A CONFIDENCE MEASURE BASED ON AGREEMENT AMONG MULTIPLE LVCSR MODELS
— CORRELATION BETWEEN PAIR OF ACOUSTIC MODELS AND CONFIDENCE —

Takehito Utsuro, Tetsuji Harada, Hiromitsu Nishizaki, and Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology
{utsuro,tetsu}@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 LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. Furthermore, among various features of acoustic models based on Gaussian mixture HMMs, it is concluded that ones such as whether or not to have short pause models, as well as different units in HMMs (e.g., triphone model or syllable model) are the most effective in achieving highly reliable confidence.

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 word selection for unsupervised adaptation schemes, 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, 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 decoders and/or different acoustic models. The results of experimental evaluation show that the agreement between the outputs with two LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. Furthermore, among various features of acoustic models based on Gaussian mixture HMMs, it is concluded that ones such as whether or not to have short pause models, as well as different units in HMMs (e.g., triphone model or syllable model) are the most effective in achieving highly reliable confidence. It is also shown that it is better to combine various features including those most effective ones than to use one of those most effective features alone.

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 [5], as well as the one named SPOJUS [6], 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. Julius is with word-trellis searches and hence has much broader search space than SPOJUS, which is with N-best searches.

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 number of Japanese phonemes for the phoneme HMMs is 43, while the number of Japanese syllables for the syllable HMMs is 124. 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 four types of HMMs are evaluated:

1. triphone model
2. phonetic tied mixture (PTM) triphone model
3. monophone model
4. syllable model
Every HMM phoneme model consists of three states and is gender-dependent (male). The number of Gaussian mixtures of an HMM state with diagonal covariance matrices is 16 for the monophone, triphone, and syllable models, and 64 for the PTM triphone model. For each of the four models above, we evaluate both HMMs with and without the short pause state1, which amount to eight 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 [7]. The number of Japanese syllables for the syllable HMMs is 116. The sampling frequencies are 12 kHz / 16 kHz and the frame shift lengths are 8 msec / 10 msec. The following three types of the sets of feature parameters are evaluated:

- $2m$ dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames, delta $m$ dimensions calculated over 9 successive frames, delta delta $m$ dimensions and delta, delta delta powers ($m = 10, 12$).
- 12 dimensional mel frequency cepstrum coefficients (MFCC), delta 12 dimensions, delta delta 12 dimensions and delta, delta delta powers.

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 / diagonal covariance matrices. We also switch between conventional HMM with self loop transition and HMM with duration control, and evaluate both of them.

Among various combinations of features such as the sampling frequencies, frame shift lengths, feature parameters, covariance matrices, and self loop transition / duration control, we carefully choose nine 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, as in the case of the acoustic models used with the decoder Julius, for each of the nine models, we evaluate both HMMs with and without the short pause states2, 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 NHK3 broadcast news scripts (about 120,000 sentences).

2.4. Evaluation Data Sets
The evaluation data sets consist of newspaper utterance, which is 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 [5] from the JNAS (Japanese Newspaper Article Sentences) speech data [8].

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:

<table>
<thead>
<tr>
<th></th>
<th>word correct (%)</th>
<th>word accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julius</td>
<td>93.9(max) to 73.8(min)</td>
<td>91.3(max) to 70.3(min)</td>
</tr>
<tr>
<td>SPOJUS</td>
<td>91.1(max) to 79.5(min)</td>
<td>86.2(max) to 55.3(min)</td>
</tr>
</tbody>
</table>

For each of the four models above, we evaluate both HMMs dependent (male). The number of Gaussian mixtures of an HMM is 116.

Every HMM phoneme model consists of three states and is gender-dependent (male). The number of Gaussian mixtures of an HMM state with diagonal covariance matrices is 16 for the monophone, triphone, and syllable models, and 64 for the PTM triphone model. For each of the four models above, we evaluate both HMMs with and without the short pause state1, which amount to eight 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 [7]. The number of Japanese syllables for the syllable HMMs is 116. The sampling frequencies are 12 kHz / 16 kHz and the frame shift lengths are 8 msec / 10 msec. The following three types of the sets of feature parameters are evaluated:

- $2m$ dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames, delta $m$ dimensions calculated over 9 successive frames, delta delta $m$ dimensions and delta, delta delta powers ($m = 10, 12$).
- 12 dimensional mel frequency cepstrum coefficients (MFCC), delta 12 dimensions, delta delta 12 dimensions and delta, delta delta powers.

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 / diagonal covariance matrices. We also switch between conventional HMM with self loop transition and HMM with duration control, and evaluate both of them.

Among various combinations of features such as the sampling frequencies, frame shift lengths, feature parameters, covariance matrices, and self loop transition / duration control, we carefully choose nine 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, as in the case of the acoustic models used with the decoder Julius, for each of the nine models, we evaluate both HMMs with and without the short pause states2, 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 NHK3 broadcast news scripts (about 120,000 sentences).

2.4. Evaluation Data Sets
The evaluation data sets consist of newspaper utterance, which is 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 [5] from the JNAS (Japanese Newspaper Article Sentences) speech data [8].

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:

<table>
<thead>
<tr>
<th></th>
<th>word correct (%)</th>
<th>word accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julius</td>
<td>93.9(max) to 73.8(min)</td>
<td>91.3(max) to 70.3(min)</td>
</tr>
<tr>
<td>SPOJUS</td>
<td>91.1(max) to 79.5(min)</td>
<td>86.2(max) to 55.3(min)</td>
</tr>
</tbody>
</table>

3. A METRIC FOR EVALUATING CONFIDENCE
This section gives 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 two outputs $Hyp_1$ and $Hyp_2$ of two LVCSR models, each of which is represented as a sequence of hypothesized words. Next, two sequences $Hyp_1$ and $Hyp_2$ of hypothesized words are aligned by DP matching. Then, words that are aligned together and have an identical lexical form are collected into a list named agreed word list. Suppose that we have two sequences $Hyp_1$ and $Hyp_2$ of hypothesized words as below:

$Hyp_1 = w_{11}, \ldots, w_{1k}$
$Hyp_2 = w_{21}, \ldots, w_{2l}$

Then, the agreed word list is constructed by collecting those words $w_{1j} (= w_{2j})$ that satisfy the constraint: $w_{1j}$ and $w_{2j}$ are aligned together by DP matching, and $w_{1j}$ 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.

$Recall = \frac{\# of correct words in the agreed word list}{\# of words in the reference sentence}$
$Precision = \frac{\# of correct words in the agreed word list}{\# of words in the agreed word list}$

4. 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.
4.1. Agreement between Two Decoders

First, we evaluate correlation between difference of decoders and confidence. We classify 325 pairs of all the 26 LVCSR models (eight for the decoder Julius and 18 for the decoder SPOJUS) according to the pairs of decoders, i.e., Julius-SPOJUS, Julius-Julius, and SPOJUS-SPOJUS. Then, 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. Figure 1 gives this plot for those with recall values above a threshold (80% for the newspaper sentence utterances) for each of the decoder pairs. It is quite clear from this result that agreement between the outputs from different two decoders can achieve higher confidence than the same two decoders. This is mainly because the two decoders Julius and SPOJUS have search spaces quite different in their breadth.

The maximum precision values for the same decoder pairs are, however, just 1% lower than that of the different decoder pairs, achieving sufficiently high confidence. We have quite similar results also with the broadcast news speech4, where their maximum precision is almost 95%. As for the recall values for those maximum precision pairs, they are around 84% for the newspaper sentence utterances and 64% for the broadcast news speech. This means that, for the newspaper sentence utterances, nearly 99% precision is achieved by decreasing the word correct rate (= recall rate) by only 7%, and for the broadcast news speech, nearly 95% precision is achieved by decreasing the word correct rate (= recall rate) by only 8%.

4.2. Agreement between Two Acoustic Models

Next, among various features of acoustic models, this section examines which ones are the most effective in achieving high confidence. For this purpose, out of the 325 pairs of all the 26 LVCSR models, we focus on pairs with the same decoders (i.e., Julius-Julius or SPOJUS-SPOJUS), which have differences only in their acoustic models.

Although most results in this paper are given only for the newspaper sentence utterances, we have obtained quite similar results also for the broadcast news speech and the claims of this paper hold for both speech data.

4.2.1. Contribution of Difference of a Single Feature of Acoustic Models

First, in order to evaluate contribution of difference of a single feature of acoustic models to achieving high confidence, we examine precision values for minimal pairs which have difference in only one of those features. Features of acoustic models examined are as follows: for the decoder Julius, units in HMMs (i.e., the four types listed in Section 2.2.1), and with/without short pause state, while for the decoder SPOJUS, all the features described in Section 2.2.2, i.e., sampling frequencies, frame shift lengths, feature parameters, covariance matrices, self loop transition / duration control, and with/without short pause states. For each of those features, we evaluate the precision/recall of the agreement between minimal pairs and plot their precision values in descending order. Figure 2 gives this plot for those with recall values above the threshold. In terms of the maximum precision values, the best performing features are “with/without short pause states” for the decoder SPOJUS and “units in HMMs” for the decoder Julius. “With/without short pause states” for the decoder Julius performs

---

4: Although most results in this paper are given only for the newspaper sentence utterances, we have obtained quite similar results also for the broadcast news speech and the claims of this paper hold for both speech data.
4.2.2. Summary

Next, in order to evaluate contribution of difference of features of acoustic models in general cases, we classify difference of features of acoustic models into the following six categories:

1. pairs of models with more than one differences in their features including with/without short pause states (decoder pair: Julius-Julius, SPOJUS-SPOJUS)
2. pairs of models with difference only in with/without short pause states (decoder pair: SPOJUS-SPOJUS)
3. pairs of models with difference only in units in HMMs (decoder pair: Julius-Julius)
4. pairs of models with short pause states which have more than one differences in their features (decoder pair: SPOJUS-SPOJUS)
5. pairs of models without short pause states which have more than one differences in their features (decoder pair: SPOJUS-SPOJUS)
6. pairs of models other than all of the above, with difference in only one feature (decoder pair: Julius-Julius, SPOJUS-SPOJUS)

This result suggests that the difference in the language model might help achieving high confidence and thus future works definitely include introducing various language models such as those with syntactic structures, trigger models, and topic categories into this task.

Then, Figure 4 gives the precision value of a single model pair with the maximum precision value for each of the six categories. From these results, putting more emphasis on the performance against the harder speech, i.e., the broadcast news speech, we conclude to represent the degree of the contribution of those acoustic features according to the inequalities below:

\[ 1 > 2, 3, 4 > 5 > 6 \]

5. 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 LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. This paper reported evaluation results that were significantly different from those of our previous study [9]. Our previous study [9] 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]. On the other hand, this paper reported evaluation results with 26 distinct acoustic models, which were much more than the seven models studied in [9], and, more importantly, identified the features of acoustic models most effective in achieving high confidence.

6. REFERENCES