An Empirical Study on Multiple LVCSR Model Combination by Machine Learning

Takehito Utsuro[†] **Yasuhiro Kodama**[‡] **Tomohiro Watanabe**[††]
Hiromitsu Nishizaki[‡‡] **Seiichi Nakagawa**[††]
†Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan
‡Sony Corporation ††Toyohashi University of Technology ‡‡University of Yamanashi

Abstract

This paper proposes to apply machine learning techniques to the task of combining outputs of multiple LVCSR models. The proposed technique has advantages over that by voting schemes such as ROVER, especially when the majority of participating models are not reliable. In this machine learning framework, as features of machine learning, information such as the model IDs 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 39 % against the best performing single model and that of up to 23 % against ROVER. We further empirically show that it performs better when LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates.

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., (Kemp and Schaaf, 1997)) 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 (Fiscus, 1997; Evermann and Woodland, 2000). 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, 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.

Our previous study 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* (Kemp and Schaaf, 1997). We also reported evaluation results with 26 distinct acoustic models and identified the features of acoustic models most effective in achieving high confidence (Utsuro et al., 2002). 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, this paper proposes to apply machine learning techniques to the task of combining outputs of multiple LVCSR models. As a machine learning technique, the Support Vector Machine (SVM) (Vapnik, 1995) learning technique is employed. 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 model IDs which output the hypothesized word, its part-of-speech, and the number of syllables 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 and others (Fiscus, 1997; Evermann and Woodland, 2000), 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 the number of syllables, 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 39 % against the best performing single model and that of 23 % against ROVER; for the broadcast news speech, a relative word error reduction of 13 % against the best performing single model and that of 8 % against ROVER. We further empirically show that it performs better when LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates[1].

2 Specification of Japanese LVCSR Systems

2.1 Decoders

As decoders of Japanese LVCSR systems, we use the one named Julius, which is provided by IPA Japanese dictation free software project (Kawahara and others, 1998), as well as the one named SPOJUS (Kai et al., 1998), which has been developed in Nakagawa lab., Toyohashi Univ. of Tech., Japan. 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 Nakagawa laboratory, Toyohashi University of Technology, Japan (Nakagawa and Yamamoto, 1996). The acoustic models are gender-dependent (male) syllable unit HMMs. Among various combinations of features of acoustic models[2], 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 (Kawahara and others, 1998) from the JNAS (Japanese Newspaper Article Sentences) speech data (Itou and others, 1998), 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:

[1] Compared with our previous report (Utsuro et al., 2003), the major achievement of the paper is this empirical result. Utsuro et al. (2003) examined the correlation between each word's confidence and the word's features, and then introduced the framework of combining outputs of multiple LVCSR models by SVM learning.

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

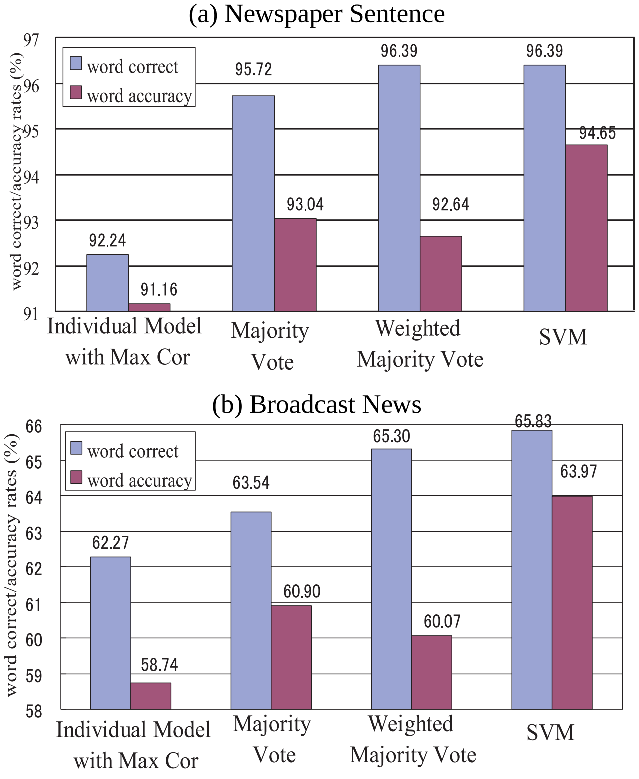

Figure 1: Comparison among Combination by SVM / (Weighted) Majority Votes / Individual Models

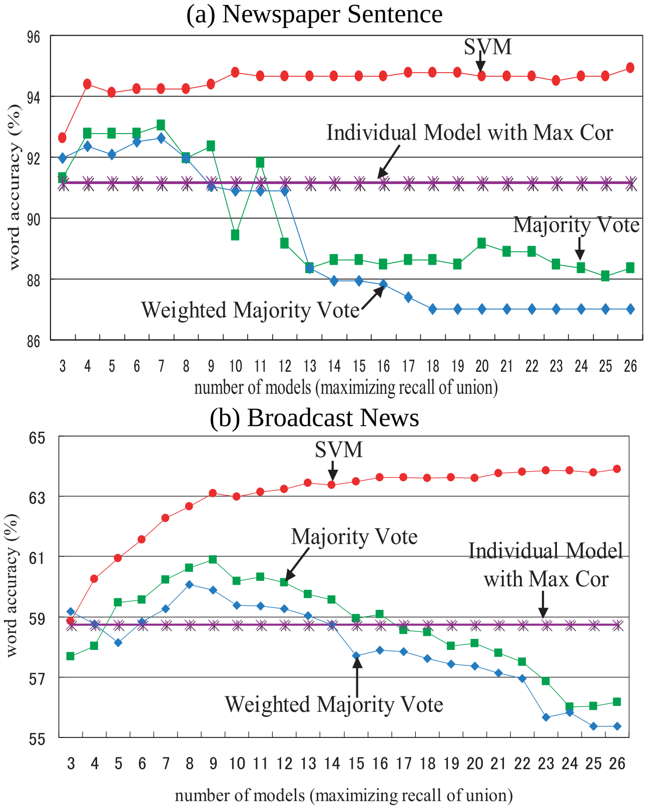

Figure 2: Comparing Methods for Combining Outputs of n $(3 \leq n \leq 26)$ Models

newspaper sentence utterances		
decoder	word correct (%)	word accuracy (%)
Julius	93.0(max) to 72.7(min)	90.4(max) to 69.4(min)
SPOJUS	90.2(max) to 78.1(min)	85.3(max) to 51.0(min)
broadcast news speech		
decoder	word correct (%)	word accuracy (%)
Julius	71.7(max) to 49.0(min)	68.8(max) to 39.7(min)
SPOJUS	70.7(max) to 55.4(min)	62.8(max) to 36.2(min)

3 Combining Outputs of Multiple LVCSR Models by SVM

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. We divide each of the data sets described in Section 2.4 into two halves[3], 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[4]. As features of the SVM learning, we use the model IDs which output the word, the part-of-speech of the word, and the number of syllables[5]. As classes of the SVM learning, we use whether each hypothesized word is correct or incorrect. Since 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 1. All the results in Figure 1 are the best performing ones among those for combining outputs of n $(3 \leq n \leq 26)$ models. The results of model combination by SVM are indicated as "SVM". As a baseline performance, that of the best performing single model with respect to word correct rate ("Individual Model with Max Cor") is shown. (Note that their word recognition rates are those for the half of the whole data set, and thus different from those in Section 2.5.) For both speech data, model combination by SVM sig-

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

[4] We used SVM^{light} (`http://svmlight.joachims.org/`) as a tool for SVM learning. We compared linear and quadratic kernels and the linear kernel performs better.

[5] Contribution of the parts-of-speech and the numbers of syllables was slight. 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.

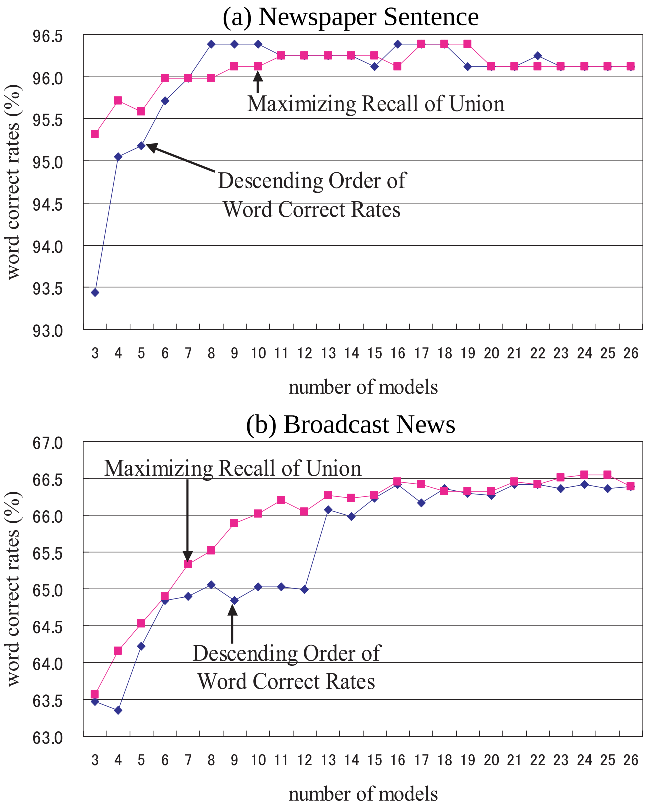

Figure 3: Comparison between Maximizing Recall of Union / Descending Order of Word Correct Rates

nificantly outperforms the best performing single model. In terms of word accuracy rate, relative word error reduction are 39 % for the newspaper sentence utterances and 13 % for the broadcast news speech. Figure 1 also shows the performance of ROVER (Fiscus, 1997) 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, for each individual model, word correct rate for each sentence is estimated and used as the weight of hypothesized words. Model combination by SVM mostly outperforms ROVER for both speech data. In terms of word accuracy rate, relative word error rate reduction are 23 % for the newspaper sentence utterances and 8 % for the broadcast news speech[6].

Figure 2 plots the changes of word accuracy rates against the increasing number of models which participate in LVCSR model combination. Here, LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates. (As we show later, the former outperforms the latter.) It is quite clear from this result that the difference of model combination by SVM and (weighted) majority votes becomes much larger as more and more models participate in model combination. This is because the majority of participating models become unreliable in the second half of the curves in Figure 2.

Figure 3 compares the model selection procedures, i.e., choosing models so as to cover as many correctly recognized words as possible (indicated as "Maximizing Recall of Union"), and choosing models in descending order of their word correct rates (indicated as "Descending Order of Word Correct Rates"). The former performs better in the first half of the curves. This result indicates that, even if recognition error words increase in the outputs of models participating in LVCSR model combination, it is better to cover as many correctly recognized words as possible. This is because, in the model combination by high performance machine learning techniques such as SVM learning, reliable and unreliable hypothesized words are easily discriminated through the training process.

4 Concluding Remarks

This paper proposed to apply the SVM learning technique to the task of combining outputs of multiple LVCSR models. The proposed technique has advantages over that by voting schemes such as ROVER, especially when the majority of participating models are not reliable.

References

G. Evermann and P. Woodland. 2000. Posterior probability decoding, confidence estimation and system combination. In *Proc. NIST Speech Transcription Workshop*.

J. G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pages 347–354.

K. Itou et al. 1998. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proc. 5th ICSLP*, pages 3261–3264.

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

T. Kawahara et al. 1998. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proc. 5th ICSLP*, pages 3257–3260.

T. Kemp and T. Schaaf. 1997. Estimating confidence using word lattices. In *Proc. 5th Eurospeech*, pages 827–830.

S. Nakagawa and K. Yamamoto. 1996. Evaluation of segmental unit input HMM. In *Proc. 21st ICASSP*, pages 439–442.

T. Utsuro, T. Harada, H. Nishizaki, and S. Nakagawa. 2002. A confidence measure based on agreement among multiple LVCSR models — correlation between pair of acoustic models and confidence —. In *Proc. 7th ICSLP*, pages 701–704.

T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, and S. Nakagawa. 2003. Confidence of agreement among multiple LVCSR models and model combination by SVM. In *Proc. 28th ICASSP*, volume I, pages 16–19.

V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

[6] Remarkable improvements are achieved especially in word accuracy rates. This is due to the strategy of requiring the confidence of hypothesized words to be higher than a certain threshold, where insertion error words tend to be discarded.