Speech Recognition using Long-Term Phase Information

Kazumasa Yamamoto, Eiichi Sueyoshi, and Seiichi Nakagawa

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
{kkyama,sueyoshi,nakagawa}@slp.cs.tut.ac.jp

Abstract
Current speech recognition systems use mainly amplitude spectrum-based features such as MFCC for acoustic feature parameters, while discarding phase spectral information. The results of perceptual experiments, however, suggested that phase spectral information based on long-term analysis includes certain linguistic information. In this paper, we propose the use of phase features based on long-term analysis for speech recognition. We use two types of parameters: the delta phase parameter as a group delay and analytic group delay features. Isolated word and continuous digit recognition experiments were performed, resulting in a greater than 90% word or digit accuracy for each of the experiments. The experimental results confirmed that a long-term phase spectrum includes sufficient information for recognizing speech. Furthermore, combining likelihoods of MFCC and long-term group delay cepstrum outperformed the baseline MFCC relatively 20% for clean speech.

Index Terms: speech recognition, phase information, long-term analysis, group delay

1. Introduction
Current speech recognition systems use mainly amplitude spectrum-based features, such as Mel-Frequency Cepstral Coefficients (MFCC), for the acoustic feature parameters. Phase spectral information, on the other hand, is usually discarded. The importance of phase in human speech recognition has been reported in [1, 2, 3]. Oppenheim et al. [4] and Liu et al. [1] concluded that the importance of the amplitude spectrum or phase spectrum depends on the length of the analysis window. In [1], Liu et al. investigated the role of phase information in the human perception of intervocalic plosives. Their experimental results showed that the perception of intervocalic plosives varies from amplitude dominance to phase dominance as the Fourier analysis window size increases and that the crossover lies somewhere between 192 ms and 256 ms. In [2], Paliwal and Alsteris also investigated the relative importance of short-term magnitude and phase spectra on speech perception. Human perception experiments were conducted to measure intelligibility of speech tokens synthesized from either the magnitude or phase spectrum. It was shown that even for shorter windows, the phase spectrum can contribute as much as the magnitude spectrum to speech intelligibility if the shape of the window function is properly selected. Furthermore, in [5], recognition experiments were carried out with MFCC derived from speech synthesized from the long-term phase spectrum. The results showed the potential for speech recognition with the phase spectrum. In addition, a feature parameter set using a long analysis window has been proposed for robust speech recognition [6]. In [3], Shi et al. analyzed the effects of uncertainty in the phase of speech signals on the word recognition error rate of human listeners. Their experimental results indicate that a small amount of phase error or uncertainty does not affect the recognition rate. Various other researchers have put a great deal of effort into modeling and incorporating phase in the recognition process [7, 8]. For speaker identification, we proposed the use of phase spectral information, thereby improving the recognition performance [9].

In this paper, we propose the use of phase features based on long-term (greater than 100 ms) analysis for speech recognition. The phase features were extracted as group delay spectral features. We used two types of parameters: the delta phase parameter for frequency as a group delay and an analytic group delay cepstrum.

2. Long-Term Phase Spectral Features
2.1. Phase spectrum
When applying analysis window $w(\tau)$ to speech wave $x(t)$ at time $t$, a complex spectrum $S(f, t)$ can be obtained by Fourier transform for frequency $f$ as follows:

$$ S(f, t) = \int_{-\infty}^{\infty} x(t + \tau)w(\tau)\exp(-jft)d\tau = |S(f, t)|\exp(j\theta(f, t)), $$

(1)

where $|S(f, t)|$ and $\theta(f, t)$ denote, respectively, the amplitude and phase spectra of $x(t)$. The phase spectrum $\theta(f, t)$ is defined as:

$$ \theta(f, t) = \angle S(f, t). $$

(2)

The phase value changes according to the clipping position of the analysis window even for the same frequency. Therefore, the phase information cannot be used directly for speech recognition.

2.2. Group delay features
In this paper, we investigated group delay features as the phase information. Group delay is defined as the differentiation of the phase for frequency, that is,

$$ G(f) = -\frac{d\theta(f)}{df}, $$

(3)

where $\theta(f)$ is the phase spectrum for $x(t)$ at frequency $f$.

The value of the phase spectrum obtained by Equation 2 ranges between $-\pi$ and $\pi$, which is not true phase. There are some discontinuous points in the phase spectrum and the differentiation cannot be done at these points. Therefore, the phase spectrum needs to be smoothed by so-called “phase unwrapping” as in Figure 1. However, since the phase obtained in this way is a discrete signal, differentiation is difficult even for the unwrapped phase.
In this paper, the group delay features were calculated using the following two methods:

- Delta phase parameters for frequency as group delay features, and
- Analytic group delay features.

2.2.1. Delta phase parameters for frequency as group delay features

As illustrated in Figure 2, the phase spectral slope, i.e., delta phase parameters for frequency $\Delta f$, was calculated for the unwrapped phase spectrum to approximate the differentiation of the phase spectrum, using the following equation:

$$
\Delta f = \frac{\sum_{k=1}^{K} k(\theta(l_f + k) - \theta(l_f - k))}{2\sum_{k=1}^{K} k^2},
$$

where $l_f$ denotes the index for a discrete Fourier transform (DFT) analysis bin corresponding to frequency $f$, and $K$ is the width of the delta-window for the index. $K$ depends on the frequency resolution of the DFT. For example, the effective frequency band is between 0 and 8000 Hz for speech signals sampled at 16 kHz. After the DFT with a 256 ms (i.e., 4096 pts) window has been completed for the signal, the effective frequency bin indices range from 0 to 2047. If we were to use 30-dimensional parameters, these bins would be divided equally into 91 bands, with each band width containing 66 bins ($\approx 2048/31$). Consequently, $K$ is set to 35 ($= 66/2$) in this case.

2.2.2. Analytic group delay features

The second method for group delay estimation involves an analytic estimation [10] expressed as:

$$
G_{ana}(f) = \frac{X_R(f)Y_R(f) + X_I(f)Y_I(f)}{|X(f)|^2},
$$

where $X_R(f)$ and $X_I(f)$ denote, respectively, the real and imaginary parts of the Fourier complex spectrum for speech signal $x(t)$, and $Y_R(f)$ and $Y_I(f)$ denote, respectively, the real and imaginary parts of the Fourier complex spectrum for the $t$-multiplied speech signal $tx(t)$.

In the recognition experiments, the group delay features were not used directly; instead two smoothed features were used. One of these was the moving-averaged group delay features. After applying the moving average process to the group delay features, the resulting features were sampled out along with equal frequency spans to obtain parameters of the required dimensions. The other was a feature, cepstrally transformed by a discrete cosine transform (DCT). We call this cepstral parameter the “Long-Term Group Delay Cepstrum” (LTGDC for short) due to the use of a long-term analysis window.

3. Recognition Experiments

3.1. Experiments on an isolated word database

3.1.1. Experimental setup

CIABR-VCV, an in-room child speech database [11], was used in the first experiment. Only “contents A”, containing about 30 word utterances for each speaker, was used. 125 speakers were used for training and 20 speakers for testing. The sampling frequency of these data was 16 kHz.

A Hamming analysis window was used in this experiment; the selected window lengths were 25 ms, 100 ms, and 256 ms, while the frame shift was fixed at 10 ms. The feature parameters were 10/20/30-dimensional $\Delta f$ parameters described in Section 2.2.1, 30-dimensional moving-averaged group delay features, and 10/20-dimensional LTGDCs described in Section 2.2.2. The window for the moving-average was 4 or 9 pts. For comparison with these parameters, a 38-dimensional MFCC (12MFCC +12ΔMFCC +12ΔΔMFCC +ΔPow + ΔΔPow) was used as a baseline. The length of the analysis window for the MFCC was 25 ms.

As acoustic models, 30 whole-word HMMs were used. Each HMM had 23 states, each of which comprised 8 Gaussian mixture components.

3.1.2. Experimental results

The word recognition rate using the baseline MFCC was 99.5%.

The experimental results using the group delay features are given in Tables 1, 2, and 3. Although the recognition results were particularly poor for all the group delay features using a short analysis window, i.e., the same length as that for the MFCC (25 ms), more than 90% accuracy in the recognition results was achieved with a long analysis window. These results confirmed that we could recognize speech using only the phase information.

In the comparison of the three kinds of group delay features, the performance of the moving-averaged group delay features was slightly worse, while the other two features, LTGDC and $\Delta f$ parameters, achieved a similar recognition performance. The difference in dimensions of the two features did not result in a big difference in the recognition results.
In this section, a long analysis window (frame) was used for the recognition experiments. Consequently, an experiment with a varying frame shift for the $\Delta f_p$ parameters was conducted to investigate the effect of the frame shift length. The experimental results for various frame shifts, given in Table 4, are very similar to those for the 256 ms analysis window.

### 3.2.2. Experimental results

The experimental results are presented in Table 5. Table 5 (c) gives the digit accuracy using MFCC (baseline), while (b) shows the digit accuracy using only LTGDC. Figures in bold indicate an improvement of over 10% compared with the MFCC baseline. The recognition rate for clean speech was over 90% due to their phase being random.

### 3.2.2. Experimental results

The experimental results are given in Table 5 (c). Figures in bold indicate an improvement of over 10% compared with the MFCC baseline. The method using the combined likelihoods of MFCC and LTGDC outperformed that using MFCC alone. The relative improvement was over 10% for the higher SNR data, especially over 20% for “Clean”. We have been sufficiently encouraged by the results to conduct further research on the use of phase information.

### 4. Conclusion

In this paper, we proposed the use of phase features based on long-term analysis for speech recognition. The phase features were extracted as group delay spectral features. The experimental results confirmed that the long-term phase spectrum included sufficient information to recognize speech. Furthermore, combining likelihoods of MFCC and LTGDC outperformed that using MFCC alone for the higher SNR data.

In future works, we will investigate combining the LTGDC and MFCC vectors in the feature domain, and applying the LTGDC for LVCSR experiments. Additionally, in [13] Kua et al. proposed a modified group delay:

$$G_{mod}(f) = \text{sign} \left| \frac{X_R(f)Y_R(f) + X_I(f)Y_I(f)}{|S_c(f)|^2} \right|^\beta,$$

where $|S_c(f)|^2$ is a cepstrally smoothed power spectrum, $\beta$ and $\gamma$ are coefficients for smoothing, respectively. Although we did not use the modified group delay in this paper, we intend considering it in future work.
Table 5: Digit accuracy for CENSREC-1 database
(a) Digit accuracy using MFCC (baseline) [%]

<table>
<thead>
<tr>
<th>SNR</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>20dB</td>
<td>97.88</td>
<td>72.16</td>
<td>84.97</td>
<td>97.78</td>
<td>80.99</td>
<td>83.52</td>
<td>80.11</td>
<td>82.78</td>
</tr>
<tr>
<td>15dB</td>
<td>84.77</td>
<td>53.51</td>
<td>55.56</td>
<td>88.58</td>
<td>57.54</td>
<td>60.85</td>
<td>57.26</td>
<td>55.29</td>
</tr>
<tr>
<td>10dB</td>
<td>57.02</td>
<td>37.58</td>
<td>37.22</td>
<td>60.78</td>
<td>38.66</td>
<td>39.78</td>
<td>42.38</td>
<td>38.26</td>
</tr>
<tr>
<td>0dB</td>
<td>10.53</td>
<td>6.17</td>
<td>10.53</td>
<td>7.46</td>
<td>7.77</td>
<td>7.46</td>
<td>7.46</td>
<td>7.28</td>
</tr>
<tr>
<td>Ave.</td>
<td>55.72</td>
<td>38.12</td>
<td>41.88</td>
<td>57.12</td>
<td>40.72</td>
<td>43.98</td>
<td>44.13</td>
<td>41.74</td>
</tr>
</tbody>
</table>

(b) Digit accuracy using only LTGDC [%]

<table>
<thead>
<tr>
<th>SNR</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>90.70</td>
<td>90.60</td>
<td>92.31</td>
<td>92.38</td>
<td>90.70</td>
<td>90.60</td>
<td>92.38</td>
<td>92.38</td>
</tr>
<tr>
<td>20dB</td>
<td>61.47</td>
<td>83.65</td>
<td>81.96</td>
<td>42.36</td>
<td>82.38</td>
<td>72.61</td>
<td>84.70</td>
<td>81.06</td>
</tr>
<tr>
<td>15dB</td>
<td>42.55</td>
<td>75.39</td>
<td>67.79</td>
<td>22.43</td>
<td>73.17</td>
<td>56.32</td>
<td>73.43</td>
<td>67.20</td>
</tr>
<tr>
<td>10dB</td>
<td>23.67</td>
<td>59.04</td>
<td>43.63</td>
<td>6.54</td>
<td>57.54</td>
<td>37.88</td>
<td>51.18</td>
<td>42.61</td>
</tr>
<tr>
<td>5dB</td>
<td>10.93</td>
<td>36.70</td>
<td>18.76</td>
<td>-2.16</td>
<td>36.94</td>
<td>20.34</td>
<td>28.93</td>
<td>21.32</td>
</tr>
<tr>
<td>0dB</td>
<td>4.30</td>
<td>16.51</td>
<td>11.06</td>
<td>-5.52</td>
<td>18.08</td>
<td>8.74</td>
<td>14.91</td>
<td>11.72</td>
</tr>
<tr>
<td>Ave.</td>
<td>28.58</td>
<td>54.26</td>
<td>44.64</td>
<td>12.73</td>
<td>53.62</td>
<td>39.18</td>
<td>50.63</td>
<td>44.78</td>
</tr>
</tbody>
</table>

(c) Digit accuracy with rescoring and likelihood combination of MFCC and LTGDC [%]

<table>
<thead>
<tr>
<th>SNR</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>99.79</td>
<td>99.67</td>
<td>99.94</td>
<td>99.75</td>
<td>99.79</td>
<td>99.67</td>
<td>99.94</td>
<td>99.75</td>
</tr>
<tr>
<td>20dB</td>
<td>98.22</td>
<td>75.79</td>
<td>87.56</td>
<td>97.84</td>
<td>84.19</td>
<td>85.85</td>
<td>85.53</td>
<td>83.26</td>
</tr>
<tr>
<td>15dB</td>
<td>87.14</td>
<td>56.05</td>
<td>59.14</td>
<td>88.71</td>
<td>60.79</td>
<td>63.30</td>
<td>63.30</td>
<td>59.06</td>
</tr>
<tr>
<td>10dB</td>
<td>59.78</td>
<td>39.39</td>
<td>39.34</td>
<td>61.15</td>
<td>41.14</td>
<td>41.87</td>
<td>44.65</td>
<td>40.51</td>
</tr>
<tr>
<td>0dB</td>
<td>12.47</td>
<td>7.80</td>
<td>10.17</td>
<td>10.74</td>
<td>8.41</td>
<td>12.91</td>
<td>15.87</td>
<td>11.45</td>
</tr>
<tr>
<td>Ave.</td>
<td>57.45</td>
<td>40.33</td>
<td>44.72</td>
<td>57.00</td>
<td>43.02</td>
<td>45.76</td>
<td>46.66</td>
<td>44.06</td>
</tr>
</tbody>
</table>

5. Acknowledgments
This work was supported in part by the Ono Acoustics Research Fund.

6. References