EXPLORING HIGH-PERFORMANCE SPEECH RECOGNITION IN NOISY ENVIRONMENTS USING HIGH-ORDER TAYLOR SERIES EXPANSION

Guo-Hong Ding+ and Bo Xu+*  
High-Tech Innovation Center+, National Laboratory of Pattern Recognition*  
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080  
{ghding, xubo}@hitic.ia.ac.cn

ABSTRACT
In this paper, high-order Taylor Series expansion is proposed to explore the most effective formulas of log-spectral compensation. The power feature, which is crucial to speech recognition in noisy environments and can’t be compensated in usual feature compensation, is processed similarly to spectral subtraction. The modeling accuracy of speech log-spectral Gaussian Mixture Model (GMM) is also discussed and carefully treated. Experimental results show that the log-spectral compensation can greatly improve recognition performance in noisy environments and with the acoustic model trained using multi-condition data, the recognition performance is superior to that in matched conditions.

1. INTRODUCTION
Noise robustness is crucial to all types of speech recognition applications. Feature compensation is one of the major aspects for robust speech recognition in noisy environments. There exist two major methods in feature compensation: the stereo-based method, such as FDCDN, RATZ and SPLICE [10], and the structure-based method, such as CDCN [4] and the VTS-based feature compensation [5, 6].

The stereo-based feature compensation needs a set of stereo pairs to train the noisy Gaussian Mixture Model (GMM) and estimate the biases between corrupted features and uncorrupted ones. However, there exist too many noise types and noise levels, so the training data to build the GMM and to estimate the biases are formidable. In the case, it is difficult to give an accurate description of noise distribution with just a GMM and as a result it is hard to obtain accurate estimates of uncorrupted speech features in unknown noisy environments. For the reason, the stereo-based method is sensitive to noise type and a switch to different noisy environments is needed to achieve desirable recognition performance [9].

In the usual structure-based feature compensation [5, 6], the nonlinear environmental function is utilized to estimate environmental parameters in the ML framework. Then the noisy distribution is constructed. Finally, uncorrupted speech features are estimated in the MMSE criterion. In the method, the environmental function is used to describe the influence of additive noise and channel distortion on speech signals and the noisy distribution can be obtained from the clean speech log-spectral GMM and the estimated environmental parameters. Thus the method can deal with unknown noisy environments without environmental priori knowledge.

The structure-based method can be implemented in the log-spectral domain [5] or in the cepstral domain [6]. For the reason of discrete cosine transform (DCT), cepstra are considered more uncorrelated than log-spectra. However, for the same reason, there exist many matrix calculations in cepstral compensation, which may lead to heavy computation requirements. In the log-spectral domain, the environmental function can be divided into independent functions of the variables of the vector and the variables can be processed separately. Thus log-spectral compensation is more suitable to give real-time implementations in speech recognition applications. Clearly the poor independence of log-spectra may affect the clustering accuracy in constructing the log-spectral GMM. The paper explores structure-based method in the log-spectral domain. Speech features are clustered in the cepstral domain and then they are transformed into the log-spectral domain to construct the log-spectral GMM. Thus the clustering accuracy in the log-spectral domain is the same as that in the cepstral domain.

In this paper, noise statistics are obtained when speech is absent. Noise log-spectra are modeled as single Gaussian distribution and its mean and variance are used to construct noisy distribution and to estimate the uncorrupted log-spectra using high-order Taylor Series expansion.

Furthermore, since the power feature, which is crucial to speech recognition in noisy environments, is difficult to compensate in the structure-based method. In this paper, a subtraction similar to spectral subtraction is treated on the power feature and it greatly improves the recognition performance, especially in low SNRs.

2. ENVIRONMENTAL FUNCTION IN THE LOG-SPECTRAL DOMAIN
The environmental model has been analyzed in detail in [5, 6]. In the log-spectral domain, the effect of additive noise and channel distortion on the clean feature satisfy

$$y = x + q + \log(I + e^{\alpha e^t})$$

where $y = [y_0, ..., y_{n-1}]$, $x = [x_0, ..., x_{n-1}]$, $n = [n_0, ..., n_{n-1}]$ and $q = [q_0, ..., q_{n-1}]$ are the log-spectral vectors of noisy speech, clean speech, additive noise and channel distortion, respectively.

The above model that relates $x$, $y$, $n$ and $q$ is nonlinear. In this paper, only additive noise is considered and channel distortion is assumed equal to zero. Then the nonlinear environmental model is rewritten as

$$y = x + \log(I + e^{\alpha e^t})$$

(1)

3. CONSTRUCTION OF NOISY DISTRIBUTION USING HIGH-ORDER TAYLOR SERIES EXPANSION
In the log-spectral domain, noise log-spectra can be modeled as
\[ p(n) = \prod_{i=1}^{N} N(n_i; \mu_{ni}, \sigma_{ni}^2) \]  

where \( N(n_i; \mu_{ni}, \sigma_{ni}^2) \) denotes the Gaussian probability density function with mean \( \mu_{ni} \) and variance \( \sigma_{ni}^2 \), while the speech is modeled as GMM

\[
p(x) = \sum_{n} c_n \prod_{i} N(x_i; \mu_{ni}, \sigma_{ni}^2) \tag{3}
\]

In developing the high-order Taylor series expansion of the environmental function, we just deal with the distribution of speech log-spectra as single Gaussian distribution for simplicity and the mixture components will be considered in the next section. Besides, from equation (1), it is obvious that the environmental function can be divided into several separate functions and thus we can consider just one variable of the vector. In the rest of section 3, the environmental function is written as

\[
y = x + \log(1 + \exp(n - x)) \tag{4}
\]

where \( x \) and \( n \) can be modeled as \( p(n) = N(n; \mu_{n}, \sigma_{n}^2) \) and \( p(x) = N(x; \mu_{x}, \sigma_{x}^2) \), respectively.

3.1 High-order Taylor Series expansion of environmental function

The last term of equation (4) can be written as

\[
f(n - x) = \log(1 + \exp(n - x))
\]

Define

\[
\Delta = n - x
\]

then it is obvious that \( \Delta \) follows Gaussian distribution

\[
p(\Delta) = N(\Delta; \mu_{\Delta}, \sigma_{\Delta}^2)
\]

where \( \mu_{\Delta}, \sigma_{\Delta}^2 \) are modeled as \( \mu_{\Delta} = \mu_{n} - \mu_{x}, \sigma_{\Delta}^2 = \sigma_{n}^2 + \sigma_{x}^2 \).

Equation (5) can be expanded using M-order Taylor Series as

\[
f(\Delta) = \sum_{i=0}^{M} a_i (\Delta - \mu_{\Delta})^i
\]

where \( a_i = \frac{f^{(i)}(\mu_{\Delta})}{i!} \) is the expansion coefficient, and \( f^{(i)}(\mu_{\Delta}) \) denotes the i-order Taylor coefficient. According to the definition of \( f(\Delta), f^{(i)}(\mu_{\Delta}) \) can be written as

\[
log(1 + \exp(\mu_{\Delta})) \quad \text{if } i = 0
\]

\[
1 - \frac{1}{1 + \exp(\mu_{\Delta})} \quad \text{if } i = 1
\]

\[
\sum_{j=1}^{i} A_{ij} \frac{1}{1 + \exp(\mu_{\Delta})^j} \quad \text{if } i \geq 2
\]

When \( i \geq 2 \), the coefficients in equation (6) can be derived as

\[
A_{ii}^{(i)} = 1, A_{i(i-1)}^{(i)} = -\frac{1}{i}, A_{j(j-1)}^{(i)} = -j \cdot A_{j(j-1)}^{(i-1)}, 2 \leq j \leq i
\]

Equation (7) gives iterative formulas to calculate coefficients \( A_{i}^{(i)} \) and the iterations are initiated with

\[
A_{i}^{(1)} = 1, A_{i}^{(2)} = -1
\]

3.2 Calculation of noisy log-spectral mean

According to the environmental function, the noisy log-spectral mean can be derived as

\[
\mu_{y} = \mu_{x} + E[f(\Delta)]
\]

where

\[
E[f(\Delta)] = \int f(\Delta) N(\Delta; \mu_{\Delta}, \sigma_{\Delta}^2) d\Delta
\]

\[
= \sum_{i=0}^{M} a_i M_{\Delta}^{(i)}
\]

where \( M_{\Delta}^{(i)} \) is the moment about the mean (The definition is referred to [2]), which can be solved as

\[
M_{\Delta}^{(i)} = \begin{cases} 0 & \text{if } i \text{ is odd} \\ \frac{1}{i(1-j)!} & \text{otherwise} \end{cases}
\]

3.3 Calculation of noisy log-spectral variance

The noisy variance can be formulated as

\[
\sigma_{y}^2 = E[y^2] - \mu_{y}^2
\]

\[
= E[(x + f(\Delta))^2] - \mu_{y}^2
\]

\[
= E[x^2] + 2E[xf(\Delta)] + E[f(\Delta)^2] - \mu_{y}^2
\]

It is obvious that

\[
E[x^2] = \mu_{x}^2 + \sigma_{x}^2
\]

\[
E[f(\Delta)^2] = \sum_{i=0}^{M} a_i M_{\Delta}^{(i)}
\]

\[
E[xf(\Delta)] = \int xf(\Delta) N(x; \mu_{x}, \sigma_{x}^2) dx
\]

\[
= \int \sum_{i=0}^{M} a_i (x - \mu_{x})^i N(x; \mu_{x}, \sigma_{x}^2) dx
\]

\[
= \mu_{x} E[f(\Delta)] + \sum_{i=0}^{M} a_i \frac{1}{i(1-j)!} M_{\Delta}^{(i)}\
\]

where \( C_{i,j} = \frac{1}{i!} \) is the Binomial Coefficient [2].

In equation (7), \( M_{\Delta}^{(i)} \) and \( A_{ij}^{(i)} \) are defined similar to \( M_{\Delta}^{(i)} \) in equation (10) and \( \sigma_{\Delta}^2 \) is replaced by \( \sigma_{x}^2 \) and \( \sigma_{y}^2 \), respectively.

4. LOG-SPECTRAL COMPENSATION BASED ON HIGH-ORDER TAYLOR SERIES EXPANSION

Equation (1) can be rewritten as

\[
x = y - f(\Delta)
\]

then the uncorrupted log-spectra can be estimated as

\[
\hat{x} = E[x | y] = y - E[f(\Delta) | y]
\]

Assume that the \textit{a posteriori} probability density function (pdf) of \( \Delta \) given \( y \) follows the Gaussian distribution as

\[
p(\Delta | y) = N(\Delta; \mu_{\Delta y}, \sigma_{\Delta y}^2)
\]
The *a posteriori* pdf can be obtained similarly to that in [1]. In [1], the process of estimating a suitable expansion point is iterated several times. But it can’t achieve good recognition performance. The reason lies in that the *a posteriori* mean of \( \Delta \) given \( y \) is heavily dependent on \( y \) and thus the expansion may fall into an error convergent area. Thus Taylor series approximation of the environmental function expanded at point \( \mu_\Delta = \mu_y - \mu_\mu \) is more suitable and as a result the priori distribution of \( \Delta \) is appropriate to give a solution of equation (13).

Since the expansion point is \( \mu_\Delta \), the approximation can be implemented in a brief form. The last term of equation (13) can be written as

\[
E[f(\Delta) \mid y] = \int f(\Delta)N(\Delta; \mu_\Delta, \sigma^2_{\Delta})d\Delta
\]

\[
\approx \sum_{i=0}^{M} a_i M_{i}\Delta^i
\]

(14)

where

\[
P(m \mid y) = \frac{c_m p(y \mid m)}{\sum_{m'} c_{m'} p(y \mid m')}\]

\[
E[x \mid y, m] = y - E[f(\Delta) \mid y, m]
\]

(16)

where \( \Delta = [\Delta_0, ..., \Delta_{\Delta_{\max}}] \) is the speech log-spectral vector subtracted by the noise vector. Eq. (16) can be solved by extending Eq. (14) to mixture components and to multiple dimension cases.

5. **PRACTICAL CONSIDERATIONS**

In this section, some practical issues on the log-spectral compensation are considered to make it more effective.

5.1 Compensation on the power feature

In the structure-based method, MFCC can be compensated efficiently, but it is difficult to process the power feature. Thus in some systems [11], the power feature is substituted by the 0th MFCC \( c_0 \). Although speech recognition system with \( c_0 \) can provide almost the same performance as that with the power feature in clean environments, the performance of the former degrades more greatly than the latter in noisy environments. The reason lies in that the power feature is more robust to noise and it can give better segmentation of speech signals.

In this paper, the power feature is compensated similar to spectral subtraction, that is

\[
\hat{p}_i = \begin{cases} 
  p_n - p_c & \text{if } p_n - p_c > Q p_n \\
  Q p_n & \text{otherwise}
\end{cases}
\]

(17)

where \( \alpha \) is floor factor, \( p_n \) is the noisy power of current frame, \( \hat{p}_i \) is the estimated uncorrupted power and \( p_c \) is the noise power estimated using a rough VAD.

5.2 Construction of speech log-spectral GMM

As it stated in the introduction, the difference of log-spectra and cepstra lies in that uncorrelation assumption is more suitable for cepstra and that the cepstral clustering is more accurate.

In constructing speech log-spectral GMM, we try to improve the modeling accuracy. Speech features are clustered in the cepstral domain and then they are transformed into the log-spectral domain to construct the log-spectral GMM. Thus the clustering accuracy of the GMM in the log-spectral domain is the same as that in the cepstral domain.

6. **EXPERIMENTAL EVALUATION**

The baseline recognition system is a speaker-independent large vocabulary continuous speech recognizer. The acoustic model is a class-triphone model, which is trained based on decision tree and the recognition is implemented using a one-pass decoder. Both the modeling and the decoding are described detailed in [3].

In the baseline system, the 39-d feature consists of log-power and 12-d MFCC \( c_1-c_{12} \), their 1st derivatives and 2nd derivatives.

The training and test data are the standard Mandarin continuous speech corporuses supported by Chinese 863 project and they are resampled at 8kHz with 16bit resolution. The test corpus contains 240 sentences from 4 speakers. We obtain the noisy speech by adding noise data, which are chosen from the NOISEX92 database [8], to the clean speech with noise amplitudes being adjusted to achieve different SNRs.

The multi-condition training takes the same measure as in [7]. The training data are equally split into 20 subsets. Each subset contains a few sentences of all training speakers. The 20 subsets represent 4 different noise types at 5 different SNRs. The 4 noises are white, babble, factory and leopard. The SNRs are 20dB, 15dB, 10dB, 5dB and the clean condition.

All the original uncorrupted test data are degraded by four noises, namely white, babble, h16 and destroyerops. The first two noisy data have the same noises as used for the multi-condition training, which leads to a high match of training and test data. In the rest of the paper, the average results on the two noisy data are denoted as Test A. For noisy data degraded by f16 and destroyerops, there exists a mismatch between training and test data for the multi-condition training. This will show the influence on recognition when considering different noises than the one used for training. In the rest of the paper, the average recognition results on the two noisy data are denoted as Test B.

6.1 Baseline recognition results

Table 1 gives the experimental results of the baseline system in clean and noisy environments. The results are obtained in mismatched conditions and are the lower limit of all effective compensation techniques. In mismatched conditions, the average recognition accuracy is 36.58% and 45.09% for Test A and Test B, respectively.

Table 2 gives the experimental result of the recognition system in matched conditions, that is, the test data and the training data are degraded by the same noise with the same level. The results give the upper limit for all model compensation techniques, such as PMC and some adaptation algorithms etc. In matched conditions, the average
recognition accuracy is 53.23% and 59.68% for Test A and Test B, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10dB</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>20dB</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 1. Baseline recognition accuracy (%) in clean and noisy environments

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>1.24</td>
<td>38.36</td>
</tr>
<tr>
<td>10dB</td>
<td>70.15</td>
<td>54.07</td>
</tr>
<tr>
<td>20dB</td>
<td>78.59</td>
<td>85.63</td>
</tr>
</tbody>
</table>

Table 2. Recognition accuracy (%) in matched conditions of the baseline system

6.2 Experimental results of different-order Taylor series approximation

The log-spectral compensation approaches operate on the non-truncated 20-dimension log-spectral vectors, from which the 12-d basic MFCC (c1-c12) is obtained after DCT. For all the log-spectral compensation approaches evaluated here, a GMM consisting of 256 components, trained on a subset of the training data, is employed. The power feature is processed as Eq. (17) shows.

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>17.23</td>
<td>62.79</td>
</tr>
<tr>
<td>10dB</td>
<td>79.67</td>
<td>54.07</td>
</tr>
<tr>
<td>20dB</td>
<td>80.63</td>
<td>85.63</td>
</tr>
</tbody>
</table>

Table 3. Recognition accuracy (%) of different-order Taylor series approximation with the acoustic model trained using clean data

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>9.86</td>
<td>50.62</td>
</tr>
<tr>
<td>10dB</td>
<td>76.19</td>
<td>63.74</td>
</tr>
<tr>
<td>20dB</td>
<td>79.54</td>
<td>85.53</td>
</tr>
<tr>
<td>0dB</td>
<td>9.62</td>
<td>54.47</td>
</tr>
<tr>
<td>10dB</td>
<td>78.52</td>
<td>67.57</td>
</tr>
<tr>
<td>20dB</td>
<td>81.72</td>
<td>85.56</td>
</tr>
<tr>
<td>0dB</td>
<td>9.87</td>
<td>52.72</td>
</tr>
<tr>
<td>10dB</td>
<td>78.62</td>
<td>66.71</td>
</tr>
<tr>
<td>20dB</td>
<td>80.24</td>
<td>86.07</td>
</tr>
<tr>
<td>0dB</td>
<td>10.51</td>
<td>53.64</td>
</tr>
<tr>
<td>10dB</td>
<td>78.13</td>
<td>65.70</td>
</tr>
<tr>
<td>20dB</td>
<td>80.6</td>
<td>86.26</td>
</tr>
<tr>
<td>0dB</td>
<td>10.13</td>
<td>54.26</td>
</tr>
<tr>
<td>10dB</td>
<td>78.17</td>
<td>66.79</td>
</tr>
<tr>
<td>20dB</td>
<td>80.91</td>
<td>85.76</td>
</tr>
</tbody>
</table>

Table 4. Recognition accuracy (%) of different-order Taylor series approximation with the acoustic model trained using multi-condition data

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB</td>
<td>19.15</td>
<td>61.94</td>
</tr>
<tr>
<td>10dB</td>
<td>78.63</td>
<td>60.61</td>
</tr>
<tr>
<td>20dB</td>
<td>80.61</td>
<td>85.28</td>
</tr>
<tr>
<td>0dB</td>
<td>23.34</td>
<td>64.82</td>
</tr>
<tr>
<td>10dB</td>
<td>78.81</td>
<td>69.99</td>
</tr>
<tr>
<td>20dB</td>
<td>80.93</td>
<td>84.90</td>
</tr>
<tr>
<td>0dB</td>
<td>21.90</td>
<td>64.06</td>
</tr>
<tr>
<td>10dB</td>
<td>78.99</td>
<td>70.70</td>
</tr>
<tr>
<td>20dB</td>
<td>81.45</td>
<td>85.02</td>
</tr>
<tr>
<td>0dB</td>
<td>22.13</td>
<td>64.69</td>
</tr>
<tr>
<td>10dB</td>
<td>79.11</td>
<td>70.65</td>
</tr>
<tr>
<td>20dB</td>
<td>81.40</td>
<td>84.58</td>
</tr>
</tbody>
</table>

Table 5. Recognition accuracy (%) of different-order Taylor series approximation with the acoustic model trained using multi-condition data

7. CONCLUSION

In this paper, the structure-based feature compensation is explored and desirable experimental results are obtained. The novelty of the paper lies in the introduction of high-order Vector Taylor Series expansion in the log-spectral compensation to discover the best compensation formulas, then the power feature and the construction of speech log-spectral GMM are discussed and processed. The experimental results show that the approach can achieve better recognition performance than that in matched conditions.

8. REFERENCES