaper Sentence)

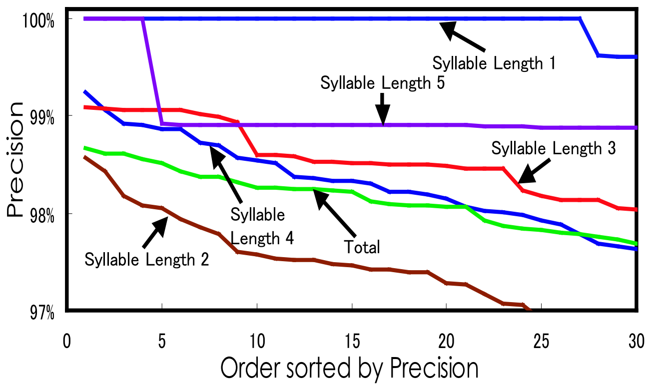

Fig. 2: Distribution of Precision per Syllable Length of Words (Newspaper Sentence)

utterances, Figure 1 gives this plot for each of the most frequent four parts-of-speech categories, i.e., *verb*, *noun*, *particle*, and *auxiliary verb*, as well as for all the part-of-speech categories together in one plot ("Total") (we have similar results for the broadcast news speech)[3].

Generally speaking, to our surprise, functional words such as auxiliary verbs and particles tend to have higher confidence than content words such as verbs and nouns for both speech data, although there exist a few exceptional cases. This tendency coincides well with the perplexity distribution per part-of-speech given in Figure 3 (a). It is also very important to note that model pairs achieving the highest precision values vary according to the part-of-speech categories. For the newspaper sentence utterances, model pairs with the highest precision for the part-of-speech categories other than verbs achieve higher precision than the total best precision. Estimating from the distribution of Figure 1, it seems quite possible to overcome the total best precision by switching the model pair to the one best performing against the part-of-speech of the word at current position.

4.2. Syllable Lengths of Words

Next, this section examines the correlation between each word's confidence and its syllable length. For each of the syllable lengths

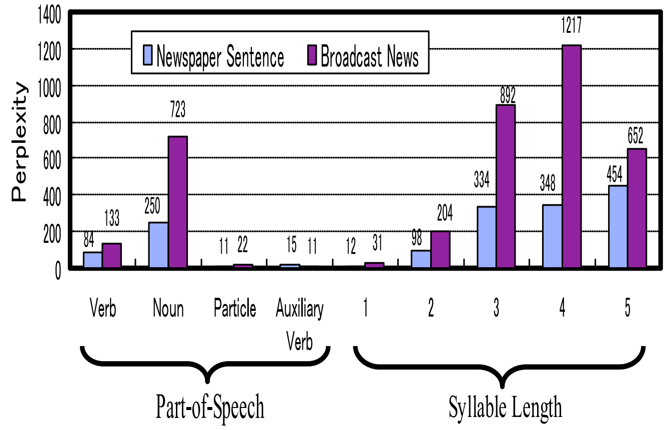

Fig. 3: Distribution of Perplexities per Word Feature

from 1 to 5, we evaluate the 325 LVCSR model pairs in terms of confidence of agreement between the outputs of the two constituent models. For each of the 325 LVCSR model pairs, we evaluate the precision/recall of the agreement between their outputs and plot their precision values in descending order. For the newspaper sentence utterances, Figure 2 gives this plot for each of the syllable lengths from 1 to 5, as well as for all the syllable lengths together in one plot ("Total").

Although the perplexity distribution per syllable length given in Figure 3 (b) shows that the perplexity becomes smaller as the syllable length becomes shorter, Figure 2 shows that the tendency of confidence distribution among different syllable lengths seems rather complicated. (Those tendencies are somehow different between the newspaper sentence utterances and the broadcast news speech.) However, it is still true that model pairs achieving the highest precision values vary according to the syllable lengths. Thus, again, it seems quite possible to overcome the total best precision by switching the model pair to the one best performing against the syllable length of the word at current position.

5. COMBINING OUTPUTS OF MULTIPLE LVCSR MODELS BY SVM

Based on the analysis of the previous section, this section describes the results of applying SVM learning technique to the task of combining outputs of multiple LVCSR models considering the confidence of each word[4]. We divide each of the data sets described in Section 2.4 into two halves[5], where one half is used for training and the other half for testing. A Support Vector Machine is trained for choosing the most confident one among several hypothesized words from the outputs of the 26 LVCSR models[6]. As features of the SVM learning, we use the pairs of the models which output the word, the part-of-speech of the word, and the syllable length of the word[7]. As classes of the SVM learning, we use whether each hypothesized word is correct or incorrect. Since

[3] We exclude the model pairs with recall values below a threshold (80% for the newspaper sentence utterances) from the experimental results in Figures 1 and 2. Then, Figures 1 and 2 show plots for the model pairs within the range of top 30 or 40.

[4] We compared the performance of SVM learning with much simpler machine learning techniques such as decision list learning [13], and found that SVM learning outperforms decision list learning.

[5] It is guaranteed that the two halves do not share speakers.

[6] We used SVM^{light} (`http://www.cs.cornell.edu/People/tj/svm_light/`) as a tool for SVM learning.

[7] We also evaluated the effect of acoustic and language scores of each hypothesized word as features of SVM, where their contribution to improving the overall performance was very little.

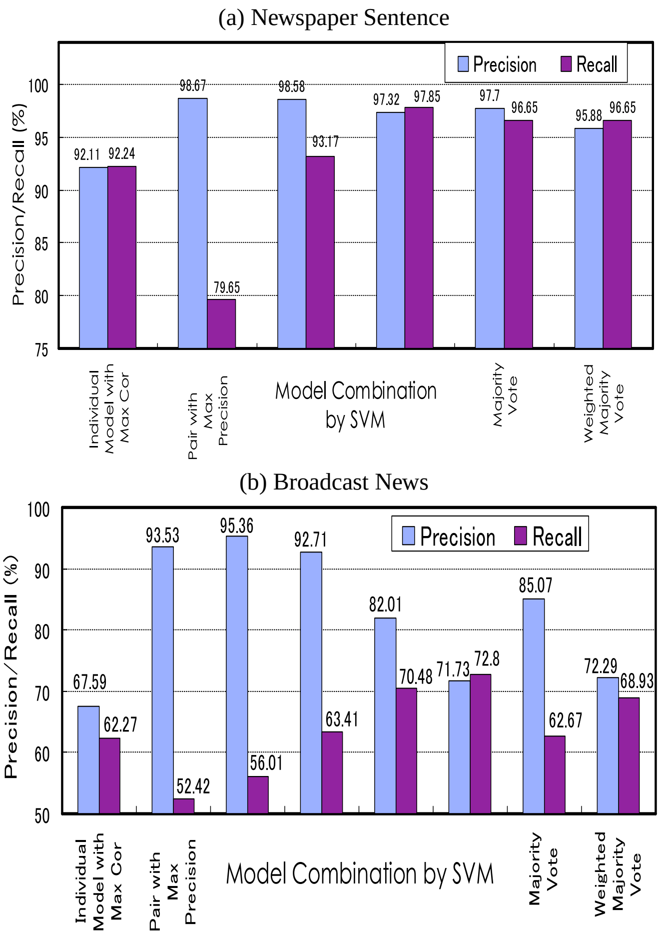

Fig. 4: Evaluation Results of Combining Outputs of Multiple LVCSR Models

Support Vector Machines are binary classifiers, we regard the distance from the separating hyperplane to each hypothesized word as the word's confidence. The outputs of the 26 LVCSR models are aligned by Dynamic Time Warping, and the most confident one among those competing hypothesized words is chosen as the result of model combination. We also require the confidence of hypothesized words to be higher than a certain threshold, and choose the ones with the confidence above this threshold as the result of model combination.

The results of the performance evaluation against the test data are shown in Figure 4 as "Model Combination by SVM", where two or four results are given by changing thresholds of the confidence of each hypothesized word. Furthermore, as baseline performances, that of the best performing single model with respect to word correct rate ("Individual Model with Max Cor"), and that of the model pair with the highest precision value ("Pair with Max Precision") [6, 7] are also shown. The recall rate of model combination by SVM is higher than that of the "Pair with Max Precision" when their precision rates are comparative. Furthermore, for both speech data, model combination by SVM significantly outperforms the best performing single model. Relative word error reduction are 72 % for the newspaper sentence utterances and 39 % for the broadcast news speech (the best correct (= recall) rate achieved by model combination by SVM was 97.85 % for the newspaper sentence utterances and 72.80 % for the broadcast news speech). Figure 4 also shows the performance of ROVER [2] as another baseline, where "Majority Vote" shows the performance of the strategy of outputting no word at a tie, while "Weighted Majority Vote" shows the performance when the word correct rate of each individual model is used as the weight of hypothesized words. As can be seen from those results, model combination by SVM mostly outperforms ROVER for both speech data. Relative word error rate reduction are 36 % for the newspaper sentence utterances and 14 % for the broadcast news speech. For the purpose of further improving the performance of model combination by machine learning such as SVM learning, we are currently working on incorporating richer information (such as the majority voting results by ROVER, and acoustic/language scores of each word) into the machine learning framework as features.

6. CONCLUDING REMARKS

This paper studied features for confidence measures that are extracted from outputs of *more than one* LVCSR models. We examined the correlation between each word's confidence and the word's features such as its part-of-speech and syllable length. We also showed that model combination by SVM achieved a relative word error reduction of up to 72 % against the best performing single model and that of up to 36 % against ROVER.

7. REFERENCES

[1] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. 5th Eurospeech*, 1997, pp. 827–830.

[2] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997, pp. 347–354.

[3] H. Schwenk and J.-L. Gauvain, "Combining multiple speech recognizers using voting and language model information," in *Proc. 6th ICSLP*, 2000, vol. II, pp. 915–918.

[4] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum Bayes-risk ASR voting strategies," in *Proc. 6th ICSLP*, 2000, pp. 139–142.

[5] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. NIST Speech Transcription Workshop*, 2000.

[6] Y. Kodama, T. Utsuro, H. Nishizaki, and S. Nakagawa, "Experimental evaluation on confidence of agreement among multiple Japanese LVCSR models," in *Proc. 7th Eurospeech*, 2001, pp. 2549–2552.

[7] T. Utsuro, T. Harada, H. Nishizaki, and S. Nakagawa, "A confidence measure based on agreement among multiple LVCSR models — correlation between pair of acoustic models and confidence —," in *Proc. 7th ICSLP*, 2002, vol. I, pp. 701–704.

[8] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[9] T. Kawahara et al., "Sharable software repository for Japanese large vocabulary continuous speech recognition," in *Proc. 5th ICSLP*, 1998, pp. 3257–3260.

[10] A. Kai, Y. Hirose, and S. Nakagawa, "Dealing with out-of-vocabulary words and speech disfluencies in an n-gram based speech understanding system," in *Proc. 5th ICSLP*, 1998, pp. 2427–2430.

[11] S. Nakagawa and K. Yamamoto, "Evaluation of segmental unit input HMM," in *Proc. 21st ICASSP*, 1996, pp. 439–442.

[12] K. Itou et al., "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *Proc. 5th ICSLP*, 1998, pp. 3261–3264.

[13] D. Yarowsky, "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French," in *Proc. 32nd ACL*, 1994, pp. 88–95.