Efficient Beam Width Control to Suppress Excessive Speech Recognition Computation Time Based on Prior Score Range Normalization

KOBASHIKAWA Satoshi¹, HORI Takaaki², YAMAGUCHI Yoshikazu¹, ASAMI Taichi¹, MASATAKI Hirokazu¹ and TAKAHASHI Satoshi¹

¹NTT Cyber Space Laboratories, NTT Corporation, Japan
²NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract
This paper proposes a method that efficiently controls the beam width and so yields computation times that permit the practical automatic transcription of massive volumes of speech data. In particular, we focus on the fact that a lot of time is wasted in attempting to recognize poor quality speech data which will yield erroneous transcripts and thus provide no useful data for subsequent text processing. To stabilize the recognition time regardless of speech quality, our proposal controls the score beam width efficiently based on overall score spread against each target speech data on the premise of stored speech; it formulates the prior score range within the beam width and maximizes computation efficiency by normalizing the range associated with the survival rate of hypotheses. The technique proposed herein can rapidly estimate the range by using just monophones prior to speech recognition decoding. Experiments with several SNRs and real call-center speech sets confirm a reduction in computation time while matching the accuracy of existing techniques.

Index Terms: speech recognition, decoding parameter optimization, beam search

1. Introduction
Massive volumes of speech data are stored on a daily basis; the typical call center will store several tens of thousands of calls per day. Speech recognition technologies can transcribe the speech automatically; the items become searchable via the transcripts [1]. Several studies analyze the customer needs by applying text mining to stored conversational spoken documents [2][3]. However, the computation time needed to recognize a lot of conversational speech data can be excessive. In particular, poor quality speech data require a long time to recognize since they produce hypotheses with no significant score difference and thus degrade the pruning efficiency in beam search. Due to the erroneous transcripts created by their poor quality, they are of no use in subsequent spoken document processing and should be removed by applying confidence measures [4][5]; otherwise, recognition time is wasted in producing transcripts that are not useful.

To minimize the time taken for speech recognition processing, several conventional techniques have been proposed that optimize the decoding parameters [6] under a speed constraint [7]. Several adaptive pruning techniques have been proposed [8][9] to tackle the speech variability problem [10]; they adapt the parameters during decoding. However, both techniques use development data to determine the parameters. Thus, they fail to fully optimize the parameters if the target speech quality is substantially different from the development data.

This paper aims to reduce the computation time for recognizing speech data that will yield erroneous speech recognition results. To this end, histogram pruning [11] is a common approach. Since it is not as effective as score beam pruning, pruning criteria are often used in combination [12]. Histogram pruning is performed frame-wise based on instantaneous score distribution, so beam width should be wide to keep recognition accuracy adequate. On the premise that we are processing stored speech as in a call center, we target each speech data (i.e. phone call) directly instead of depending on development data. Our proposal controls the score beam width for each data (call) based on the overall (i.e. not instantaneous) score spread which is estimated just before speech recognition. The proposed technique formulates the score range within the beam width, and suppresses the speech recognition time by maintaining the range. The score range is rapidly estimated by using only those Gaussians that belong to monophones.

We evaluate the efficiency of the proposed technique in spontaneous speech recognition tasks with several speech data qualities at various SNRs: 0 dB to ∞ dB. Experiments show that our technique satisfies the speed constraint regardless of the quality of the target speech, and matches the accuracy achievable with ideally-optimized parameters. In addition, the proposed technique reduces the computation time needed to recognize the speech data recorded in an actual call center with effectively no drop in accuracy.

The rest of this paper is organized as follows; the proposed technique is described in Section 2. Section 3 introduces the experiments conducted to confirm the effectiveness of the proposed technique. Our conclusion is drawn in Section 4.

2. Proposed approach
Score beam pruning is the most well-known method to control computation time. It retains only those hypotheses whose score is close (within the score beam width) to the best state hypothesis [13]. Since the likelihood score distribution of hypotheses varies according to the speech quality, the recognition computation time fluctuates with speech quality. The computation time is stable if the survival rate of hypotheses is constant. The proposed approach estimates the prior score range for each target speech data that will yield erroneous speech recognition by using a limited number of Gaussians in an acoustic model. On the assumption that the prior score spread is proportional to the latter score spread in the subsequent speech recognition, we stabilize the recognition computation time by reducing the score beam width so as to keep the score range within the beam width.
2.1. Framework of proposed approach

The framework of the proposed system is shown in Fig. 1. The conventional system uses previously-fixed decoding parameters (e.g. $B_{\text{base}}$) optimized by using development data. In contrast, the proposed approach uses the input target speech to estimate the prior score range within the beam width, and then performs speech recognition by using the controlled score beam width, $B_{\text{prop}}$. To estimate the prior score range rapidly, it uses only monophones as in [13], and calculates the average of the monophones’ score spread as the prior score spread from the likelihood score (i.e. $S_{\text{best}}$). The score range $S$ is given by the following equation, where $S_{\text{base}}$ is the difference score from the best likelihood score ($\approx$ probability) i.e. score range $S$: the size of the area within the beam width in Fig. 2. The range of likelihood scores is spread due to the variance of the Gaussian distribution. By normalizing the score range $S$, we can keep the hypothesis survival rate constant, and therefore hold the computation time steady.

To calculate score range $S$, we swap the vertical and horizontal axes as in Fig. 3, and the range is found by integrating the inverse function $g(\cdot) = f^{-1}(\cdot)$ as follows;

$$S = 2 \int_{B_S}^{b_S} f^{-1}(\tilde{y})d\tilde{y} = 2 \int_{0}^{b_S} g(\tilde{y})d\tilde{y} = 2G(\tilde{y})|_{0}^{b_S} = 2G(B_S) = \text{const.} \quad (1)$$

where $\tilde{y}$ is the difference score from the best likelihood score (i.e. $\tilde{y} = y_{\text{best}} - y$) and $G(\cdot)$ is the integration function of $g(\cdot)$.

The likelihood score based on Gaussian distribution $\mathcal{N}(x; \mu, \sigma)$ is expressed by the following equation, where $-\log \sqrt{2\pi\sigma}$ corresponds to vertex $y_{\text{best}}$.

$$y = \log \left( \frac{1}{\sqrt{2\pi\sigma}} \exp \left( \frac{(x - \mu)^2}{2\sigma^2} \right) \right)$$

$$= -\log \sqrt{2\pi\sigma} - \frac{(x - \mu)^2}{2\sigma^2} = \sqrt{2\pi\sigma} y_{\text{best}} - \frac{(x - \mu)^2}{2\sigma^2} \quad (2)$$

Thus, difference score $\tilde{y}$ corresponds to the second term in Eq. (2) and can be converted as per the following equation where $\tilde{x} = x - \mu$ and $\alpha = \frac{1}{2\sigma^2}$.

$$\tilde{y} = y_{\text{best}} - y = \frac{(x - \mu)^2}{2\sigma^2} = \alpha \tilde{x}^2 \quad (3)$$

Then, by converting Eq. (3) to $\tilde{x} = \frac{1}{\alpha} y \tilde{y}$, the inverse function $g(\tilde{y})$ becomes

$$g(\tilde{y}) = \frac{1}{\alpha} \tilde{y}^\frac{1}{2}. \quad (4)$$

From Eq. (4), integral function $G(\tilde{y})$ is given by the following equation.

$$G(\tilde{y}) = \frac{2}{3\alpha^{\frac{3}{2}}} \tilde{y}^{\frac{3}{2}} \quad (5)$$

By substituting formula Eq. (5) into Eq. (1), we obtain the following.

$$S = 2G(B_S) = \frac{4}{3\alpha^{\frac{3}{2}}} B_S^3 = \text{const.} \quad (6)$$

Therefore, score range $S$ depends on beam width $B_S$ and coefficient $\alpha$ which is associated with variance $\sigma^2$ of the score distribution.

The score distribution changes with speech quality. In the case of clear (noisy) speech, the difference score from the best likelihood score becomes large (small) and so the score variance becomes narrow (wide) as shown in Fig. 4.
As score range \( S \) is constant as shown by Eq. (6), the beam width relation between target \( B_{\text{target}} \) and base \( B_{\text{base}} \) is shown by using \( \alpha_{\text{target}} \) and \( \alpha_{\text{base}} \), as per the following equation. In Fig. 4, the base beam width is fixed by using clear speech; the target speech is noisy.

\[
S = \frac{4}{3} \alpha_{\text{target}} B_{\text{target}}^2 = \frac{4}{3} \alpha_{\text{base}} B_{\text{base}}^2 \\
(7)
\]

The target beam width \( B_{\text{target}} \) is calculated from the base beam width \( B_{\text{base}} \) as follows;

\[
B_{\text{target}} = \left( \frac{\alpha_{\text{target}}}{\alpha_{\text{base}}} \right)^{1/2} B_{\text{base}} \\
(8)
\]

A review of Eq. (3) shows that there is a proportional relationship between \( \hat{y} \) and \( \alpha \); \( \hat{y} \propto \alpha \). \( \hat{y} \) means the difference from the best likelihood score, i.e. score spread. The score spread is considered as the difference between best and worst likelihood scores among hypotheses. To reduce the computation time, we calculate the score spread by using the average monophone’s score spread \( \bar{\gamma} \) instead of that of triphones.

\[
\frac{\alpha_{\text{target}}}{\alpha_{\text{base}}} = \frac{\bar{s}_{\text{target}}}{\bar{s}_{\text{base}}} = \frac{\gamma_{\text{mono}}}{\gamma_{\text{mono}}} \\
(9)
\]

Thus, the target beam width can be calculated from the ratio of the average monophone’s score spread to normalize the score range within beam width as follows;

\[
B_{\text{target}} = \left( \frac{\bar{s}_{\text{target}}}{\bar{s}_{\text{base}}} \right)^{1/2} B_{\text{base}} \\
(10)
\]

Our proposed approach is more efficient than histogram pruning [11]. Histogram pruning restricts the surviving number of hypotheses based on the histogram-based score distribution. However, it prunes based on instantaneous score histograms frame by frame, so a larger beam width is required to maintain equivalent accuracy. The surviving number of hypotheses changes frame by frame, and thus histogram pruning can only work at the frames wherein the surviving number exceeds the beam width after constructing the score histogram. Instead, the proposed score beam pruning is more efficient since it works immediately if the hypothesis’s is not close to the best hypothesis. Furthermore, our proposed technique can stably control the score beam width by using the overall score spread from all frames of the target speech data.

### 3. Experiments

#### 3.1. Experimental settings

Table 1 shows the evaluation task, Table 2 shows the speech analysis conditions, and Table 3 shows the acoustic model parameters used in the experiments.

Using the metrics of the average recognition rate and computation time, we investigate the impact of the beam width optimization proposal on speech data with several SNR values, from 0 dB to \( \infty \) dB with white noise. The effect of the proposed beam width optimization is confirmed in a three-way comparison; “conventional”: previously-fixed beam width optimized by using development data, “ideal”: beam width is optimized in a preliminary step by using target SNR speech data with maximum recognition rate under the speed constraint as per [7], it simulates the ideal condition as the development and target data are the same, and “proposed”: proposed beam width optimization. Here, the speed constraint is to keep the computation time under the case of \( \infty \) dB. All optimized techniques use histogram pruning [11] together with number beam width optimized by using development data.

We also evaluate the effectiveness of the “proposed” technique compared to the above-mentioned “conventional” one by using speech recorded in an actual call center as Table 4. The remaining conditions are the same as in the above investigation.

#### 3.2. Experimental results and discussions

The beam optimization performance recorded is shown in Table 5 and Fig. 5. Table 5 shows the average recognition rate [%] of speech with the target SNR. The horizontal axis in Fig. 5 is SNR as speech quality, and the vertical axis is the average computation time normalized by that of the conventional method. The performance on speech recorded in an actual call center is shown in Table 6; the average recognition rate/accuracy [%] and computation time [x RT(Real Time)].
As shown in Table 5, there is no significant difference in speech recognition rate. The effect of our beam optimization proposal is shown in Fig. 5. The “conventional” computation time depends on SNR, and the computation time exceeds that of ∞ dB at several SNR conditions; the increase is 50% at 20 dB. In contrast, the “proposed” computation time remains under that of ∞ dB at all SNR conditions. Since several regions are buried in noise and are recognized as non-speech periods (pause) at low SNR (< 15 dB), both methods reduce the computation time compared to the 20dB condition. However, even at low SNR, “proposed” provides a significant reduction in computation time with no significant degradation in recognition rate.

Table 6 shows the efficiency of the speech recorded in an actual call center. The proposed technique reduces the computation time by 21.4% with equivalent accuracy. This means the proposed technique could reduce excessive computer resource consumption significantly.

4. Conclusion
This paper proposed an efficient score beam width optimization technique that is performed just before speech recognition. It formulates the score range within the score beam width, and keeps the range constant by controlling the beam width with score spread so as to maintain the decoding computation time. Because the score spread is calculated by using the limited number of GMMs belonging to monophones, it offers high speed. Recognition experiments using spontaneous speech show that the proposed technique maintains the decoding speed regardless of speech quality while offering equivalent recognition accuracy. Furthermore, the proposed technique recognizes the speech data recorded in an actual call center while reducing the computation time significantly with effectively no drop in accuracy.

5. References