Confidence Measure and Rejection Based on Correctness Probability of Recognition Candidates

Norihide Kitaoka,1,* Ichiro Akahori,1 and Seiichi Nakagawa2

1Denso Corporation, Kariya, 448-8661 Japan
2Faculty of Engineering, Toyohashi University of Technology, Toyohashi, 441-8580 Japan

SUMMARY

We propose a confidence measure expressing the correctness probability of recognition candidates, to be applied to a speech dialog system using a dialog strategy based on the degree of confidence of speech recognition candidates. We assume a function expressing the correctness probability of the feature parameters associated with the recognition candidate, and its parameters are estimated by using the square error of the correct (1) and incorrect (0) data as the evaluation function. For two feature quantities, namely, the likelihood ratio of the recognized word and syllable-concatenated model, and the variance of the syllable duration inside a word, the correctness probability can be expressed by a sigmoid function. The correctness probability can also be expressed for a combination of these two feature parameters. Rejection experiments involving evaluation of the degree of discrimination between the correct and incorrect data show that the degree of discrimination is higher when the two feature parameters are combined than when they are used individually. © 2004 Wiley Periodicals, Inc. Syst Comp Jpn, 35(11): 91–103, 2004; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20046

Key words: speech recognition; correctness probability; confidence measure; likelihood ratio; variance of duration.

1. Introduction

A problem that often arises in speech dialog systems is that inconsistency in state between the system and the users occurs due to misrecognition by the system, so that the dialog breaks down. To avoid this difficulty, a dialog strategy that takes account of misrecognition may be considered. But if the dialog always considers the possibility of misrecognition, the number of utterances for recognition will be increased, and the dialog will generally become redundant and confused.

To deal with this problem, a dialog strategy using a confidence measure may be considered. This strategy may consist of changing the response in accordance with the confidence measures of the recognition candidates [1, 5]. The ratio of the likelihood of the acoustic model used in recognition to the likelihood of a competition model in the same region is often used as a confidence measure for speech recognition candidates [6]. For example, it has been shown to be effective for distinguishing between the results of in-vocabulary word recognition or of inputting an out-of-vocabulary word. The likelihood ratio constitutes an approximation of the posterior probability of the recognition candidate given a speech observation sequence [7]; however, its meaning is difficult to understand intuitively, and in addition, since it can take a wide range of values, it is difficult to use as a confidence measure of correct or incorrect data.

© 2004 Wiley Periodicals, Inc.
Many other feature quantities have also been proposed as confidence measures for speech recognition candidates, including the duration of subword segments such as phonemes or syllables [2, 3], the use of a score assigned by a language model [2], and the use of candidates from multiple recognizers in which the weights of the language model and the acoustic model are different [4].

When considering the construction of a dialog system by using these confidence measures, the term “correctness probability” should express the extent of intuitive reliance on the recognition candidate in terms of the value of the confidence measure. Thus, in this paper we study a method of determining the “correctness probability” from various confidence measures such as the likelihood ratio. In this case the correctness probability is the probability that a recognition candidate that is output under certain conditions by the recognition system will be correct. The conditions are quantitatively expressed in terms of certain feature parameters to be described later. In this way, a dialog strategy can reflect the recognition conditions. If the correctness probability of a recognition candidate is 95%, the dialog can proceed with confidence, but if it is 50%, it may be necessary to consider misrecognition.

As feature parameters used to obtain the correctness probability, we shall employ the likelihood ratio described above and the variance of syllable duration inside a word. The latter feature makes use of the property of the Japanese language that the duration of syllables is roughly constant over short periods of time, such as within a word morae isochronism. We first express a method of obtaining the correctness probability from a feature parameter, and then extend the method for multiple feature parameters.

Even if the correctness probability is accurately obtained for feature parameters such as the likelihood ratio, it makes no sense if the ability to discriminate correct and incorrect data is poor. For example, if a correctness probability is always approximately 50%, the system has no leeway in selecting the dialog strategy. Therefore, to evaluate the discrimination ability, we shall evaluate rejection performance of misrecognitions by applying threshold to the correctness probability. In addition, the application of the correctness probability to the rejection of out-of-vocabulary words will be considered, and the performance will also be evaluated.

2. Confidence Measures Based on the Correctness Probability

2.1. The correctness probability

We first define the correctness probability. Suppose that a recognition candidate \( W \) is output by a speech recognition system. When this recognition candidate \( W \) is correct, the situation is expressed by \( C(W) = 1 \), and the situation in which it is incorrect will be expressed by \( C(W) = 0 \). In addition, let the value of some feature parameter \( x \) for \( W \) be \( x_W \). We then consider the probability that \( W \) is correct when the value of feature parameter \( x \) for recognition candidate \( W \) is \( x_W \):

\[
p(C(W) = 1| x = x_W)
\]

This value may be considered as intuitively expressing the reliability of recognition candidate \( W \) when the feature quantity \( x = x_W \) is obtained.

For example, let us consider the use of confidence measures in a spoken dialog system incorporating speech recognition. If the dialog system accepts recognition candidates with a correctness probability of 90%, the dialog can proceed without confirmation by concluding that such candidates are almost certainly correct. But when the correctness probability is 50%, misrecognition is considered, and it may be necessary to confirm to the user whether the previous recognition result was correct or not.

2.2. Expression of the correctness probability and its estimation

When the value \( x_W \) of feature parameter \( x \) is obtained for recognition candidate \( W \), let us consider a method of determining the correctness probability of \( W \).

[Method 1: estimation using Bayes’ rule]

The correctness probability may be considered as a posterior probability as follows:

\[
p(C(W) = 1| x = x_W) = \frac{p(x = x_W, C(W) = 1)}{p(x = x_W)}
\]

where

\[
p(x = x_W) = p(x = x_W, C(W) = 1) + p(x = x_W, C(W) = 0)
\]

and

\[
p(x = x_W, C(W) = 1) = p(x = x_W | C(W) = 1) \cdot p(C(W) = 1)
\]

\[
p(x = x_W, C(W) = 0) = p(x = x_W | C(W) = 0) \cdot p(C(W) = 0)
\]

The first terms on the right-hand sides of Eqs. (4) and (5) are the conditional probability that feature parameter \( x \)
takes the value \( x_W \) when the recognition candidate is correct or incorrect, respectively, and can be estimated in advance by assuming a probability function with, say, a normal distribution. The second terms on the right-hand sides of Eqs. (4) and (5) are respectively the recognition rate and the misrecognition rate. By using these, we can determine the posterior probability, that is, the correctness probability, from Eq. (2). In the experiments described below, the probability density function is assumed to be a normal distribution.

[Method 2: direct estimation by assuming the functional form of the correctness probability]

Let us consider the case in which \( x \) takes discrete values. For all possible \( x \), the samples in which \( x_W = x \) are collected, and the correctness probability \( N_C/(N_C + N_I) \) can be determined beforehand from the number of samples of correct recognition \( N_C \) and misrecognition \( N_I \) among them. However, when \( x \) is continuous, it is impossible to determine the probability by obtaining many samples for a specific value.

We therefore assume that the correctness probability is a function \( f(x) \) of \( x \), namely,

\[
p(C(W) = 1|x = x_W) \equiv f(x_W)
\]

In this case the problem of estimating the correctness probability in advance becomes the problem of estimating the parameters of \( f(x) \).

We denote the sample recognition candidates for parameter estimation as \( \{W_1, W_2, \ldots, W_N\} \), the feature quantities corresponding to them as \( \{x_W^1, x_W^2, \ldots, x_W^N\} \), and the sequence expressing their correctness or incorrectness as \( \{C(W_1), C(W_2), \ldots, C(W_N)\} \), \( C(W_n) \in \{0, 1\} \) (where 1 indicates correct, and 0, incorrect). We then minimize

\[
E = \sum_{n=1}^{N} (C(W_n) - f(x_W^n))^2
\]

In other words, the correctness probability function can be estimated by estimating the parameters of \( f(x) \) that minimize the squared errors between indicators of correct or incorrect and the value of the function. We can estimate the correctness probability for each bin of \( x \)-axis by counting correct and incorrect samples. Then a function which approximates the correctness probability can be derived by minimum mean square error estimation. This method can be extended by making the bin width infinitely small to derive the estimation using Eq. (7).

In method 1, we must make the estimation by assuming the probability density functions \( p(x = x_W|C(W) = 1) \), \( p(x = x_W|C(W) = 0) \); but in method 2, that assumption is not necessary, and they can be directly determined by assuming only the correctness probability function.

Recently, research using the same correctness probability \( p(C(W) = 1|x = x_W) \) as ours has been performed independently by Wendemuth and colleagues [8]. They propose a method for determining the best threshold for discrimination of correct and incorrect data [they include out-of-vocabulary (OOV) words as misrecognitions] by using the cross-entropy between the correctness probability function and the samples. However, their objective is not to express the correctness probability as a function of the feature quantity, as in our case.

2.3. Examples of estimation of the correctness probability

We performed estimations of the correctness probability for the experimental results of a large-vocabulary word recognition. The recognition experiment used the large-vocabulary recognition system described in Ref. 16. The task was a car navigation system, the vocabulary size for recognition was about 180,000 words, including place names throughout Japan (about 100,000 words), and facility names (about 70,000 words), and the commands for operating the car navigation system comprised about 200 words. The speech samples were place names recorded inside automobiles by four male speakers whose ages ranged from the 20s to the 50s, and utterances of car navigation system commands (1162 samples). The place names were utterances that included locality names down to the “cho” level (such as Aichi Prefecture, Kariya-shi, Showa-cho), and the facility names were utterances in which the name of the facility was followed by a prefecture name (such as Aichi Prefecture, Nagoya Airport). The commands were terms for navigational operations, such as “widen,” “destination setting,” and so on. Almost equal numbers of these two types (names and commands) were included in the samples. Information on the correctness or incorrectness of these recognition candidates was provided as learning data. The recognition rate of speaker-independent speech recognition was 80.1%.

[Estimation example 1: correctness probability based on likelihood ratio]

Let us consider the determination of the correctness probability of a recognized word. The ratio of the likelihood (difference of logarithmic likelihood) of a speech recognition candidate for a certain word to the likelihood of the competition model prepared separately has often been used as a confidence measure [6, 7, 13]. The likelihood ratio is a measure that can effectively classify recognition candidates into correct and incorrect data or utterances of out-of-vocabulary words. As the competition model we used a syllable-concatenation model [9] that can freely connect Japanese syllables.
The likelihood ratio was used in logarithmic form: using the logarithmic likelihood \( l_w \) of the recognition candidate, the logarithmic likelihood \( l_{sc} \) of the syllable-concatenation model, and the duration \( T \) of the word, the ratio is expressed as the difference in the time-normalized logarithmic likelihoods:

\[
x_W = (l_w - l_{sc})/T
\]  

This will henceforth be called the LLR (log likelihood ratio).

With the horizontal axis, representing the LLR, divided into 10 sections, Fig. 1(a) shows histograms of the numbers of correct and incorrect data in each section. The number of correct data \( (N_C) \) and the number of incorrect data \( (N_I) \) are shown by solid and dashed lines, respectively. The bar graph in Fig. 1(b) is the correctness probability of the recognition candidates, the logarithmic likelihood \( \ell_{sc} \) of the catenation model, and the duration \( T \) of the word, the ratio is expressed as the difference in the time-normalized logarithmic likelihoods:

\[
x_W = (l_w - l_{sc})/T
\]  

In method 1, we must determine the probability density functions \( p(x = x_{wl}(W)) \) of the correct and incorrect data. In Fig. 1(a), the probability density under the assumption of a normal distribution is shown by dotted lines. Appropriate vertical scaling was applied to fit the lines to the histograms. The correctness probability function determined from Eq. (2) by using the recognition rate \( (p(C(W) = 1) = 0.801) \) and the misrecognition rate \( (p(C(W) = 0) = 1 - p(C(W) = 1) = 0.199) \) is shown by a solid line in Fig. 1(b).

In method 2, the correctness probability function should be assumed; based on the bar graph in Fig. 1(b), a monotonically increasing function \( f(x) \) can be assumed for LLR. Here the function \( f(x) \) is assumed to be a sigmoid function

\[
f(x) = \frac{1}{1 + \exp(-ax + b)}
\]  

The errors between the sigmoid function \( f(x) \) and the training data \( C(W_n) \) are evaluated by means of Eq. (7), and the parameters \( a, b \) that minimize it are estimated. In Fig. 1(b), the learning data \( [x_{wl}, C(W)] \) are shown by circles and the function \( f(x) \) estimated from them is represented by a dashed line.

In both methods 1 and 2, we can see that the results express the sawtoothed line of the bar graph. In particular, in method 2, the expression in terms of a sigmoid function and the method of minimum mean square error parameter estimation are appropriate for this purpose.

Methods for transforming feature parameters such as LLR by means of the sigmoid function have been proposed in the past. They were used to reduce the dynamic range of the feature parameters [12] or for weighting the likelihood during decoding by treating the value as a probability [13]. In this investigation, using Eq. (7) as the evaluation function for parameter estimation, we assign it the meaning of the correctness probability.

[Estimation example 2: correctness probability based on variance of syllable duration within a word]

In Japanese, within a short time range (such as within a word), the duration of syllables is almost constant (morae isochronism) and thus the variance of the syllable duration is small.

However, this duration is hard to use in recognition with the conventional HMM. Although there are many methods for enhancing recognition accuracy by taking duration into account, in general they use the shortest and
longest duration of subword units such as syllables or phonemes, or estimate the duration distribution in advance and introduce its probability into the likelihood calculation, but the relation between the duration of subwords within a word is not considered. Irregular warping can occur in matching wrong models to the speech, resulting in misrecognition with relatively high likelihood.

Suppose that the durations of syllables \(s_1, s_2, \ldots, s_N\) within a word are \(d_1, d_2, \ldots, d_N\). We shall attempt to use the equation

\[
x_W^2 = \frac{1}{N} \sum_{n=1}^{N} d_n^2 - \left( \frac{1}{N} \sum_{n=1}^{N} d_n \right)^2
\]

(10)

that is, the variance of the syllable duration within a word (the feature quantity \(x_W\) employed is actually the standard deviation, which is the square root of the variance) is used as a measure for discriminating between correct and incorrect data. 

Henceforth this will be called the VSD (variance of syllable duration). When the speech is aligned with the correct HMM, the estimated syllable duration will be almost correct, and the variance of that duration is likely to become smaller. Therefore, in correspondence with the wrong HMM, unreasonable warping is likely to occur. In particular, this kind of warping is especially likely to occur when matching an utterance with an HMM consisting of a number of states different from that of the correct HMM.

The VSD can be obtained from the result of Viterbi alignment of the recognition candidate. 

For the VSD thus obtained, the histograms of the correct and incorrect data of the recognition candidates, divided into 10 segments in the VSD-axis direction, as in the case of the LLR, are shown by solid and dashed lines in Fig. 2(a). The correct values concentrate near 13, while the incorrect values tend to be larger, but the mode of the value is near 13. The correctness probabilities for each segment are shown by the bar graph in Fig. 2(b). We can find that the value almost monotonically decreases.

The probability density functions \(p(x = x_W | C(W) = 1)\) and \(p(x = x_W | C(W) = 0)\) of the correct and incorrect data determined by method 1 are shown by dotted lines in Fig. 2(a). The correctness probability function determined from Eq. (2) by using these functions, the recognition rate and misrecognition rate, are shown by solid lines in Fig. 2(b). We may assume that the confidence measure is higher when the VSD is smaller, and is lower when the VSD is larger. Even from the bar graph in Fig. 2(b), it is reasonable to assume a monotonically decreasing function \(f(x)\) for the VSD. In this case, the sigmoid function expressed by Eq. (9) is also assumed, and the values of \(a\) and \(b\) are estimated by means of Eq. (7). In Fig. 2, \((x_{Wc}, C(Wb))\) are represented by circles and the estimated function \(f(x)\) is shown by a dashed line.

In method 1 (solid line), there is a large deviation from the bar graph. This is because the probability density functions of the correct and incorrect data are assumed to follow the normal distribution. But in Fig. 2(a), particularly for the incorrect data, we see that the real distribution and the probability density function differ greatly. When using this method, a more suitable form must be assumed for the probability density function. In addition, because the var-
ances of the two probability density functions differ greatly, we have instead obtained a function which decreases when VSD becomes very small.

On the other hand, method 2 (dashed line) gives good results. In this method, since the correctness probability is directly assumed, it will not be affected by inappropriate assumption of the probability density functions of the correct and incorrect data. In addition, it appears important that, by assuming a sigmoid function, accurate estimation has been achieved without losing the monotonic property that the confidence measure is higher when the VSD is smaller.

Even in the bar graph, when VSD is small, the correctness probability tends to drop somewhat. In this respect, we may say that method 1 is appropriate to approximate the tendency. But since the agreement with the overall bar graph is poor, and the reliability of the probability value with small VSD is poor because of the lack of samples, we may conclude that method 2 is superior.

3. Correctness Probability Using Multiple Feature Parameters

In Section 2 we define the correctness probability as the posterior probability of LLR or VSD. But it is also possible to define the correctness probability as the posterior probability for multiple concurrent feature parameters:

$$p(C(W) = 1|x = x_W) \quad (11)$$

Here $x$ expresses the set of multiple feature parameters. Let us now consider the method for obtaining the correctness probability by using LLR and VSD.

Figures 3(a) and 3(b) show histograms for correct and incorrect recognition candidates, respectively. The recognition experiment is the same as that described in Section 2.3. Partitioning is performed in the LLR and VSD directions. The correctness probability for every mesh point determined from them is shown in Fig. 3(c).

We next investigate methods of expressing these distributions as functions.

3.1. Expression as posterior probability

Even when multiple feature parameters are used, Eq. (2) can be applied by determining the probability density functions of the correct and incorrect data, similarly to method 1 when an individual feature parameter is used. But in this case, $x$ is the set of multiple feature parameters.

Figures 3(d) and 3(e) show the estimates of the probability density functions

$$p(x = x_W | C(W) = 1) \quad (12)$$

denote the feature parameters of the correct and incorrect data, respectively, under the assumption of a normal distribution (full covariance). Figure 3(f) shows the correctness probability determined from Eq. (2) by using (d) and (e) as well as the recognition rate ($p(C(W) = 1) = 0.801$) and the misrecognition rate ($p(C(W) = 0) = 1 - p(C(W) = 1)$).

Figure 3(f) approximates (C) well. But, as when a single feature parameter was used, the correctness probability drops with decreasing VSD when the value of VSD is less than 10 in Fig. 3(f), which is not valid.

3.2. Direct determination of the correctness probability function

Let us now consider parameter estimation by assuming a function that will not lose its monotonic characteristic along every axis. We let $x = (x_1, x_2)$, and assume as the correctness probability function

$$f(x_1, x_2) = \frac{1}{1 + \exp(g(x_1, x_2))} \quad (14)$$

For $g(x_1, x_2)$ we use the following simple functions:

Linear combination: $g(x_1, x_2) = a_1 x_1 + a_2 x_2 + a_3$$Bilinear combination: $g(x_1, x_2) = a_1 x_1 x_2 + a_2 x_1 + a_3 x_2 + a_4$

The coefficients $a_i$ are estimated by minimizing Eq. (7). This method is equivalent to method 2 with a single parameter used. Figure 4 shows the curved surfaces estimated for the respective cases. The surfaces are well expressed. In addition, since monotonic behavior in the axial directions is assumed, the correctness probability will not become smaller with decreasing VSD.

3.3. Approximation with the correctness probability functions of single feature parameters

When the correctness probability function resulting from the combination of feature parameters is relatively simple, the estimation method described in Section 3.2 is possible. However, let us consider a method for obtaining values close to the correctness more simply, including cases not meeting the above description.

When the recognition candidate is correct, the correctness probabilities obtained from the individual feature parameters may both become higher. Thus, we may consider the product, representing the AND operation of the correctness probabilities of the two individual feature parameters $f_1(x_1), f_2(x_2)$. Let us express the correctness probability based on two feature parameters in terms of the weighted geometric mean

$$p(x = x_W | C(W) = 1) \quad (13)$$
Fig. 3. Estimation of correct probability function from LLR and VSD as a posterior probability.
The case of \( w = 0.5 \) is shown in Fig. 5. Several methods have been proposed for obtaining confidence measures by combining multiple feature quantities [10, 11, 14]. Even in these methods, better results are believed to be obtained by using a combination close to the AND operation, such as a product or maximum value of the confidence measures, rather than the sum.

The values obtained by dividing the values of Eq. (7) by the number of training samples, namely, the error per sample, are shown in Table 1. The cases in which the geometric mean is 1.0:0.0 or 0.0:1.0 are equivalent to the case of LLR or VSD alone. The use of the two together gives more accurate results than the use of either alone, and the case in which the weights are equal is the most accurate. The results of the posterior probability method described in Section 3.1 and the method of direct determination by a linear or bilinear combination of feature quantities described in Section 3.2 are also shown. The method involving the determination of the posterior probability has a large error, about equivalent to the use of LLR alone. The use of a linear or bilinear combination of feature quantities gives more accurate results, but the geometrical mean with equal weights also gives a good approximation.

\[
f(x_1, x_2) = f_1(x_1)^w \cdot f_2(x_2)^{1-w}
\]

Fig. 4. Expressions of a correct probability using linear and bilinear combinations of two features.

Fig. 5. Expressions of a correct probability using geometrical mean of correct probabilities for LLR and VSD.

<table>
<thead>
<tr>
<th>Method of representation</th>
<th>Mean square error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation as posterior probability</td>
<td>0.1147</td>
</tr>
<tr>
<td>Direct estimation</td>
<td></td>
</tr>
<tr>
<td>Linear combination</td>
<td>0.0961</td>
</tr>
<tr>
<td>Bilinear combination</td>
<td>0.0957</td>
</tr>
<tr>
<td>Weighted geometric mean of single correct-ness probability</td>
<td></td>
</tr>
<tr>
<td>LLR:VSD</td>
<td>0.1296</td>
</tr>
<tr>
<td>0.0:1.0</td>
<td>0.1123</td>
</tr>
<tr>
<td>0.25:0.75</td>
<td>0.1059</td>
</tr>
<tr>
<td>0.5:0.5</td>
<td>0.1064</td>
</tr>
<tr>
<td>0.75:0.25</td>
<td>0.1152</td>
</tr>
<tr>
<td>1.0:0.0</td>
<td></td>
</tr>
</tbody>
</table>
4. Rejection Based on Correctness Probability

4.1. Misrecognition rejection experiments

A confidence measure can be regarded as good when the performance of discrimination between correct and incorrect data is high. Thus, applying threshold to the correctness probabilities obtained by the methods described above, we attempted experimentally to detect and reject misrecognitions, and also to evaluate the effectiveness of the confidence measures.

The samples were 1655 utterances by male speakers, recorded separately from those used in the estimation, but under the same conditions as described in Section 2.3. The speaker-independent recognition rate was 89.7%. For reference, the correctness probability functions determined in Section 2.2 and the bar graph of the correctness probabilities determined from the test samples used in the recognition experiments are superimposed in Figs. 6 and 7. Figure 6 presents the case in which LLR is used, and Fig. 7 presents the case in which VSD is used. The correctness probabilities tend to be somewhat lower in regions where the recognition rate of the test samples is somewhat higher. But the recognition rates of the data for estimation and for the test data rarely become 10% closer; and even in such cases, the approximation is relatively good and is sufficiently effective for purposes such as dialog management.

Using \( \theta \) as the threshold, a recognition candidate for which

\[
p(C(W) = 1|x = x_W) < \theta
\]

is regarded as a misrecognition. When the threshold is varied, the rate of correct detection of misrecognitions (rejection rate of misrecognized data) and the rate of erroneous detection of misrecognitions as correct recognitions (rejection rate of correctly recognized data) will change. Since the correctness probability is used as a confidence measure, the range of variation of the threshold is from 0 to 1.

Figure 8 shows the results of rejection experiments in which LLR and VSD are used individually. The horizontal axis represents the rate of wrong rejection of correct recognitions and the vertical axis is the rate of correct detection of misrecognitions. In one method, the results obtained while varying the threshold form a curve. The detection

---

Fig. 6. Correct probabilities of test samples and functional expressions for LLR.

Fig. 7. Correct probabilities of test samples and functional expressions for VSD.

Fig. 8. Results of rejection of misrecognized data using LLR or VSD.
ability is greater when the curve approaches more closely to the upper left of the figure, indicating that the discrimination performance is higher.

When LLR is used as the feature parameter, nearly equal results are obtained by determining the posterior probability from the distributions of the correct and incorrect data (corresponding to the solid line labeled “2 norm. dist.” in this figure) and the case of direct estimation by assuming a sigmoid function (corresponding to the dashed line labeled “Sigmoid”). However, when VSD is used as the feature parameter, there is a difference in the performance of the two methods, and good results are obtained by direct estimation. The reason may be that the probability distributions of the correct and incorrect data are assumed to have normal distributions.

For the method of determining the correctness probabilities from the correct and incorrect data, the results obtained by using the LLR and VSD individually (solid line) and the results of their use in combination (dashed line) are shown in Fig. 9. We see that due to errors in the estimation of VSD, even when both are used, we could not obtain any improvement from the use of LLR alone.

The results of detection experiments involving the direct estimation of sigmoid functions (the dashed line labeled “Sigmoid with linear function” or the dot-dash line labeled “Sigmoid with bilinear function”) and the weighted geometrical means of the correctness probabilities with LLR and VSD used individually are plotted as solid lines in Fig. 10. The ratios shown in this figure are the weights of the correctness probabilities based on LLR and VSD. When LLR alone (corresponding to 1.0:0.0) and VSD alone (corresponding to 0.0:1.0) are compared, the plot approaches the upper left in the case of LLR, indicating good performance. When the weighted geometrical means are used, better results are obtained than when LLR or VSD is used individually. The best result is obtained when the same weight is used, roughly equivalent to the results obtained with correctness probabilities based on linear or bilinear combinations.

4.2. Recognition of out-of-vocabulary word utterances

When utterances of words not belonging to the vocabulary (out-of-vocabulary, OOV) are recognized by the
speech recognition system, the result is generally obtained by matching with words in the vocabulary. LLR is effective for discrimination of in-vocabulary and out-of-vocabulary words [6, 7]. In addition, the syllable durations are often expanded and contracted, as in the case of misrecognition, and since the variance is then large, VSD may also be considered effective for separation. We therefore performed experiments on the rejection of OOV words using the correctness probability expressions in Section 3 as confidence measures.

However, in essence the correctness probability expresses the probability that the recognition candidates for the words within the vocabulary are correct and there is no direct relationship to the detection of OOV words. In addition, when LLR or VSD is transformed by using a sigmoid function, since a monotonically increasing function is transformed into a monotonically increasing or monotonically decreasing function, the performance will be entirely the same as when rejection is performed by direct application of threshold processing to the conventional value of LLR [6, 7]. But in the proposed technique, since the range of values of the confidence measure is limited to the interval from 0 to 1, it suffices to set the threshold within this range.

The experiment was performed as follows. The recognition vocabulary consisted of about 180,000 place names and facility names from throughout the country, excluding the 200 navigation command words, and command utterances (800 samples) were to be recognized. Thus, all utterances would be recognized as OOV. Figure 11 shows plots of the rejection rates of OOV words versus the rate of erroneous rejection of correctly recognized data (based on the results in Section 4.1) as the threshold is varied. Compared with the detection of misrecognition, the plots are strongly inclined toward the upper left, indicating that this task is easier than the detection of misrecognition.

Misrecognition and the recognition of OOV words may be performed by the same mechanism. For example, in the recognition of OOV words, expansion and contraction of syllables similar to the case of misrecognition is likely to occur. Thus, since the correctness probability is defined on the assumption of words within the vocabulary, its meanings will be lost for OOV words; but it may be effective as a feature parameter for recognition.

In addition, as in the case of rejection of misrecognition data, LLR is more effective than VSD, but better results are obtained when the two feature parameters are combined. For example, if the rate of incorrect rejection of correct data is set at 5%, the rejection rates of OOV words are 98% and 97% when LLR and VSD are used individually, but the rejection rate is 100% when they are used with an equal weight. In addition, when the rejection rate of OOV words is set at 90%, the rejection rate of correct data is 0.6% for LLR, but is a very high value of 54% for VSD. Nonetheless, when the two are used with equal weights, the rejection rate of correct data falls to 0.4%.

5. Conclusions

In this paper, to support the construction of a speech dialog system whose dialog strategy is based on confidence measures, we have proposed a method for obtaining confidence measures that express the correctness probability of recognition candidates.

In particular, when LLR or VSD is used as the feature parameters of the recognition candidates, we have shown that when a particular functional form of the correctness probability is assumed, good results can be obtained by estimating the parameters of that function by minimizing the squared error between the function and the training data, with correctness assigned a value of 1 and incorrectness a value of 0.

We have also described a method for expressing the correctness probability for multiple feature parameters. When LLR and VSD are used, we have shown that a good representation can be obtained by directly estimating a sigmoid function with a linear or bilinear combination of LLR and VSD, or by a combination in which the correctness probability based on multiple feature parameters is approximated as the geometrical mean of the correctness probabilities of the individual feature parameters.

The use of rejection experiments to evaluate the degree of separation of correct and incorrect data by means of the correctness probability showed that the accuracy is enhanced when LLR and VSD are combined, compared with their use in isolation. Even in the rejection of out-of-vocabulary words, the accuracy is similarly enhanced when multiple feature parameters are combined.

In the future, to achieve more accurate rejection of misrecognition, it will be necessary to devise feature parameters and their combinations that enhance the discrimination performance. In the use of VSD in this paper, the effect of variation of utterance speed is not taken into account. But in the future, if the proposed technique is to be used for continuous speech recognition, it will be necessary to increase its accuracy and robustness as a feature parameter by performing normalization based on the utterance speed and the like. In addition, it may be necessary to make concurrent use of local information such as the likelihood ratio of syllable units or durations and the like.

We also plan to develop speech dialog applications using the correctness probability obtained in this manner.
REFERENCES


AUTHORS (from left to right)

Norihide Kitaoka (member) graduated from the Department of Information Science, Kyoto University, in 1992, completed the M.E. program in 1994, and joined Denso Corp. He was engaged in research on speech recognition at Denso Basic Research Laboratories. From 1997 to 2000 he was enrolled in the second half of the doctoral program at Toyohashi University of Technology, and is now a lecturer in the Department of Information and Computer Sciences. He holds a D.Eng. degree, and is a member of the Acoustical Society of Japan, IEICE, and the Information Processing Society of Japan.

Ichiro Akahori completed the M.E. program at Okayama University in 1983 and joined Denso Corp. He is engaged in research on speech recognition systems. He is a member of the Society of Instrument and Control Engineers, the Information Processing Society of Japan, and the Acoustical Society of Japan.
Seiichi Nakagawa (member) completed the doctoral program at Kyoto University in 1976. He has been a professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, since 1990. He was a visiting researcher at Carnegie-Mellon University from 1985 to 1986. He is engaged in research on speech information processing, natural language processing, and artificial intelligence. He holds a D.Eng. degree. He received 1997 and 2001 Paper Awards from IEICE and the 1998 Paper Award from IETE. He is the author of *Speech Recognition by Probabilistic Models* (IEICE), *Fundamentals and Applications of Information Theory* (Kindai Kagakusha), and *Pattern Information Processing* (Maruzen), and a coauthor of *Speech, Auditory Perception, and Neural Network Models* (Ohm Press).