Experimental Evaluation on Confidence of Agreement among Multiple Japanese LVCSR Models

Yasuhiro Kodama, Takehito Utsuro, Hiromitsu Nishizaki, Seiichi Nakagawa

Department of Information and Computer Sciences
Toyohashi University of Technology
{kodama,utsuro}@cl.ics.tut.ac.jp, {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, this paper studies features for confidence measures that are extracted from outputs of *more than one* LVCSR models. More specifically, this paper experimentally evaluates 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. The results of experimental evaluation show that the agreement between the outputs with two acoustic models which have different units in HMMs, such as phonemes and syllables, can achieve quite reliable confidence.

1. Introduction

Since current speech recognizers' outputs are far from perfect and always include 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 word selection for unsupervised adaptation schemes, automatic weighting of additional, non-speech knowledge sources, keyword based speech understanding, and recognition error rejection – repair dialogues generation in spoken dialogue systems.

Most of previous works on confidence measures (e.g., [1, 2]) 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 [3, 4]. 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, this paper studies features for confidence measures that are extracted from outputs of more than one LVCSR models.

For the purpose of estimating confidence for each hypothesized word, it is more important to examine which combination of existing LVCSR models can achieve high confidence and which combination can not, although even simple voting schemes can achieve word error reduction. Therefore, in this paper, we experimentally evaluate 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. In this evaluation of existing Japanese LVCSR models, we concentrate on evaluating confidence of the agreement among outputs with different acoustic models. The results of experimental evaluation show that the agreement between the outputs with two acoustic models which have different units in HMMs, such as phonemes and syllables, can achieve quite reliable confidence. We also experimentally evaluate several previously studied features for confidence measures such as the *acoustic stability* and the *hypothesis density* [1] for comparison, and show that the proposed measure of confidence outperforms them.

2. Specification of Japanese LVCSR Systems

2.1. 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.

2.1.1. Phoneme Models

The acoustic models based on phoneme HMMs are provided by IPA Japanese dictation free software project [5, 6]. The number of Japanese phonemes for the phoneme HMMs is 43. The speech data are sampled at 16 kHz and 16 bits. The feature parameters consist of 25 dimensions: 12 dimensional mel frequency cepstrum coefficients (MFCC), the cepstrum difference coefficients (delta MFCC), and delta power are calculated every 10 msec. The following three types of HMMs are evaluated:

1. triphone model.
2. phonetic tied mixture (PTM) triphone model,
3. monophone model,

Every HMM phoneme model consists of three states and is gender-dependent (male). The number of Gaussian mixtures of a HMM state is 16 for the monophone and triphone models, and 64 for the PTM triphone model.

2.1.2. Syllable Models

The acoustic models based on syllable HMMs are those which have been developed in our laboratory [7]. The number of Japanese syllables for the syllable HMMs is 114. The sampling frequency is 12 kHz and the frame shift length is 8 msec. The following four types of the sets of feature parameters are evaluated:

1. 20 dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames, delta 10 dimensions calculated over 9 successive frames, delta delta 10 dimensions and delta, delta delta powers (henceforth "MFCC-seg").

2. 10 dimensional mel frequency cepstrum coefficients (MFCC), delta delta 10 dimensions and delta, delta delta powers (henceforth "MFCC-frm").
3. 20 dimensional LPC mel cepstrum segmented from 4 successive frames, delta 10 dimensions calculated over 9 successive frames, delta delta 10 dimensions and delta, delta delta powers (henceforth "LPC-seg").
4. 10 dimensional LPC mel cepstrum, delta 10 dimensions, delta delta 10 dimensions and delta, delta delta powers (henceforth "LPC-frm").

The acoustic models are gender-dependent (male) syllable unit HMMs that have 5 states 4 densities, 4 Gaussian mixtures models per density with full covariance matrices.

2.2. Language Models

As the language models, the following three types of word bigram and trigram language models for 20k vocabulary size are evaluated:

1. the one trained using 75 months Mainichi newspaper articles, provided by IPA Japanese dictation free software projects [5, 6].
2. the one trained using 45 months Mainichi newspaper articles.
3. the one trained using 5 years Japanese NHK[1] broadcast news scripts (about 120,000 sentences).

2.3. Decoders

As the decoders of Japanese LVCSR systems, we use the one named JULIUS, which is provided by IPA Japanese dictation free software project [5, 6], as well as the one named SPOJUS [8], which has been developed in our laboratory. As the acoustic models of those two decoders, we use the phoneme HMMs for JULIUS and the syllable HMMs for SPOJUS. 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].

2.4. Evaluation Data Sets

The evaluation data sets consist of newspaper sentence utterance, which is relatively easier for speech recognizers, as well as 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 [5, 6] from the JNAS (Japanese Newspaper Article Sentences) speech data [9].
2. 175 Japanese NHK broadcast news (June 1st, 1996) speech sentences consisting of 6,813 words, uttered by 10 male speakers (two announcers and eight reporters).

2.5. Word Recognition Rates

Word correct ("Cor.") and accuracy ("Acc.") rates of the individual LVCSR models for the above two evaluation data sets are listed in Table 1, where the specification of language models is summarized as below:

[1] Japan Broadcasting Corporation.

[2] We temporarily use two distinct decoders for the phoneme and syllable acoustic models. Presently, we assume that the difference of the search strategies of the two decoders are negligible for our purpose of evaluating the confidence of the agreement among individual LVCSR models. It is quite straightforward to run JULIUS with context-independent syllable HMMs and currently we are working on it.

Table 1: Word Recognition Rates of Individual LVCSR Models (%)

Acoustic Models		Newspaper Sentence		Broadcast News	
		Cor.	Acc.	Cor.	Acc.
Phoneme Models (MFCC -frame)	Triphone (1st)	85.4	78.3	59.0	50.2
	Triphone (2nd)	**91.3**	**87.7**	62.6 (62.3)	52.9 (56.1)
	PTM	89.9	87.1	**63.1**	**56.4**
	Monophone	79.9	77.6	46.8	41.3
Syllable HMMs	MFCC-seg	**87.5**	**84.7**	**63.3**	**57.5**
	MFCC-frm	86.0	83.3	61.6	56.0
	LPC-seg	86.1	82.9	62.7	53.8
	LPC-frm	80.1	77.3	55.6	49.2

1. for the recognition of the newspaper sentence utterances, the language model is trained using:[3]
   (a) the 75 months newspaper articles when phoneme HMMs are used as the acoustic models,
   (b) the 45 months newspaper articles when syllable HMMs are used as the acoustic models.
2. for the recognition of the broadcast news speech, the language model is trained using the 5 years broadcast news scripts.

In the table, recognition rates of both the first and the second passes are shown for the triphone model. Furthermore, for the second pass, the table also includes with parenthesis the rates of recognizing the broadcast news speech with the language model trained using the 75 months newspaper articles.

Among the syllable HMMs, the one with MFCC-seg feature parameters has the highest word recognition rates (**boldfaced**) both for the newspaper sentence utterances and for the broadcast news speech. Among the phoneme HMMs, the triphone model has the highest word recognition rates (**boldfaced**) for the newspaper sentence utterances, while for the broadcast news speech, the PTM triphone model has the highest word recognition rates (**boldfaced**).

3. Experimental Results

This section describes the results of evaluating the agreement among the outputs of multiple LVCSR models as an estimate of confidence for each hypothesized word.

3.1. A Metric for Evaluating Confidence

First, we give the definition of our metric for evaluating confidence. In principle, the task of estimating confidence for each hypothesized word is to have an estimate of which words of the outputs of LVCSR models are likely to be correct and which are not reliable. In this paper, however, 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 n outputs $Hyp_1, \ldots, Hyp_n$ of n LVCSR models, each

[3] The reason why the sizes of these training texts differ is just temporary one. We are currently working on applying the syllable HMMs with the language model provided by IPA Japanese dictation free software project.

Table 2: Evaluation Results of Agreement between Two Acoustic Models: Lower Left: for Newspaper Sentence, Upper Right: for Broadcast News (Recall / Precision (%))

		Phoneme Models (MFCC-frm)			Syllable HMMs			
		Triphone	PTM	Monophone	MFCC-seg	MFCC-frm	LPC-seg	LPC-frm
Phoneme Models (MFCC-frm)	Triphone	—	54.4 / 84.2	40.0 / 91.6	50.8 / 93.8	49.7 / 93.6	49.9 / 92.3	45.1 / 94.0
	PTM	88.2 / 93.0	—	40.7 / 91.2	**51.5 / 93.5**	50.4 / 93.7	50.7 / 92.9	45.8 / 94.4
	Monophone	78.5 / 93.7	78.3 / 92.9	—	39.6 / 94.3	38.7 / 94.7	38.8 / 94.4	35.6 / 94.5
Syllable HMMs	MFCC-seg	**81.7 / 99.4**	80.8 / 98.9	74.7 / 98.3	—	56.9 / 82.5	55.4 / 87.3	50.8 / 88.0
	MFCC-frm	80.6 / 99.2	79.7 / 98.7	73.7 / 98.1	84.2 / 92.6	—	54.4 / 88.3	50.1 / 87.5
	LPC-seg	80.5 / 99.1	79.7 / 98.7	73.3 / 98.2	80.2 / 94.0	82.1 / 93.2	—	50.6 / 87.0
	LPC-frm	78.0 / 99.2	77.3 / 98.9	71.3 / 98.4	82.8 / 94.1	80.1 / 93.0	80.4 / 93.1	—

Table 3: Evaluation Results of Agreement among Individual Models (Recall / Precision (%))

Combination of the Models		Newspaper Sentence	Broadcast News
2 LMs (newspaper articles, broadcast news scripts) with Triphone		—	53.7 / 86.5
1st and 2nd passes (Triphone, MFCC-frame)		81.0 / 92.7	57.9 / 66.1
Weighting of Acoustic/ Language Scores	Triphone (MFCC-frame)	86.9 / 93.2 $\sim$ 82.4 / 93.9	50.7 / 87.6 $\sim$ 44.4 / 91.6
	Syllable HMM (MFCC-seg)	82.7 / 92.2 $\sim$ 76.2 / 94.5	58.5 / 81.1 $\sim$ 44.7 / 90.1
N-best Hypotheses	Triphone (MFCC-frame)	82.4 / 93.1 $\sim$ 66.0 / 94.5	64.0 / 82.0
	Syllable HMM (LPC-seg)	80.4 / 90.6 $\sim$ 39.9 / 98.3	42.4 / 74.4

of which is represented as a sequence of hypothesized words. Next, n sequences $Hyp_1, \ldots, Hyp_n$ of hypothesized words are aligned by DP matching. Then, words that are aligned together and have an identical lexical form throughout n hypotheses $Hyp_1, \ldots, Hyp_n$ are collected into a list named *agreed word list*. For example, suppose that we have two sequences Hyp_1 and Hyp_2 of hypothesized words as below:

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

Then, the *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 DP matching, 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.

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

3.2. Agreement between Two Acoustic Models

In order to evaluate the agreement between the outputs of LVCSR models with various acoustic models, for every pair of the available acoustic models, we collect the agreed word list and evaluate it against the reference sentences. Both for the newspaper sentence utterances and for the broadcast news speech as evaluation data sets, Table 2 shows recall/precision rates for all the pairs of the available acoustic models. The lower left half of the table shows the results for the newspaper sentence utterances, while the upper right half shows those for the broadcast news speech.

In general, these results clearly indicate that the agreement between the output with phoneme HMMs and that with syllable HMMs tends to achieve highest precision rates. For the purpose of highly reliable detection of correctly recognized words, achieving high precision is more important than achieving high recall. Following this point of evaluation, in the case of the newspaper sentence utterances, we can judge that the highest precision with sufficiently high recall (**boldfaced**) is achieved by the pair of the triphone model and the syllable HMMs with MFCC-seg feature parameters, those which have the highest word recognition rates among the phoneme HMMs and the syllable HMMs, respectively. Also in the case of the broadcast news speech, we can judge that the (almost) highest precision with sufficiently high recall (**boldfaced**) is achieved by the pair of the PTM model and the syllable HMMs with MFCC-seg feature parameters, those which have the highest word recognition rates among the phoneme HMMs and the syllable HMMs, respectively.

3.3. Agreement between Two Language Models

For the broadcast news speech, we also evaluate the agreement between the outputs with two language models. This evaluation is with the triphone model as the acoustic model, and the two language models are the one trained using 75 months newspaper articles and the one trained using 5 years broadcast news scripts. The result of recall/precision rates is in the row of "2 LMs" in Table 3. This result indicates that the agreement between the outputs with two language models can not achieve a precision as high as that with two acoustic models such as phoneme/syllable HMMs.

3.4. Agreement between First and Second Passes

With the triphone model as the acoustic model, we also evaluate the agreement between the outputs of the first and the second passes of the decoder. The result of recall/precision rates is in the row of "1st and 2nd passes" in Table 3. The precision is not as high as that with two acoustic models such as phoneme/syllable HMMs.

3.5. Agreement among Different Weighting of Acoustic/Language Scores

The following two sections experimentally compare the proposed measure of confidence based on the agreement among the outputs of multiple LVCSR models with those previously studied features for confidence measures.

In this section, we evaluate a feature similar to the *acoustic stability* [1] in our evaluation task, because the acoustic stability is one of the features that are effective in estimating confidence for each hypothesized word. In this evaluation, first we compute 10 alternative hypotheses with different weighting of acoustic/language model scores, where the different weighting is selected around the best weighting[4]. Then, for every possible subset of those 10 alternative hypotheses, we evaluate the agreement among the constituent hypotheses. Results with the highest recall as well as those with the highest precision are listed in Table 3. Again, in this evaluation, a precision as high as that with two acoustic models such as phoneme/syllable HMMs can not be achieved.

3.6. Agreement among N-best Hypotheses

Next, in this section, we evaluate a feature similar to the *hypothesis density* [1] in our evaluation task, because the hypothesis density is also one of the features that are effective in estimating confidence for each hypothesized word. In this evaluation, n-best (in this experiment, 200-best) hypotheses of a single LVCSR model are aligned together and the hypothesized words that appear in majority or all of the n-best hypotheses are evaluated for their confidence. For the newspaper sentence utterances, Table 3 shows the results with the highest precision together with recall closed to the best boldfaced result (81.7%) in Table 2, for each of the triphone model and the syllable HMMs (LPC-seg). For the broadcast news speech, a single best result is shown for each of the triphone model and the syllable HMMs (LPC-seg). Again, in this evaluation, a precision as high as that with two acoustic models such as phoneme/syllable HMMs can not be achieved.

3.7. Discussion

From the results of the experimental evaluation of the previous sections, it can be summarized that the agreement between the outputs with two acoustic models which have different units in HMMs, such as phonemes and syllables, performs best. One of the remarkable achievements is that the proposed feature for estimating confidence, i.e., the agreement between the outputs with two acoustic models, is found to have quite high precision with less than 10% loss of recall from the baseline (recall with a single LVCSR model). The achieved precision is over 99% when the baseline word recognition rate is around 90%, and over 93% even when the baseline word recognition rate is less than 60%.

One promising approach to utilizing this highly reliable feature for the purpose of estimating confidence for each hypothesized word is to optimally integrate many candidate features through machine learning techniques as in [1]. In the case of the broadcast news speech, especially, results in Tables 2 and 3 show that there exist several other useful features with higher precisions, such as the agreement of some other pair of two acoustic models, and the agreement among different weighting of acoustic/language scores. In this case, it is quite possible to improve recall with small loss of precision by selectively applying the most appropriate features through machine learning techniques.

[4]The lowest word recognition rates among those weighting are: in the case of the newspaper sentence utterances, 86.9% (Cor.) / 79.3% (Acc.) with the triphone, and 80.1% (Cor.) / 77.3% (Acc.) with the syllable HMMs (MFCC-seg): and in the case of the broadcast news speech, 61.9% (Cor.) / 45.9% (Acc.) with the triphone, and 52.3% (Cor.) / 46.3% (Acc.) with the syllable HMMs (MFCC-seg).

Another very interesting direction for further research is to examine acoustic features as well as linguistic features of those words that are estimated as unreliable. If some useful feature is discovered through careful examination, there could be several possible solutions to improve word recognition rates of those unreliable words: i) selectively apply an acoustic model tailored to words that are rejected by some confidence measures, ii) in order to improve confidence of those unreliable words, selectively add training data for improving the acoustic models through some active learning framework, iii) develop a language model tailored to those unreliable words, if there exist some useful linguistic features such as parts-of-speech for those unreliable words.

4. Concluding Remarks

This paper 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. The results of experimental evaluation showed that the agreement between the outputs with two acoustic models which have different units in HMMs, such as phonemes and syllables, can achieve quite reliable confidence.

5. References

[1] T. Kemp and T. Schaaf. Estimating confidence using word lattices. In *Proceedings of the 5th Eurospeech*, pages 827–830, 1997.

[2] F. Wessel, K. Macherey, and H. Ney. A comparison of word graph and N-best list based confidence measures. In *Proceedings of the 6th Eurospeech*, pages 315–318, 1999.

[3] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, 1997.

[4] H. Schwenk and J.-L. Gauvain. Combining multiple speech recognizers using voting and language model information. In *Proceedings of the 6th ICSLP*, volume II, pages 915–918, 2000.

[5] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 5th ICSLP*, pages 3257–3260, 1998.

[6] K. Itou, K. Shikano, T. Kawahara, K. Takeda, A. Yamada, A. Itou, T. Utsuro, T. Kobayashi, N. Minematsu, M. Yamamoto, S. Sagayama, and A. Lee. IPA Japanese dictation free software project. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1343–1350, 2000.

[7] S. Nakagawa and K. Yamamoto. Evaluation of segmental unit input HMM. In *Proceedings of the 21st ICASSP*, pages 439–442, 1996.

[8] 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 *Proceedings of the 5th ICSLP*, pages 2427–2430, 1998.

[9] K. Itou, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, S. Itahashi, and M. Yamamoto. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proceedings of the 5th ICSLP*, pages 3261–3264, 1998.