Short- and Long-term Dynamic Features for Robust Speech Recognition

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura

Tokyo Research Laboratory, IBM Japan, Ltd.
{fukuda1, ichikaw, nisimura}@jp.ibm.com

Abstract

The short-term temporal information in speech is widely used for automatic speech recognition (ASR) systems in the form of dynamic features. Long-term temporal information has also been focused on recently and is used to complement traditional short-term features (typically from 25 to 100 ms). There are several approaches to represent long-term temporal information in ASR systems. However, those systems use high-dimensional feature spaces to capture the long-term temporal information. This paper describes an attempt to incorporate long-term temporal information into a feature parameter set by combining conventional dynamic features extracted from both short- and long-term cepstrum sequences. The proposed method includes the temporal contexts of phonemes by using long-term features and the spectral variations within phonemes as short-term features. In an experiment on the realistic speech corpus CENSREC-2, the proposed method yielded higher performance than a standard cepstral mean normalization (CMN) approach also modifies the modulation spectrum by eliminating the DC component from the cepstral coefficients. A straightforward approach for incorporating more temporal context into the feature parameter set is to simply widen the delta window. In the next section, we discuss the effectiveness of dynamic features as a long-term temporal information representation, and Section 3 includes an outline of the proposed method, experimental results, and provides some discussion. Finally, Section 4 finishes with our conclusions.

1. Introduction

Many approaches have been investigated for robust automatic speech recognition (ASR) systems. Current ASR systems show high performance in low-noise environments with basic tasks such as digit tasks, Boolean tasks, or command tasks. This is mainly due to the powerful support of a stochastic classifier, or an HMM-based classifier, and large speech corpora. However, the HMM classifier can often fail to provide adequate performance in real-world environments such as in automobiles. The most straightforward approach for overcoming the problems is to design the HMM classifier using data recorded in the same situation where the recognizer will be deployed. This is a very simple and reliable method to design robust ASR systems, but there is still room for improvement.

Feature-based approaches such as dynamic features [1] and RASTA [2] have been proposed for robustness. The dynamic features are widely used for ASR systems and equivalent to filtering processes that emphasize the modulation spectrum around 10 Hz for the temporal trajectory of the spectral envelopes [3]. The RASTA filter passes the modulation spectrum around 10 Hz for the temporal trajectory equivalent to filtering processes that emphasize the dynamic features are widely used for ASR systems and still room for improvement.

2.1. Dynamic feature extraction

The dynamic features are generally calculated with a short window length consisting of several consecutive frames and are combined with static cepstral coefficients. In the following, the first-order and second-order derivatives of the cepstral sequence appear as $\Delta$ and $\Delta^2$ cepstrum, respectively. The $\Delta$ cepstrum $d_1(t)$ at time $t$ is estimated as

$$d_1(t) = \sum_{\theta=0}^{\Theta-1} \left( \theta \cdot (c_{t,\theta} - c_{t-1,\theta}) \right) / 2 \sum_{\theta=0}^{\Theta-1} \theta^2,$$

where $c_t$ is the cepstral coefficient at time $t$. In Equation (1), consecutive frames of $2\Theta+1$ are used to extract the $\Delta$ cepstrum. The window lengths consisting of $2\Theta+1$ frames for $\Delta$ cepstrum and $\Delta\Delta$ cepstrum are often called a delta and an acceleration window, respectively. In a standard ASR system, the value of $\Theta$ is set to from two to four based on the frame size, the frame rate, and other parameters. In this paper, the same equation is used for the $\Delta\Delta$ cepstrum to obtain the $\Delta\Delta$ cepstrum coefficients. A straightforward approach for incorporating more temporal context into the feature parameters is to simply widen the delta window. In the next section, we discuss the effectiveness of dynamic features when long deltas or large windows are used. The $\Delta$ cepstrum

information spanning more than 500 ms by using multi-layered perceptrons (MLPs) [5]. Their approach uses temporal patterns consisting of consecutive frames of log critical band energies and feeds these patterns into the MLPs. Similar approaches were investigated in [6, 7]. Using psychological tests, Poeppel showed that human beings use two types of temporal information extracted from both short (20 to 40ms) and long (150 to 250ms) temporal windows to understand spoken language [8]. These prior studies suggest that long-term temporal information can support accurate ASR systems. However, generally it is necessary to deal with a large-dimensional feature vector to exploit the long-term temporal information, and thus these systems tend to be complicated.

In this paper, we attempt to incorporate long-term temporal information into the feature parameter set in a simple way. In our proposed method, the feature parameter set is composed of a combination of two types of $\Delta$ cepstrum calculated from different delta window lengths. The first-order dynamic features, or $\Delta$ cepstrum sequences, are robust to the additive noise interference [9]. For greater noise-robustness, then the $\Delta$ cepstrum obtained from enlarged delta window lengths are used.

This paper is organized as follows. Section 2 shows the effectiveness of the dynamic features as a long-term temporal information representation, and Section 3 includes an outline of the proposed method, experimental results, and provides some discussion. Finally, Section 4 finishes with our conclusions.
parameters obtained from short and long delta window lengths are called the short-term and long-term dynamic features, respectively.

2.2. Experimental setup

A Japanese connected digit corpus called CENSREC-2 from the IPSJ-SIG SLP noisy speech recognition evaluation working group in Japan was used in our experiments [10]. The recognition task of the CENSREC-2 database involves continuous digit strings recorded in automotive environments under conditions of idling, low-speed driving on city streets, and high-speed driving on expressways. The CENSREC-2 database has four types of data sets reflecting the positions of the microphones and the noise conditions. In our experiment, speech data collected at the same microphone position in the same noise conditions for training and testing was used. This microphone was attached under the ceiling of the car above the driver’s head. The car environment includes air-conditioner noises, an audio CD player, and open windows. There are a total of 7,195 sentences uttered by 33 male and 40 female speakers in the training set, and 2,964 sentences uttered by 19 male and 12 female speakers in the test set. The evaluation category was zero for CENSREC-2, which means no change at the back-end system.

The input speech was sampled at 16 kHz and each 25 ms speech segment was pre-emphasized by the filter \( H(z) = 1 - 0.97z^{-1} \) every 10 ms. A 512-point FFT for each speech segment was used after the speech segment was Hamming-windowed. The resultant FFT power spectrum was integrated into the outputs of band-pass filters with 24 channel mel-scaled center frequencies. Then the outputs of the band-pass filters were converted into 12 static cepstrums (the MFCC) by using DCT. Finally, \( \Delta \)cepstrum and \( \Delta \Delta \)cepstrum were extracted with Equation (1). The static MFCC was processed with CMN for every utterance in both the training and testing data. A total of 11 digit-HMMs with 18 states were used together with a silence model with 5 states and a pause model with 3 states.

2.3. Experimental results

The experimental results are shown in Fig. 1. In the figure, MFCC25 shows the feature parameter set with 25 dimensions combined with MFCC, \( \Delta \)cepstrum, and Apower. MFCC38 represents the feature parameter set with 38 dimensions combined with MFCC, \( \Delta \)cepstrum, \( \Delta \Delta \)cepstrum, Apower, and \( \Delta \)Apower. The dotted line indicates the result with only 12 static MFCCs. The acoustic models are designed for each window length. As shown in the figure, MFCC25 significantly improved the performance in comparison with the static MFCCs only. MFCC25 had the best performance when the delta window length was seven \((\Theta = 3)\), and showed a word accuracy of 88.02\%. A spectrum segment consisting of seven frames roughly corresponds to the duration of a phoneme and is used in the standard ASR system. However, MFCC25 including \( \Delta \)cepstrum didn’t improve the performance as the delta window length was increased.

Next we considered the MFCC38 including \( \Delta \Delta \)cepstrum. In this experiment, \( \Delta \Delta \)cepstrum was calculated from the \( \Delta \)cepstrum sequence extracted with \( \Theta = 3 \), which showed the greatest accuracy in MFCC25. As can be seen in the figure, MFCC38 had slightly lower performance than MFCC25 because of noises. But the ASR performance can be improved in noisy environments by optimally weighting the likelihood of static MFCC, \( \Delta \)cepstrum, and \( \Delta \Delta \)cepstrum in the HMM decoding process [9]. In terms of long-term acceleration features, MFCC38 also didn’t improve the performance when the long acceleration window was used. This result means that using long-term dynamic features alone is not effective for noise-robust ASR.

2.4. Long-term dynamic feature extraction as a filtering process

Here we discuss the long-term dynamic feature extraction in terms of a filtering process of a modulation spectrum. Previous studies showed that the accurate acoustic models can be built by properly filtering the modulation spectrum \([11, 12, 13, 14]\). Durliman et al. showed that the high-pass filtering above 4 Hz and the low-pass filtering below 16 Hz for the modulation spectrum did not reduce the speech intelligibility \([11, 12]\). Kanedera et al. found that most of the useful linguistic information of speech lies in the modulation frequencies ranging from 1 to 16 Hz and especially from 2 to 10 Hz \([14]\). In agreement with these findings, a linear regression calculation used in the dynamic feature extraction is also regarded as a filtering process that emphasizes the important modulation spectrum for ASR. Fig. 2 shows the frequency responses of linear regression filtering with 7 frames \((\Theta = 3)\) which extracts short-term dynamic features and with 17 frames \((\Theta = 8)\) which extracts long-term dynamic
features. In the figure, the attenuation amplitude of the filtering process is adjusted in order to compare the two different window lengths. This shows the modulation frequencies around 10 Hz are emphasized by short-term linear regression filtering with 7 frames while those around 2 Hz are emphasized by long-term linear regression filtering with 17 frames. The short-term dynamic features enhance the linguistic information with preserving speech intelligibility. In contrast, the long-term dynamic features focus on slowly changing spectral variations by transforming the modulation spectrum. In [8], Poeppel suggested that both short-term variations and long-period spectral information were important for listening to spoken words. In a similar manner, the combination of short-term and long-term dynamic features has the potential to improve the ASR performance.

3. Combination of dynamic features

3.1. Overview of proposed method

In Section 2, we discussed the effectiveness of dynamic features in terms of the representation of long-term temporal information. From the experimental results, we found that ∆cepstrum had no effect on performance when the delta window length was increased. In this section, we investigate a combination of two ∆cepstrum parameters extracted from both the short and long delta window lengths. The ∆cepstrum is eliminated in the proposed feature parameter set, and instead, the long-term ∆cepstrum is applied to the feature parameter. Hence the total number of dimensions is the same as the standard feature parameter set with 38 dimensions. The long-term ∆power is also used instead of ∆power. The speech database and the outline of the experiment are the same as in Section 2.2.

3.2. Experimental results

Fig. 3 illustrates these experimental results. In the experiment, ∆cepstrum calculated with θ=3, which showed the best performance in MFCC25, was used for the short-term dynamic features of the proposed method. In the figure, the delta window length was changed for the other ∆cepstrum, or long-term dynamic features. Here, MFCC25 and MFCC38 are the same configurations as shown in Fig. 1. From the figure, we can see that a combination of the two kinds of dynamic features obtained from short window lengths is less effective than a combination of short- and long-term dynamic features that yielded improved performance. In this experiment, the combination of short-term dynamic features with θ=3 and long-term dynamic features with θ=8 had the highest accuracy of 88.98% and outperformed the standard feature parameter set with MFCC, ∆cep., ∆∆cep., ∆power, and ∆∆power.

Table 1 shows the detailed experimental results. We compare the proposed method with the standard feature parameter set of MFCC38. The proposed method achieved higher performance over MFCC38 in all cases, and especially improved the performance in the high-speed driving case. In MFCC38, the values θ=3 for the delta window and θ=2 for the acceleration window were typically used for the ASR system. The proposed method showed 17.5% error reduction for the overall accuracy in comparison with MFCC38. In particular, our proposed method significantly improved performance for the “Audio On” case in the high-speed driving condition. The singing of the audio CD was a very severe noise source for the ASR, but the singing voices under the high-speed conditions have less impact than under the idling conditions because the singing voices are partially masked with the stationary noises of the high-speed driving that the ASR system trained.

3.3. Discussion

Table 2 shows the detailed experimental results with substitution errors, deletion errors, and insertion errors. From the table, we can see that the proposed method showed lower error rates in all of the cases except for the deletion error in
Table 2. Substitution, deletion, and insertion errors.

<table>
<thead>
<tr>
<th>Table 2. Substitution, deletion, and insertion errors.</th>
<th>Error rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car speed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline MFCC38:</td>
</tr>
<tr>
<td></td>
<td>MFCC+ short Accep. + short ΔAccep. + short ΔAP</td>
</tr>
<tr>
<td></td>
<td>Proposed method:</td>
</tr>
<tr>
<td></td>
<td>MFCC+ short Accep. + long Accep. + short ΔP+ long ΔP</td>
</tr>
<tr>
<td>Substitution Error</td>
<td></td>
</tr>
<tr>
<td>Idling</td>
<td>3.69</td>
</tr>
<tr>
<td>Low speed</td>
<td>4.86</td>
</tr>
<tr>
<td>High speed</td>
<td>7.17</td>
</tr>
<tr>
<td>Overall</td>
<td>5.18</td>
</tr>
<tr>
<td>Deletion Error</td>
<td></td>
</tr>
<tr>
<td>Idling</td>
<td>1.32</td>
</tr>
<tr>
<td>Low speed</td>
<td>3.83</td>
</tr>
<tr>
<td>High speed</td>
<td>6.59</td>
</tr>
<tr>
<td>Overall</td>
<td>3.86</td>
</tr>
<tr>
<td>Insertion Error</td>
<td></td>
</tr>
<tr>
<td>Idling</td>
<td>5.49</td>
</tr>
<tr>
<td>Low speed</td>
<td>3.98</td>
</tr>
<tr>
<td>High speed</td>
<td>3.50</td>
</tr>
<tr>
<td>Overall</td>
<td>4.31</td>
</tr>
</tbody>
</table>

Insertion errors under high-speed conditions were lower than those under low-speed conditions, whereas deletion errors were higher under high-speed conditions. The error rate of overall insertion errors was 3.08%. The insertion error rate decreased significantly by incorporating long-term dynamic features into the feature parameter set.

4. Conclusion

In future work, we will investigate the robustness by using other recognition tasks and their relationships with speech rates. Also, we will evaluate our method using spontaneous speech corpora.

Acknowledgement

The present study was conducted using the CENSREC-2 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

References