EVALUATION OF SPECTRAL SUBTRACTION WITH SMOOTHING OF TIME DIRECTION ON THE AURORA 2 TASK

Norihide Kitaoka and Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, Aichi 441-8580, Japan
E-mail: {kitaoka, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

To reduce the effects of additive noises, spectral subtraction (SS) is often used. We discuss SS on the power spectral domain. This method has two problems to make the estimation of clean speech difficulties: (1) There exists the estimation error between true power spectrum of noise and estimated one (2) The correlation between speech and noise also exists because of the phase difference.

To overcome these problems, we proposed a spectral subtraction using a smoothing method of time direction. We consider the average of estimated speech power spectra over some frames as the estimated speech power spectrum. This operation makes the estimation of noise more accurate. We can reduce the effect of correlation between speech and noise.

In this paper, we tested this method on the AURORA 2 database, which consists of English connected digit added with various realistic noises. We achieved 47.26% relative improvement of word accuracy with acoustic models trained under clean condition and 11.95% with models trained under multi-condition.

1. INTRODUCTION

One of the problems on the application of the speech recognition system in the real world is the degradation of the performance by surrounding noises. Some techniques are proposed to reduce the effects of additive noise [1][2]. Spectral subtraction (SS) [3] is an effective method for stationary additive noises with low computational costs.

We discuss SS in the power spectral domain. This method has two problems: (1) The power spectrum of noise cannot be estimated accurately, so there is estimation error between true power spectrum of noise and estimated one, (2) There exists the correlation between speech and noise because of the phase difference, which makes difficult to estimate the power spectrum of clean speech accurately.

We proposed spectral subtraction using a smoothing method of time direction to solve these problems [4]. This method makes the estimation of speech more accurate. We can also reduce the effect of correlation between speech and noise. Experimental results on the AURORA 2 noisy connected digit recognition task are shown below.

2. PROBLEMS OF SS

The observation signal \( x \) is assumed to be the sum of speech signal \( s \) and noise \( n \), namely, \( x = s + n \). Spectral subtraction in power spectral domain is defined as below:

\[
|\tilde{S}_i(t)|^2 = |X_i(t)|^2 - \alpha|\tilde{N}_i|^2, \tag{1}
\]

where \( |\tilde{S}_i(t)|^2 \) and \( |X_i(t)|^2 \) are the \( i \)-th components of the estimated power spectrum of speech and the power spectrum of observed signal at the time \( t \), respectively, \( |\tilde{N}_i| \) is the \( i \)-th component of a priori estimated power spectrum of noise, and \( \alpha \) is the over estimation factor. We can express \( |X_i(t)|^2 \) as:

\[
|X_i(t)|^2 = |S_i(t) + N_i(t)|^2
= |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)|\cos \theta_i(t), \tag{2}
\]

where \( |S_i(t)| \) and \( |N_i(t)| \) are the true values for speech and noise, and \( \theta_i(t) \) is the phase difference between speech and noise. We suppose that the speech and the noise do not correlate each other. The definition of Equation (1) stands on the fact that the expectation value of \( \cos \theta_i(t) \) in Equation (2) equals zero. However, considering \( \cos \theta_i(t) \) as a random variable ranging \(-1 \) to \(1 \) and assuming that \( \theta_i(t) \) distributes uniformly, the probability density function \( f(\phi) = \frac{1}{\pi \sqrt{1 - \phi^2}} \) becomes [4]

\[
f(\phi) = \frac{1}{\pi \sqrt{1 - \phi^2}}. \tag{3}
\]

This function is plotted in Figure (1). This is a concave function with sole minimum at \( \theta = 0 \). So the term including \( \cos \theta_i(t) \) in Equation (2) cannot be removed even if the noise power can be accurately estimated.

Yoma et al. considered the correlation term as a hidden information and weigh the matching of the template against the testing pattern with the reliability based on the estimation of the hidden information [5]. A modified method of PMC and an adaptation method of HMM to the noisy environment based on Taylor series expansion considering the correlation terms were proposed by the literature [6][7].
Fig. 1. Pdf of $\phi = \cos \theta$ assuming that $\theta$ distributes uniformly between $-\pi$ and $\pi$.

3. SS WITH THE SMOOTHING METHOD OF TIME DIRECTION (SS-SMT)

To reduce the effect of correlation between speech and noise, we propose to use SS with the smoothing of time direction (SS-SMT). We define the smoothing method as below:

$$|X_i(t)|^2 = \sum_\tau \beta_\tau |S_i(t-\tau) + N_i(t-\tau)|^2,$$

(4)

$$= \sum_\tau \beta_\tau |S_i(t-\tau)|^2 + \sum_\tau \beta_\tau |N_i(t-\tau)|^2 + 2 \sum_\tau \beta_\tau |S_i(t-\tau)||N_i(t-\tau)| \cos \theta_i(t-\tau).$$

(5)

Assuming speech and noise are stable for the period $T$,

$$\sum_\tau \beta_\tau |S_i(t-\tau)|^2 \approx |S_i(t)|^2,$$

(7)

$$\sum_\tau \beta_\tau |N_i(t-\tau)|^2 \approx |N_i(t)|^2,$$

(8)

$$\sum_\tau \beta_\tau |S_i(t-\tau)||N_i(t-\tau)| \cos \theta_i(t-\tau) \approx |S_i(t)||N_i(t)| \sum_\tau \beta_\tau \cos \theta_i(t-\tau).$$

(9)

So Equation (6) becomes

$$|X_i(t)|^2 \approx |S_i(t)|^2 + |N_i(t)|^2 + 2|S_i(t)||N_i(t)| \sum_\tau \beta_\tau \cos \theta_i(t-\tau).$$

(10)

Replacing $|X_i(t)|^2$ in Equation (1) with $|X_i(t)|^2$, Equation (1) becomes

$$|\tilde{S}_i(t)|^2 = \frac{|X_i(t)|^2 - \alpha |\tilde{N}_i|^2}{|S_i(t)|^2 + |N_i(t)|^2 - \alpha |\tilde{N}_i|^2}.$$  

(11)

where $\phi = \sum \beta_\tau \cos \theta_i(t-\tau)$. Assuming phase differences between speech and noise of successive frames don’t correlate each other, the pdfs of $\phi$ are shown in Figure 2 ($\beta_\tau = 1/T$). The function $f(\phi)$ has the peak at zero and the variance of this term becomes smaller than the original one. The standard deviation is $1/\sqrt{2T}$ for each $T$ when the successive frames are uncorrelated. In fact, $\phi$ led from the phase difference between the components of real speech and noise follows $1/\sqrt{2T}$ as Figure 3. So we can assume the third term of Equation (6) is almost zero and the Equation (6) becomes

$$|\tilde{S}_i(t)|^2 \approx |S_i(t)|^2 + |N_i(t)|^2.$$  

(12)

Replacing $|X_i(t)|^2$ in Equation (1) with $|X_i(t)|^2$, Equation (1) becomes

$$|\tilde{S}_i(t)|^2 = \frac{|X_i(t)|^2 - \alpha |\tilde{N}_i|^2}{|S_i(t)|^2 + |N_i(t)|^2 - \alpha |\tilde{N}_i|^2}.$$  

(13)

So we can estimate the speech signal more accurately if we can estimate $|\tilde{N}_i|$ accurately.

4. EXPERIMENTAL RESULTS

Recognition experiments were performed on the Aurora 2 speaker independent connected digit recognition task [8]. Baseline system consisted of FE2.0 front-end included in Aurora 2 CD-ROM and HTK ver 3.0 [9].

We trained digit HMMs with either 8440 clean utterances (clean condition training) or with 8440 clean and noisy utterances (multi-condition training). Each digit HMM had 16 states with 3 Gaussian mixture emission probability. A 3-state silence model and a 1-state short pause model were also prepared.
Feature vectors extracted by FE2.0 contains 12 MFCCs and log energy. We used pre-emphasis factor 0.97. Frame length and frame shift length were set to 25 ms and 10 ms respectively. HTK uses these features and their first and second derivatives.

Test set consist of three subsets of utterances under known noises (set A), utterances under an unknown noises (set B) and noisy utterances through unknown channel (set C). SNR were set to 20, 15, 10, 5, 0, −5 dB.

In the experiments, we used spectral subtraction with smoothing of time direction and cepstral mean normalization (CMN) [10]. For spectral subtraction, we estimated noise spectrum from first ten frames of each speech file. Over estimation factor and the number of frames used in smoothing were tuned for each method and smoothing coefficients β, was set to 1/T for τ = 1, · · · , T.

We averaged the recognition accuracy improvements over all the set from 20 dB to 0 dB and compared it with that of the baseline system. The results are summarized in Table 1.

Applying only CMN, we could obtain good improvement. This is because CMN can normalize not only channel distortion but the distortion by additive noises to some extent. In fact, accuracy for set A,B and C were all improved under clean training condition. Under multi-condition, on the contrary, CMN could improve only set C. This means that multi-condition training already absorbed almost all the spectral distortion which could be compensated by CMN.

SS could improve the recognition performances of clean condition and multi-condition. Smoothing method also worked well. Under the multi-condition, residual noises (called ‘musical noises’) were also modeled in the acoustic models, so the effect of smoothing was small. Under the clean condition, residual noises were not modeled in the acoustic models and the smoothing method attenuated them.

Figure 4 shows the effects of the number of frames T used in smoothing on relative recognition improvements (clean training).

Table 2 shows the detail of recognition results by SS-SMT + CMN. We can find significant improvements for noisy utterances, but some degradations especially for clean utterances were also found. Fig. 4 shows the effects of the number of frames used in smoothing. T = 1 means the conventional SS. T = 3 improved the accuracy for all SNR. T = 5 also improved the performances under noisy conditions, but that under clean conditions got worse because the time resolution got worse.

5. CONCLUSION

We proposed a smoothing method of time direction on speech recognition under noisy environments using spectral subtraction on the power spectral domain. This method, SS-SMT, reduces the effect of the correlation between speech and noise. We also show the effect of SS-SMT experimentally. We achieved 47.26% relative improvement under clean condition training on AURORA 2 English connected digit recognition and 11.95% under multi-condition training.

We assumed stable noises, but realistic noise often non-stable. We will develop an adaptive noise estimation method to catch up with changes of noises.
### Table 2. Recognition results using SS-SMT.

#### (a) clean condition (absolute)

<table>
<thead>
<tr>
<th>Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>90.05</td>
<td>98.15</td>
<td>98.90</td>
<td>99.12</td>
</tr>
<tr>
<td>Subway</td>
<td>96.94</td>
<td>98.89</td>
<td>99.01</td>
<td>99.07</td>
</tr>
<tr>
<td>Babble</td>
<td>94.17</td>
<td>93.14</td>
<td>96.64</td>
<td>93.73</td>
</tr>
<tr>
<td>Car</td>
<td>94.29</td>
<td>92.35</td>
<td>94.80</td>
<td>95.05</td>
</tr>
<tr>
<td>Exhibition</td>
<td>95.06</td>
<td>96.98</td>
<td>74.68</td>
<td>94.74</td>
</tr>
<tr>
<td>Average</td>
<td>95.87</td>
<td>97.64</td>
<td>98.86</td>
<td>97.46</td>
</tr>
</tbody>
</table>

#### (b) clean condition (relative)

<table>
<thead>
<tr>
<th>Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>89.52</td>
<td>6.90%</td>
<td>28.62%</td>
<td>3.46%</td>
</tr>
<tr>
<td>Subway</td>
<td>90.05</td>
<td>99.90</td>
<td>97.62</td>
<td>99.90</td>
</tr>
<tr>
<td>Babble</td>
<td>89.72</td>
<td>98.05</td>
<td>95.65</td>
<td>98.05</td>
</tr>
<tr>
<td>Car</td>
<td>90.20</td>
<td>91.81</td>
<td>92.15</td>
<td>91.81</td>
</tr>
<tr>
<td>Exhibition</td>
<td>89.21</td>
<td>80.40</td>
<td>73.95</td>
<td>80.40</td>
</tr>
<tr>
<td>Average</td>
<td>88.58</td>
<td>66.57</td>
<td>58.86</td>
<td>66.57</td>
</tr>
</tbody>
</table>

#### (c) multi condition (absolute)

<table>
<thead>
<tr>
<th>Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>76.20</td>
<td>76.39</td>
<td>80.20</td>
<td>78.39</td>
</tr>
<tr>
<td>Subway</td>
<td>86.10</td>
<td>94.08</td>
<td>97.96</td>
<td>95.03</td>
</tr>
<tr>
<td>Babble</td>
<td>15.36</td>
<td>12.27</td>
<td>66.53</td>
<td>41.92</td>
</tr>
<tr>
<td>Car</td>
<td>10.00</td>
<td>15.43</td>
<td>78.05</td>
<td>27.85</td>
</tr>
<tr>
<td>Exhibition</td>
<td>73.00</td>
<td>72.27</td>
<td>95.13</td>
<td>79.65</td>
</tr>
<tr>
<td>Average</td>
<td>50.38</td>
<td>60.00</td>
<td>64.44</td>
<td>55.91</td>
</tr>
</tbody>
</table>

#### (d) multi condition (relative)

<table>
<thead>
<tr>
<th>Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>25.61</td>
<td>60.10%</td>
<td>50.26%</td>
<td>32.92%</td>
</tr>
<tr>
<td>Subway</td>
<td>14.11</td>
<td>50.50%</td>
<td>54.84%</td>
<td>46.89%</td>
</tr>
<tr>
<td>Babble</td>
<td>31.50</td>
<td>50.50%</td>
<td>74.68%</td>
<td>62.14%</td>
</tr>
<tr>
<td>Car</td>
<td>58.57</td>
<td>50.50%</td>
<td>67.43</td>
<td>60.37%</td>
</tr>
<tr>
<td>Exhibition</td>
<td>31.50</td>
<td>50.50%</td>
<td>28.38</td>
<td>57.53%</td>
</tr>
<tr>
<td>Average</td>
<td>50.38</td>
<td>50.38</td>
<td>50.38</td>
<td>50.38</td>
</tr>
</tbody>
</table>