Mutual Information Based Dynamic Integration of Multiple Feature Streams for Robust Real-Time LVCSR

Shoei SATO†‡†(a), Member, Akio KOBAYASHI†, Kazuo ONOE†, Nonmembers, Shinichi HOMMA†, Toru IMAI†, Tohru TAKAGI†, and Tetsunori KOBAYASHI††, Members

SUMMARY We present a novel method of integrating the likelihoods of multiple feature streams, representing different acoustic aspects, for robust speech recognition. The integration algorithm dynamically calculates a frame-wise stream weight so that a higher weight is given to a stream that is robust to a variety of noisy environments or speaking styles. Such a robust stream is expected to show discriminative ability. A conventional method proposed for the recognition of spoken digits calculates the weights from the entropy of the whole set of HMM states. This paper extends the dynamic weighting to a real-time large-vocabulary continuous speech recognition (LVCSR) system. The proposed weight is calculated in real-time from mutual information between an input stream and active HMM states in a search space without an additional likelihood calculation. Furthermore, the mutual information takes the width of the search space into account by calculating the marginal entropy from the number of active states. In this paper, we integrate three features that are extracted through auditory filters by taking into account the human auditory system’s ability to extract amplitude and frequency modulations. Due to this, features representing energy, amplitude drift, and resonant frequency drifts, are integrated. These features are expected to provide complementary clues for speech recognition. Speech recognition experiments on field reports and spontaneous commentary from Japanese broadcast news showed that the proposed method reduced error words by 9.2% in field reports and 4.7% in spontaneous commentaries relative to the best result obtained from a single stream.

key words: speech recognition, stream integration, entropy, mutual information, active hypotheses

1. Introduction

Recent studies on automatic speech recognition (ASR) have expanded the practical uses of this technology. NHK (Nippon Hoso Kyokai or Japan Broadcasting Corporation) developed a real-time large-vocabulary continuous speech recognition (LVCSR) system for a closed-captioning service [1]. The system, followed by human correction, can generate captions for news programs and sports commentaries in real-time. The system has definitely been a success with read speech, but it is not accurate enough under noisy and spontaneous speech conditions for practical application. A variety of algorithms have been proposed to improve the recognition performance under these conditions. This paper presents a method for increasing the acoustic robustness for these varieties of speech.

It is known that our brain shrewdly integrates various cues, such as amplitude and frequency modulation, so that we can recognize unclear speech without difficulty [2]. Conventional speech recognition systems, however, are based on joint probabilities of spectrum-oriented features. In the literature, many attempts at integrating features have been proposed to capture various properties of speech. These attempts are classified into three kinds of integration.

The first group analytically integrates features into a single stream. This method obtains a discriminative set of features by using a projection matrix obtained by statistical analysis, such as principal components analysis (PCA), linear discriminant analysis (LDA) [3], heteroscedastic discriminant analysis (HLDA) [4], or maximum likelihood linear transformation (MLLT) [5]. A combined projection of HLDA and MLLT improves recognition accuracy [6].

The second group clusters the features into streams, and integrates the likelihoods of feature streams statically. The acoustic score of an input frame is a weighted sum of the log-likelihoods calculated from individual feature streams. Stream weights are defined in advance and can be optimized based on criteria, such as maximum likelihood (ML) [7], maximum mutual information (MMI) [8], minimum classification error (MCE) [9], or maximum entropy (ME) [10]. Stream weights optimized globally by using the ME criterion were reported to show the best performance among them [10].

The last group improves the likelihood integration so that the method optimizes the stream weights frame by frame. Dynamic stream weights, calculated from the entropy of the search space at a given input frame, were proposed for recognizing spoken digits [11], [12]. Selective stream weights and proportional weights to the inverse entropy were proposed in [11]. They calculate the entropy of each stream from posterior probabilities of the whole set of HMM states at every input frame. In the recognition of spoken digits, it is feasible to calculate observation probabilities of the whole set of HMM states because there is a constant search space of several hundred states. On the other hand, a real-time LVCSR system would have to deal with an infeasible amount of likelihood calculations to obtain the entropy from the whole set of HMM states.

The proposed method extends the dynamic weights to real-time LVCSR. The method estimates the entropy from probabilities of active HMM states at each input frame, so

Manuscript revised October 9, 2007.
†The authors are with NHK (Japan Broadcasting Corporation) Science and Technical Research Laboratories, Tokyo, 157–8510 Japan.
‡†The authors are with Waseda University, Tokyo, 169–8555 Japan.
a) E-mail: satou.s-gu@nhk.or.jp
DOI: 10.1093/ietisy/e91-d.3.815

Copyright © 2008 The Institute of Electronics, Information and Communication Engineers
that the dynamic weights can be obtained without any additional calculation of state probabilities. Furthermore, the dynamic marginal entropy given by a number of active states is taken into account. The method, therefore, obtains the weights from mutual information defined as the difference between the marginal entropy and the entropy given each feature stream. It gives a higher weight to a stream that is robust to a variety of noisy environments or speaking styles according to input speech characteristics.

In this study, we integrate features extracting amplitude modulation (AM) and frequency modulation (FM) of speech resonance, motivated by physiological evidence that the human auditory system extracts such modulation information [13]. Although these modulation features are not commonly used for speech recognition, several methods that demodulate AM and FM signals have been proposed in the literature. The Teager energy cepstrum coefficients (TECC) [14] are DCT coefficients of resonant frequencies, i.e., the center frequency of the formants, weighted by its energy. The resonant frequency is extracted from the output signal through band pass filters by using the energy separation algorithm derived from the Teager energy operation [15]. The comprehensive modulation spectra (CMS) [16] represents similar features to the TECC, but it extracts resonant frequencies by using the time derivatives of the phases obtained through orthogonal band-pass filters. The frequency modulation percentage (FMP) [17] was proposed to capture fluctuations in speech resonances. The feature is the ratio of the second moment over the first moment of instantaneous frequency extracted by the energy separation algorithm (ESA) [15]. Amplitude modulation also can be captured in a similar manner. Psychophysical experiments showed that these modulation features provide auxiliary cues to conventional spectrum-oriented features [18]. The feature streams treated in this study are based on a similar extraction algorithm to the FMP feature extraction, but they are calculated as the ratios of amplitude- or resonant-drift to the amplitude or frequency of the resonant.

The rest of this paper is organized as follows. Section 2 presents an outline of the proposed method and details of its stream weighting, and Sect. 3 describes the features we used. Experimental results on broadcast news are given in Sect. 4. Section 5 discusses the issues related to the features and implementation.

2. Integration of Multiple Streams

Figure 1 shows the outline of the method. The proposed LVCSR decoder integrates various feature streams, such as energy, AM, and FM, as described in Sect. 3. The frame-wise entropy of a stream is calculated from the likelihoods of active states in hypotheses generated in the search process. The number of active states also yields the entropy before observing an input feature, and the mutual information of an input stream is calculated as the entropy reduction caused by an observed feature. The method dynamically yields a log-linear stream weight proportional to the mutual information in order to give a higher weight to a discriminating stream or a stream giving more information. The likelihoods of the multiple features are integrated into an acoustic score in a log-linear domain. The Viterbi search is executed while changing the stream weights frame by frame.

2.1 Stream Weights Based on Mutual Information

As shown in Fig. 1, the proposed method gives a higher weight to a stream \( k \in \{1 \cdots K\} \) with larger mutual information, given an observation vector \( x_k(n) \), where \( n \) is frame index, \( k \) is stream index, and \( K \) is number of streams. The stream given a higher weight is supposed to be less affected than the others by background noise or spontaneous speech. The mutual information \( I_k(n) \) of the search space given \( x_k(n) \) is defined as follows,

\[
I_k(n) = H^0(n) - H_k(n),
\]

where \( H^0(n) \) is the marginal entropy of search space \( \Lambda(n) \), i.e., the entropy before observing a feature vector, and \( \Lambda(n) \) is a set of active HMM states \( \lambda \in \Lambda(n) \) at frame \( n \). In this paper, \( H_k(n) \), which is the entropy of \( \Lambda(n) \) given an observation vector \( x_k(n) \), is used instead of the conditional entropy so as to capture the information yielded by \( x_k(n) \). Therefore, the mutual information \( I_k(n) \) defined above can be calculated as the entropy reduction of the search space \( \Lambda(n) \) caused by observing a feature vector \( x_k(n) \).

The proposed method, in another sense, maps the entropy \( H_k(n) \) to a non-negative value of stream reliability by using a linear monotonic decreasing function. In this sense, the marginal entropy \( H^0(n) \) is utilized as an offset of the linear transformation so that the reliability will be non-negative. The offset can not be determined in advance in the method yielding the entropy from only active HMM states \( \Lambda(n) \), because the dynamic range or the maximum value of the entropy is not constant but dependent on the frame \( n \) or number of active states \( L(n) \). Therefore, we assume the prior probabilities of \( \lambda \in \Lambda(n) \) to be a uniform distribution.
to ensure that the reliability is a non-negative value. Consequently, $H^0(n)$ is only dependent on $L(n)$ and is calculated as

$$H^0(n) = \log L(n).$$

(2)

The forward Viterbi path scores in $\Lambda(n)$ may be used for the marginal entropy calculation. However, in such a case, the difference between the entropies calculated by the forward Viterbi scores before and after observing $x_k(n)$ does not provide a measure of the discriminative ability yielded by $x_k(n)$. Because $I_k(n)$ is the entropy reduction as defined in Eq. (1), it is not always positive for any observation $x_k(n)$. Hence, the forward Viterbi path scores cannot be used in our method since they could give a negative $I_k(n)$ even for a discriminatively important frame.

The entropy given a stream $x_k(n)$ can be written as follows,

$$H_k(n) = - \sum_{\lambda \in \Lambda(n)} P(\lambda|x_k(n)) \log P(\lambda|x_k(n)),$$

(3)

where we define a posterior probability $P(\lambda|x_k(n))$ given $x_k(n)$ as follows,

$$P(\lambda|x_k(n)) = \frac{\tilde{P}(x_k(n)|\lambda)}{\sum_{\lambda_j \in \Lambda(n)} \tilde{P}(x_k(n)|\lambda_j)},$$

(4)

assuming

$$P(\lambda_1) = \cdots = P(\lambda_{|\Lambda|}) = \frac{1}{L(n)}$$

(5)

and

$$P(x_k(n)) = \frac{1}{L(n)} \sum_{\lambda \in \Lambda(n)} \tilde{P}(x_k(n)|\lambda).$$

(6)

As integrated streams have potentially different discriminative abilities, their likelihoods should be balanced with each other so that the contribution of a stream which degrades the recognition performance will be underestimated. Therefore, $\tilde{P}(x_k(n)|\lambda)$ is defined to compensate the discriminative ability of stream $k$:

$$\tilde{P}(x_k(n)|\lambda) = P(x_k(n)|\lambda)^{w_k},$$

(7)

where $w_k$ is the static weight of the stream $k$ and $P(x_k(n)|\lambda)$ is the observation probability. As described in Sect. 1, these static weights can be optimized by ME or MCE using training data in advance. The dynamic stream weights described in the following are calculated from these compensated likelihoods.

The proposed integration multiplies each compensated likelihood in a log-linear domain by the dynamic weight $W_k(n)$, and obtains the acoustic score by summing the weighted log-likelihoods.

$$\log P(x(n)|\lambda) = \sum_{k=1}^{K} W_k(n) \log \tilde{P}(x_k(n)|\lambda).$$

(8)

Finally, the forward Viterbi search is executed using the following recursive equation,

$$a_i(n) = \max_j a_j(n-1) a_{ji} P(x(n)|\lambda_i),$$

(9)

where $a_i(n)$ is the accumulated Viterbi score of state $\lambda_i$ at frame $n$ on the Viterbi path and $a_{ji}$ is the transition probability from state $j$ to $i$.

The proposed method uses a stream reliability based on the entropy of the search space. In the literature, there are various ways of assessing the discriminative ability of the hypotheses. For example, the N-best log-likelihood difference [19], [20] utilizes the likelihood ratios between the best and the N-best hypotheses. In [19], $N$ was chosen to be 5. The N-best log-likelihood dispersion [19]-[21] captures additional N-best likelihood ratios which are not present in the N-best log-likelihood difference. These measures ensured stream reliability for audio-visual spoken digit recognition [19]. However, the discriminative ability calculated from a small number of N-best hypotheses tends to be underestimated in LVCSR utilizing a huge number of context-dependent HMMs such as triphones because of the multiple models trained for the identical center phoneme. Furthermore, these methods require an additional sorting procedure in order to obtain the N-best likelihoods of hypothesized HMM states, whereas the proposed method does not require such additional computation.

Below, we compare the proposed dynamic weights $W_k^{MI}(n)$ based on the mutual information $I_k(n)$ with two conventional dynamic weights, $W_k^{MinH}(n)$ and $W_k^{InvH}(n)$ based on the entropy $H_k(n)$.

2.2 Selective Weighting (MinH)

This integration uses selective weights [11]; therefore, the selected stream is the one giving the minimum entropy. The method obtains the stream weight $W_k^{MinH}(n)$ as follows:

$$W_k^{MinH}(n) = \begin{cases} 1.0 & \text{if } k = \arg \min_k H_k(n) \\ 0.0 & \text{otherwise}. \end{cases}$$

(10)

2.3 Inverse Entropy (InvH)

This integration introduces inverse entropy [11] in order to obtain a higher weight for a stream giving lower entropy. The stream weight $W_k^{InvH}(n)$ is calculated as follows:

$$W_k^{InvH}(n) = \frac{1}{\sum_{j=1}^{|\Lambda|} 1/H_j(n)}.$$
manner similar to the calculation of $W^{inh}(n)$,

$$W_k^{inh}(n) = \frac{I_k(n)}{\sum_{j=1}^{l} I_j(n)}. \quad (12)$$

3. Integrated Streams

To examine the proposed stream integration, we pick up three streams based on an auditory filter bank, as shown in Fig. 2. The orthogonal auditory filter bank is implemented with asymmetrical gamma-tone filters [22] whose impulse response is given by

$$g_{kc}(t) = A t^3 \exp(-2\pi \cdot 0.019 B(f_c) t) \cos(2\pi f_c t) \quad (13)$$

and

$$g_{jc}(t) = A t^3 \exp(-2\pi \cdot 0.019 B(f_c) t) \sin(2\pi f_c t), \quad (14)$$

where $t$ is the sample index, $A$ is the gain parameter, $c$ is the index of a filter, $f_c$ is the center frequency of filter $c$ and $B(\cdot)$ is a function giving the equivalent rectangular bandwidth. The filter yields a complex output $y_c(t)$; filter $g_{kc}(t)$ yields the real part $y_{kc}(t)$ and filter $g_{jc}(t)$ yields the imaginary part $y_{jc}$. These filters can be efficiently implemented with second order IIR filters [23]. The method compensates the length of the filter response by shifting the output $y_c(t)$ with

$$T_{shift}(c) = \arg \max_t \left( g_{kc}(t)^2 + g_{jc}(t)^2 \right) \quad (15)$$

samples.

The output signal $y_c(t)$ of the $c$-th filter is decomposed into its instantaneous amplitude $a_c(t)$ and instantaneous phase $\phi_c(t)$ as follows:

$$a_c^2(t) = |y_c(t)|^2 \quad (16)$$

$$\phi_c(t) = \angle y_c(t). \quad (17)$$

Three kinds of features are calculated from $a_c^2(t)$ and $\phi_c(t)$. The integrated streams are derived from these features through the logarithm and the DCT calculations. Each stream consists of 39 parameters: 12 cepstrum coefficients (without C0) with log energy and their first- and second-order regression coefficients ($\Delta + \Delta \Delta$). Descriptions of these features are given below.

3.1 Energy

The first feature captures the expectations of filter-bank energy at frame $n$. This stream is supposed to have the same aspect as mel-frequency cepstrum coefficients (MFCC). The $c$-th element of the feature $E_G(n)$ is calculated from the amplitude $a_c(t)$ as follows,

$$E_G(c) = E_n \left[ a_c^2(t) \right] = \sum_{t=1}^{T_w} a_c^2(T^* \cdot n + t) \quad (18)$$

where $T_w$ is frame width and $T^*$ is frame shift. The features calculated from Japanese speech pronouncing “m a ch i n i m a Q t a” are shown in Fig. 3.

3.2 Amplitude Modulation (AM)

The second feature captures the time derivative of the energy. We focus on the ratio of the energy differential, which is calculated as $\frac{a_c^2(t) - a_c^2(t-1)}{a_c^2(t)}$, and energy $a_c^2(t)$ so that Weber’s law of just noticeable difference (JND) is taken into account. An element of the feature $A_M(n)$ is calculated by summing the ratio over a frame as follows:

$$A_M(n) = \left| E_n \left[ \frac{\Delta a_c^2(t)}{a_c^2(t)} \right] \right| = \sum_{t=1}^{T_w} \frac{\Delta a_c^2(T^* \cdot n + t)}{a_c^2(T^* \cdot n + t)} \quad (19)$$

This feature captures the energy derivative calculated from just one frame. The modulation frequencies captured by this feature are different from the “Energy” regression coefficients, which are calculated from five adjacent frames as the smoothed derivative. This feature should yield an additional clue to the conventional spectrum-oriented feature, since it emphasizes the raising and falling edges of the input speech (Fig. 4).
3.3 Frequency Modulation (FM)

The third feature captures the time derivative of the resonant frequency. This frequency modulation feature has an extraction algorithm similar to FMP feature extraction, but it extracts the drift of the resonant frequency.

Assuming that the instantaneous frequency \( \dot{f}_c(t) \) is extracted by \( \dot{f}_c(t) = f_c(t) - f_c(t-1) \), the feature extracts the resonant drift \( \dot{f}_c(t) \) as \( \dot{f}_c(t) = f_c(t) - f_c(t-1) \). Moreover, Weber's ratio of the drift and resonant frequency \( f_c(t)/\dot{f}_c(t) \) is weighted by the energy \( f^2_c(t) \) in order to prevent the ratio from being influenced by non-speech resonants. An element of the feature \( FM_c(n) \) is the weighted sum of the ratios in a frame.

\[
FM_c(n) = \left[ \frac{1}{E_n} \sum_{T^n} \left( f^2_c(T^n + t) \frac{\dot{f}_c(T^n + t)}{\dot{f}_c(T^n + t)} \right) \right]
\]

(20)

As shown in Fig. 5, the resonant movement is emphasized in this feature. Although it may be difficult to obtain a good recognition result from these modulation features alone, these additional clues are expected to help the recognition by integrating with the conventional stream.

4. Experiments

4.1 Experimental Setup

Experiments were performed to examine the recognition performance of the proposed stream integration. The evaluation speech data were extracted from NHK's Japanese broadcast news. The 16 most difficult news topics were selected from the programs aired in July 2004. The data consisted of read speech in quiet studios, noisy field reports, and conversational commentaries uttered by male announcers and reporters; there were 643 segments comprising 8,391 words “all”. Subsets of “field” reports (395 segments comprising 4,905 words) and “spontaneous” commentaries (231 segments comprising 3,220 words) were also evaluated. There were 97 common segments in these subsets, and the complement set of the union of these subsets included 114 segments of read speech in a quiet studio.

Triphone HMMs with 16 mixtures and about 4 K clustered states were trained from about 100 hours of NHK's multi-conditioned news data uttered by male speakers. The announcers and some reporters in the evaluation set were included in this training set. All the features described in this paper were extracted from pre-emphasized speech digitized at 16kHz and 16 bits. The features were obtained from band-limited signals passed through 40 channels of gammatone filters with a frame of \( T_w = 25 \text{ ms} \) width and \( T_s = 10 \text{ ms} \) shift. Each stream element was normalized to have zero mean and unit variance on the training data.

The n-gram language models used in this experiment were word bigrams for the first pass in the continuous speech recognition and word trigrams for the second pass. The training texts were NHK’s Japanese news manuscripts consisting of 127M words extending back over 10 years. As more recent news had a higher probability of frequently appearing, the n-gram language models were trained with a higher weighting factor to the latest news in an n-gram count level [24]. There were 61K vocabulary words, and the trigram language model showed a perplexity of 21.9, with an out-of-vocabulary rate of 0.4% against the evaluation data.

4.2 Single Stream

Table 1 shows the word error rate (WER) of the recognition results of each individual feature stream. The WER discarding insertion errors, i.e., the WER calculated from only substitution and deletion errors, are shown in parentheses. The
reason for omitting insertion errors was the difficulty of the unbiased evaluation of many repetitions consisting of several words. It is difficult to define unbiased appropriate references for these repetitions, which contain disfluencies or lack of several phonemes, and the repetitions are manually deleted whether these words are recognized correctly or not in the practical use of the LVCSR for the closed-captioning service [1]. The correction cost of these repetitions is negligible because they are correctable without any manual keyboard input following the recognition [1]. The reference used in the experiments, therefore, discarded such repetitions, which were mainly contained in the “spontaneous” subset.

To compare the recognition results obtained from the streams yielding different dynamic ranges of acoustic score, the acoustic scaling factor was optimized to minimize total word errors of insertion, substitution, and deletion, while the other search parameters were fixed; beam width was 200, maximum HMM nodes being activated was 2,000, and grammar weight was 10. Table 2 shows the optimized acoustic scaling factors and frame-averaged number of active states (#active). The equality of the search width of each experiment is confirmed. Among the individual streams, the modulation features produced larger WERs than the fine result obtained from the “Energy” stream.

Tables 3 and 4 compare the recognition results of joint feature streams joined with three features. “EAF” shows a result of a 111-dimensional stream joined with three kinds of DCT coefficients and log-power (12 × 3 + 1) with their regressions (37 × 3). “HLDA” shows a result of a stream transformed by a HLDA-MLLT [6] matrix reducing 111 dimensions to a 57 or 39-dimensional stream and de-correlating each element of the stream. The matrix was trained with the same data as used in HMM training. The dimension of the transformed feature “HLDA(57)” was optimized with the Fibonacci search method [25]. The results show that the “EAF” degraded WERs compared with the best result obtained from the individual stream, i.e., “Energy” in Table 1. In addition, “HLDA” could not improve the WERs except under the condition of spontaneous speech.

### Table 2

<table>
<thead>
<tr>
<th>Feature stream</th>
<th>Acoustic scaling factor</th>
<th>#active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy [39]</td>
<td>0.8</td>
<td>2355</td>
</tr>
<tr>
<td>AM [39]</td>
<td>1.2</td>
<td>2243</td>
</tr>
<tr>
<td>FM [39]</td>
<td>1.3</td>
<td>2211</td>
</tr>
</tbody>
</table>

Dimension of the feature vector is shown in square brackets.

### Table 3

<table>
<thead>
<tr>
<th>Feature stream</th>
<th>WER [%]</th>
<th>all</th>
<th>field</th>
<th>spon.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAF [111]</td>
<td>8.9 (7.4)</td>
<td>9.0 (7.8)</td>
<td>14.8 (12.5)</td>
<td></td>
</tr>
<tr>
<td>HLDA [39]</td>
<td>8.2 (7.1)</td>
<td>8.8 (7.8)</td>
<td>12.3 (10.6)</td>
<td></td>
</tr>
<tr>
<td>HLDA [57]</td>
<td>7.5 (6.5)</td>
<td>7.8 (6.9)</td>
<td>12.3 (10.3)</td>
<td></td>
</tr>
</tbody>
</table>

Dimension of the feature vector is shown in square brackets.

### Table 4

<table>
<thead>
<tr>
<th>Feature stream</th>
<th>Acoustic scaling factor</th>
<th>#active</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAF [111]</td>
<td>0.4</td>
<td>2425</td>
</tr>
<tr>
<td>HLDA [39]</td>
<td>1.0</td>
<td>2142</td>
</tr>
<tr>
<td>HLDA [57]</td>
<td>0.9</td>
<td>2193</td>
</tr>
</tbody>
</table>

Dimension of the feature vector is shown in square brackets.

4.3 Integrated Likelihoods

All the acoustic models integrated in these experiments were composed so as to retain the identical structure of the state-tying and probability matrices of the state transitions. Each HMM was generated from the reference HMM of the 111-dimensional stream “EAF(111)”, to which three kinds of the features were joined. The parameters of the probability density function of each state were re-estimated by using each feature stream while keeping the state-tying structure and transition matrices. These identical structure and transition probabilities accomplished the Viterbi search on a single pre-compiled state network with varying the stream weight.

Tables 5 and 6 show the results of the stream integration. In this experiment, the static stream weight $W_k$ described in Eq. (7) was unit weight for all streams $k$. “Static” shows the result of static weighting, i.e., the integration did not use dynamic weights and $W_k(n)$ was 1.0 for all $k$ and $n$. It had higher word error rates than those of “Energy(39)”, “MinH”, “InvH”, and “MI” indicate the dynamic weighting described in Sect. 2.1. These results showed that the dynamic weights did not reduce the total WERs (“all”) compared with the result obtained by “Energy”. “MinH”, however, reduced the WER of the “field” condition. The dynamic weights also reduced the WERs discarding insertion errors under the “field” condition. These results were contrastive to those of “HLDA(57)”.

The selective method “MinH” yielded percentages of frames selected to use the energy, AM, and FM features of 70%, 16%, and 15%, respectively. This showed that the energy feature was the most reliable among them and consistent with the results shown in Table 1. In a similar manner, the average weights of each stream of energy, AM, and FM

<table>
<thead>
<tr>
<th>Integration</th>
<th>Acoustic scaling factor</th>
<th>#active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.7</td>
<td>2130</td>
</tr>
<tr>
<td>MinH</td>
<td>0.9</td>
<td>2243</td>
</tr>
<tr>
<td>InvH</td>
<td>1.0</td>
<td>2278</td>
</tr>
<tr>
<td>MI</td>
<td>1.1</td>
<td>2223</td>
</tr>
</tbody>
</table>
were calculated over the evaluation frames. “InvH” yielded average weight ratios of 0.45:0.28:0.27, and “MI” yielded 0.37:0.31:0.32. “MI” degraded the WERs more than “InvH” because of its narrower weight dynamics.

Tables 7 and 8 show the results of the dynamic stream integration in conjunction with the optimized static stream weights. The optimization of the static stream weights was based on the maximum entropy criterion [10] and yielded weights $w_k$ maximizing the posterior log-likelihood $\log P(\hat{\lambda}|x)$ given training feature vectors $x = \{x(1) \cdots x(N^{\text{train}})\}$, where $N^{\text{train}}$ is the number of training frames, and the corresponding state sequence $\hat{\lambda}$.

$$\log P(\hat{\lambda}|x) = \sum_{n=1}^{N^{\text{train}}} \sum_k w_k \log P(\hat{\lambda}(n)|x_k(n)) - \sum_{\lambda} \sum_k w_k \log P(\lambda|x_k(n))$$  \hspace{1cm} (21)

The state clusters $\hat{\lambda}$ of 42 Japanese phonemes were taken into account to make use of the optimization for the LVCSR task processing thousands of states. Namely, the posterior probability $P(\hat{\lambda}(n)|x_k(n))$ in (21) was replaced by

$$\tilde{P}(\hat{\lambda}(n)|x_k(n)) = \max_{\lambda(n)} P(\lambda|x_k(n)).$$  \hspace{1cm} (22)

Optimization on a 9-hour subset of training data yielded static weights of 0.51:0.28:0.21 for energy:AM:FM.

The tables also show results for the “IEWAT” method [11], i.e., inverse entropy weighting with average entropy at each frame level as threshold. It obtained the largest improvement in experiments on recognition of connected digits in telephone speech, which integrated the outputs of artificial neural networks trained for the full-combinations of multi-stream [11], [26]. In this weighting scheme, the average entropy of all the streams for a frame is calculated by

$$\bar{H}(n) = \frac{\sum_{k=1}^{K} H_k(n)}{K}. \hspace{1cm} (23)$$

The average entropy is used as a dynamic threshold for the frame. The likelihoods of the streams having an entropy greater than the threshold are weighted much less, whereas the likelihoods of the streams having an entropy lower than the threshold are weighted inversely proportionally to their respective entropies. The stream weight $W_k^{\text{IEWAT}}(n)$ is calculated as follows:

$$W_k^{\text{IEWAT}}(n) = \frac{1}{\sum_{k=1}^{K} 1/\bar{H}_k(n)}$$  \hspace{1cm} (24)

The selective method “MinH” yielded percentages of frames selected to use the energy, AM and FM features of 99.8%, 0.2%, and 0.1%, respectively, and “InvH”, “IEWAT”, and “MI” yielded ratios of average weights to the evaluation data of 0.49:0.26:0.24, 0.99:0.008:0.001, and 0.61:0.23:0.15. These results showed that the optimized static stream weights significantly reduced the WERs of “MI” compared with the results shown in Table 5. Dynamic integration improved the WERs in both “field” and “spontaneous” subsets, while the static integration of “HLDA” improved the WER only in the “spontaneous” subset and “Static” improved the WER only in the “field” subset.

Finally, the proposed dynamic weight “MI” based on mutual information reduced WER by 4.1% relative to “MinH” which showed the best result from among the conventional stream integration. The proposed method also obtained WER reductions of 5.4%, 9.2%, and 4.7% relative to the single stream of “Energy” for the evaluation sets of “all”, “field”, and “spontaneous”, respectively.

The WER reduction of the proposed integration (MI) from the WER obtained for the single stream (Energy) for all evaluation data (all) was significant in a paired difference t-test [27] at a significance level of 0.05, but the WER reduction of the conventional integration (MinH) was not significant. The difference in WERs between the proposed integration (MI 7.0% (5.9%)) and the conventional integration (MinH 7.3% (6.2%)) was not significant, but the difference in WERs discarding insertion errors showed significant improvement.

For the dynamic weight of $W_k^{\text{MI}}(n)$ calculated directly from the uncompensated likelihood $P(x_k(n)|\lambda)$, (i.e., $\forall k$, $w_k = 1.0$ was used in the calculation of $W_k^{\text{MI}}(n)$ and optimized $w_k$ were used in the Viterbi search) the WERs shown in Table 7 (“MI”) were degraded to 7.8%, 7.3%, and 13.7% for “all”, “field”, and “spon.”, respectively. This result suggests the importance of the compensated likelihood $\tilde{P}(x_k(n)|\lambda)$ in the estimation of the dynamic weight.

Tables 9 and 10 show the results of stream integration calculating the dynamic weights from all 3,972 states of the HMM set. The differences in WERs between the results shown in Table 7 and those of Table 9 were very small, although the proposed method required only 56% of the computational cost relative to the experiment calculating
Table 9: Recognition results of the stream weighting calculated by the whole set of HMM states.

<table>
<thead>
<tr>
<th>Integration</th>
<th>WER [%]</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinH.full</td>
<td>7.3 (6.2)</td>
<td>7.0 (6.1)</td>
</tr>
<tr>
<td>InvH.full</td>
<td>7.4 (6.2)</td>
<td>7.2 (6.3)</td>
</tr>
<tr>
<td>MI.full</td>
<td>7.9 (5.9)</td>
<td>6.8 (5.9)</td>
</tr>
</tbody>
</table>

WER discarding insertion errors is shown in parentheses.

Table 10: Acoustic scaling factor and average number of active states of stream weighting calculated by the whole set of HMM states.

<table>
<thead>
<tr>
<th>Integration</th>
<th>Acoustic scaling factor</th>
<th>#active</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinH.full</td>
<td>1.6</td>
<td>3972</td>
</tr>
<tr>
<td>InvH.full</td>
<td>2.5</td>
<td>3972</td>
</tr>
<tr>
<td>MI.full</td>
<td>2.3</td>
<td>3972</td>
</tr>
</tbody>
</table>

all 3,972-state probabilities.

5. Discussion

This section discusses the issues related to the features used in this experiment and the implementation of the proposed method. Because the AM and FM features were very weak compared with the energy feature, good results were not obtained from the conventional static integration of the joint feature “EAF”, the transformed feature “HLDA”, or the static stream weight “Static”. Though carefully designed features may improve the recognition performance of these integration schemes, the major advantage of the proposed method was the experimental fact that the dynamic weights improved WERs by integrating these weak streams, i.e., the AM stream and the FM stream. It is considered that the weak streams improved the WERs owing to their local robustness against particular kinds of environmental conditions or speaking styles. Therefore, a stream robust to a certain kind of noise for specific phonemes has a chance to improve the WER by the proposed method, even if the noise is unseen in the HMM training. Of course, the proposed method would perform better if it integrated more accurate features or features whose vector components are optimized to represent complementary speech aspects.

Figure 6 compares the proposed dynamic weights $W_{\text{MI}}(n)$ and the conventional dynamic weights $W_{\text{linH}}(n)$. The top graph shows the envelope of an input waveform. The second graph shows the number of active states in the log scale together with recognition results. The WER of the result shown in (E), which is the result for “Energy” and $W_{\text{linH}}(n)$, was 50%. The WER of the result shown in (C), which is the result of $W_{\text{MI}}(n)$, was 0%. The third and the bottom graphs show $W_{\text{MI}}(n)$ and $W_{\text{linH}}(n)$ for the same time period as the top graph.

$W_{\text{MI}}(n)$ maps the entropy $H_k(n)$ to a non-negative reliability by using a linear monotonic decreasing function, while $W_{\text{linH}}(n)$ maps it by using the inverse function. $W_{\text{linH}}(n)$ utilizes a more nonlinear function than $W_{\text{MI}}(n)$. The difference in $W_{\text{linH}}(n)$ between “AM” and “FM” was smaller than the difference in $W_{\text{MI}}(n)$. It is considered that the nonlinear characteristic of the inverse function did not capture the difference in the reliabilities of these streams, whereas the proposed method yielded better estimation of reliabilities of these streams by using the linear function and the static compensation weight $W_k$ than the conventional method.

The other issue is the computational cost when implementing the proposed method on a real-time system. As mentioned above, the proposed method, which takes active states into account, does not increase the number of states requiring a likelihood calculation. The number of likelihood calculations, however, increases with the number of streams; for example, the proposed integration under the experimental conditions of this study required three times more likelihood calculations than those of a single stream such as “Energy”. In this case, the pool of state-likelihoods, from which the method obtained the dynamic weights, can help the implementation of the proposed method. To obtain the entropy of the active states, the proposed method pools the likelihoods of all active states before updating Viterbi scores described in Eq. (9). This likelihood pool can be computed in parallel although the Viterbi updates can not be computed in parallel because of their merging process of the Viterbi scores. In this method, the process of the Viterbi update can be implemented to compute the scores by referring to the
pooled likelihoods after pooling. In addition, a recent multicore processor makes parallelization easy.

6. Conclusion

This paper proposed a new method of real-time stream integration utilizing mutual information given input feature vectors in the search space, i.e., active HMM states of hypotheses. The method calculates frame-wise dynamic stream weights from the mutual information and dynamically integrates the likelihoods of multiple feature streams. Motivated by physiological evidence, we integrated three streams extracting energy, AM, and FM.

Transcription experiments implementing this method for Japanese broadcast news were performed for the purpose of automatic captioning. The results showed that the proposed dynamic weighting reduced error words by 5.4% relative to the result of the single stream of "Energy" feature. The experiments also showed that a combination of the proposed integration with static stream weights optimized by ME reduced error words by 4.1% relative to the conventional dynamic integration based on the minimum entropy ("MinH").

Future work will involve experiments using more accurate features and optimization of the combination of feature vector components to complement speech aspects.

Acknowledgments

The coefficients of the gamma-tone filter were suggested by Prof. Toshio Irino with Wakayama University.

References

Shoei Sato received the B.E. degree and M.E. degree in 1993 from Tohoku University. He has been with the NHK Science and Technical Research Laboratories since 1995. He researched digital satellite broadcasting systems. He is currently engaged in automatic speech recognition research. Since April 2006, he has been a PHD candidate under the supervision of Prof. Tetsunori Kobayashi at Waseda University. He is a member of the Acoustic Society of Japan.

Akio Kobayashi received his B.E. degree in 1991 from Waseda University. He joined NHK the same year and has been with the Science and Technical Research Laboratories since 1996. He is currently engaged in speech recognition research. He is a member of the Acoustic Society of Japan (ASJ), the Information Processing Society of Japan, and the Association of Natural Language Processing of Japan.

Kazuo Onoe received the B.E. degree in electrical engineering in 1993 from Waseda University. He joined Japan Broadcasting Corporation (NHK) in 1993. Since 1997, he has been with the Science and Technical Research Laboratories where he has been engaged in speech recognition research.

Shinichi Homma received his B.E. degree in electronic and communication engineering from Waseda University, Tokyo, Japan, in 1992. He joined Japan Broadcasting Corporation (NHK) in 1992. Since 1998 he has been with the Science and Technical Research Laboratories, where he has been engaged in research on speech recognition.

Toru Imai received his B.E. degree in electrical engineering in 1987 and his Ph.D. degree in information and science in 1999 from Waseda University. He joined Japan Broadcasting Corporation (NHK) in 1987. Since 1996, he has been with the Science and Technical Research Laboratories. In 1996, he was a visiting scientist of BBN in Massachusetts. He is currently a senior research engineer of the Human Science Research Division at the Laboratories, where he is currently engaged in speech recognition research. He received the Awaya Prize from the Acoustical Society of Japan (ASJ) in 1998, the Excellent Paper Award from IEICE in 2002, and the Niwa Takayanagi Prize from the Institute of Image Information and Television Engineers (ITE). He is a member of IEEE, ASJ, ITE and the Information Processing Society of Japan.

Tohru Takagi received his B.E. and M.E. degrees in electrical engineering from the University of Electro-Communications, Tokyo, Japan. He also received the Dr. Eng. degree from Toyohashi University of Technology in 2006. He joined NHK (Japan Broadcasting Corporation) in 1981. Since 1984, he has been with the Science and Technical Research Laboratories, where he has been engaged in research on speech synthesis, voice conversion, speech perception, speech rate conversion, and development of a speech recognition-error correction system.

Tetsunori Kobayashi received the B.E., M.E., and Dr. Eng. degrees in electrical engineering from Waseda University, Tokyo, Japan, in 1980, 1982, 1985, respectively. He was a Lecturer (1985–1987) and an Associate Professor (1987–1991) at Hosei University, Tokyo, Japan. In 1991, he joined Waseda University as an Associate Professor. Since 1997, he has been a Professor there. He was a visiting scientist at the Spoken Language System Group, Laboratory for Computer Science, Massachusetts Institute of Technology from 1994 through 1995. His research interests include perceptual computing and intelligent robotics. He received the IEICE Paper Award in 2003. He is a member of the Information Processing Society of Japan and the Institute of Electrical and Electronics Engineers.