Robust MMSE-FW-LA ASR Scheme at low SNRs

Tao Xu and Zhigang Cao
State Key Lab on Microwave & Digital Communications, Dept. of Electronic Engineering,
Tsinghua University, Beijing, 100084, China
xutao@sat.mdc.tsinghua.edu.cn

ABSTRACT

In this paper, a novel feature weight (FW) algorithm for robust automatic speech recognition (ASR) is proposed. In this algorithm every feature will be weighted according to their credible probability, especially, the weight factors are formulated and obtained from the gain coefficients generated as by-products of speech enhancement based on minimum mean square error (MMSE) estimation. Moreover a new robust ASR scheme is presented. In this scheme the MMSE-based speech enhancement, the FW algorithm and the Log-Add (LA) model compensation will be integrated together. Experimental evaluations show that this MMSE-FW-LA scheme can achieve significant improvement in recognition across a wide range of signal-to-noise ratios (SNR), especially in very low SNR conditions.

1. INTRODUCTION

Speech-recognition systems are very sensitive to the mismatch between the testing and training conditions. The mismatch, which causes performance deterioration of the speech recognition, lies in three aspects viz. time domain, feature domain and model domain. Consequently, there are three kinds of approaches for improving the recognition performance in noisy environment.

In the signal domain, speech enhancement acting as a front-end processor can improve recognition accuracy measurably [1]. However, in the low SNR conditions the performance still degrades dramatically due to the restriction of the algorithm itself.

In the feature domain, missing data approach and its improved version were proposed [2][3]. This approach takes advantage of the fact that, since the signal energy of speech and noise are different in different frequency bands, noise affects different regions of a spectrographic representation of speech differently. In this paper, we propose a novel feature weight (FW) algorithm, in which a credible probability as weight will be given to every feature and these distorted features are treated differently according to their credible probability during the pattern match. We will solve two problems in this robust algorithm: how to get the credible probability or weight, and how to embed it into the HMM-based recognition system.

The paper is organized as follows. In section 5, we present the weight computation formula. In section 3, the FW algorithm is embedded into the HMM-based recognition system using the Mel Frequency Cepstral Coefficient (MFCC). In section 4, MMSE-FW-LA scheme is presented. Some ASR experiment results are given in section 5. Section 6 is the conclusion.

2. CREDIBLE PROBABILITY MEASUREMENT

In order to estimate the feature’s credible probability, we utilize the accessorial information acquired from the MMSE based speech enhancement [5]. Let $A_k(n)$, $R_k(n)$ denote the $k$th spectral amplitude component of the $n$ th frame clean and noisy speech respectively. In MMSE-based speech enhancement, the estimate of $A_k(n)$ can be given as :

$$\hat{A}_k(n) = G_k(n) \cdot R_k(n)$$

(1)

$$G_k(n) = \frac{1}{2} \sqrt{\frac{\pi}{\gamma_k}} M^{-0.5} \cdot \frac{1 - \gamma_k \zeta_k}{1 + \zeta_k}$$

(2)

where the $G_k(n)$ is called gain coefficient and the $\zeta_k$, $\gamma_k$ are
interpreted as a priori SNR and a posteriori SNR of the $k$th spectral component of current frame respectively. $M(a; c; x)$ is the hypergeometric confluent function. As the higher the priori and posteriori SNR, the bigger the gain coefficient and vice versa, we can take the gain coefficient as some measurement of the local SNR of every time-frequency component.

In the log-spectra domain, the credible probability of every feature is connected with the local SNR of some time-frequency span directly, while in the cepstral domain, the orthogonal transform, such as DCT, will spread credibility measurement over all features. So the feature weigh should be performed in the log spectral domain. As the weight factors are determined by the local SNR of the time-frequency regions where the log-spectra features are extracted, they can be calculated following the procedure of the log-spectra feature extraction and formulized as:

$$w_{mn}(n) = \sum_{k=0}^{H-1} G_{mn}(n) H_m(k) / \sum_{k=0}^{H-1} H_m(k)$$ (3)

then normalized such that

$$\sum_{m=0}^{M-1} w_{mn}(n) = 1$$ (4)

where the $G_{mn}(n)$ is the gain coefficient (2), and can be obtained in the speech enhancement. $H_m(k)$ is the coefficient of the $m$th triangle filter at the $k$th frequency component in the power spectrum domain, $M$ is the number of the filters, $N$ is the size of FFT and $n$ denotes the $n$th frame.

![Figure 1: log-spectra feature distortion and weight factors of one voiced speech frame against white noise (SNR=0dB)](image)

In Figure 1, we can observe that the weight factors have the similar shape with the log-spectra feature vector extracted from the clean speech and the two peak values represent the first and the second formants. The higher the distortion, the lower the weight factors, and vice versa. Thus the weight can extrude some low-distorted log-spectra features, especially, the discriminative information of the first two formants, which is of great importance in the speech comprehension. In the silent and unvoiced speech segment, the weight factors obtained by this method are inconsistent with the distortion of log-spectra feature obviously. In this paper, the weight factors will be deemed to ones at the silent segment compulsively.

3. FEATURE WEIGHT ALGORITHM

Since corruption caused by noise will be smeared by some orthogonal transforms, such as DCT, algorithms utilizing the missing data method is usually restricted in the log-spectral domain. However, this is a serious drawback because the recognition accuracy using log-spectra is much worse than that obtained using the cepstra features. Meanwhile, the log-spectra does not have the near independence of the cepstra features on which most traditional ASR systems are based. This means that diagonal Gaussian component is not itself good approximation to the feature’s distribution, and a greater number of components are required to produce proper acoustic models. In paper [6], a new technique has been proposed for performing the missing data operation in the frequency domain while utilizing standard Mel-cepstra features. We will adopt the technique in this paper as follows.

The Viterbi recognition algorithm essentially is to find the maximum likelihood state sequence (the direct computation of likelihoods leads to underflow, hence, the log likelihood is used instead), so it is critical to find the robust distance measures against noise disturbances for reliable recognition of noisy speech. The log-likelihood (Mahalanobis distance) of the input feature vector $f^c$ given the gaussian mixture component parameter is calculated as:

$$D = c(\Sigma) - \frac{1}{2}(f^c - u)^T \Sigma^{-1} (f^c - u)$$ (5)

where $c(\Sigma)$ represents the constant term of the log-likelihood estimation formula. $u$ and $\Sigma$ denote the mean vector and diagonal covariance matrix of gaussian distribution respectively (since the near independence of cepstra features). The superscript $c$ and $l$ (in the following formula) denote the cepstral and log-spectral domain respectively.

The robust distance measure can be implemented by the addition of a weight matrix $W = \text{diag} \{ w_1, w_2, \ldots \}$, and each element $w_i$ denotes the weight of each log-spectra feature. The feature weight formula in log-spectral domain can be expressed as:

$$D = c(\Sigma^l) - \frac{1}{2}(f^l - u^l)^T W^T \Sigma^{-1} W (f^l - u^l)$$ (6)

In order to weight the log-spectra while using the cepstra feature, a simple modification of (5) will be made and the formula (7) is obtained:
where $d^l = C^{-1}(f^l - u^l)$ and $C$ denotes the DCT matrix. That is, the cepstra difference vector is first reverted to the log-spectral domain, where it is weighted and then transformed back to the cepstral domain.

4. **MMSE-FW-LA Scheme**

In [7], the MMSE-LA scheme was proposed, in which the residual noise after MMSE-based speech enhancement can be regarded as additive stationary noise approximately and LA algorithm compensated the model mismatch caused by these residual noise. Based on this work, an improved MMSE-FW-LA scheme is presented. Figure 2 shows its flow chart:

![Flow Chart of the MFCC-FW-LA scheme](image)

First, MMSE-based speech enhancement is used to estimate the short time spectra amplitude (STSA) of clean speech and suppress the corrupted additive noise. Meanwhile the voice activity detection (VAD) result is reserved. Secondly, MFCC features are extracted from those enhanced speech. Thirdly, the residual noise is modeled with a single gaussian mixture state. As only the mean of model is required, we estimate it by calculating the mean of all MFCC feature vectors taken from silent segments. Fourthly, a residual noise compensated speech model is acquired using the LA algorithm [3]. Finally, the compensated speech model, MFCC features of enhanced speech and weight factors of log-Spectra features enter a Viterbi decoder adapted for FW algorithm and recognition results are acquired.

5. **Experiment Results**

Speaker-Independent TI-Digits recognition experiments are carried out with a FW adapted Viterbi recognizer to evaluate the MMSE-FW-LA scheme. The contaminated speech for test is generated by artificially adding different levels of noise to the clean speech. All noise signals come from a Noisex-92 database. The model we used is continuous density HMM (CDHMM) with left-to-right structure. 500 connected digits utterances from 15 speakers and 100 connected digits utterances from 4 speakers unseen in the training set are used for training and testing respectively. The features are 13-dimension static MFCC features with their delta parameters. Only static features are weighted.

A. Comparison of the FW algorithm with MMSE and LA algorithm

In Figure 3, the baseline performance, the performance using FW, MMSE and LA algorithm respectively are shown as a function of SNR with added white noise. As can been seen, all these algorithms improve recognition accuracy, while at the low SNRs (below 10dB), the FW algorithm improves the recognition accuracy more remarkably, and at the high SNRs the LA algorithm afford more significant performance amelioration.

![Word recognition rates against white noise of Baseline, using MMSE, FW and LA algorithms separately at a range of SNRs](image)

B. Combining FW with MMSE and LA algorithm.

Figure 4 shows the recognition performance combining FW algorithm with the front-end MMSE-based speech enhancement and LA model compensation algorithm respectively. As a fact, the three algorithms attempt to eliminate the mismatch in the signal, feature and model domains differently, their combination can offer further improvement of...
recognition performance.

C. Comparison of the MMSE-LA and MMSE-FW-LA scheme

Figure 5 shows that the MMSE-LA and MMSE-FW-LA scheme all improve the recognition accuracy over LA algorithm dramatically. However, at low SNRs (below 5dB), the MMSE-FW-LA scheme is superior to the MMSE-LA scheme, and the lower the SNR, the more remarkable the superiority.

Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>-5</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6</td>
<td>8</td>
<td>13</td>
<td>30</td>
<td>65.33</td>
</tr>
<tr>
<td>MMSE</td>
<td>14.67</td>
<td>24</td>
<td>54</td>
<td>80.33</td>
<td>91</td>
</tr>
<tr>
<td>FW</td>
<td>24.33</td>
<td>46.33</td>
<td>65</td>
<td>76.67</td>
<td>82</td>
</tr>
<tr>
<td>LA</td>
<td>22.33</td>
<td>34.33</td>
<td>67.67</td>
<td>84.67</td>
<td>92.67</td>
</tr>
<tr>
<td>MMSE-FW</td>
<td>46.33</td>
<td>76</td>
<td>85</td>
<td>91.67</td>
<td>92.67</td>
</tr>
<tr>
<td>FW-LA</td>
<td>32</td>
<td>57.33</td>
<td>77.67</td>
<td>87.33</td>
<td>94.67</td>
</tr>
<tr>
<td>MMSE-LA</td>
<td>65.33</td>
<td>79.67</td>
<td>89.33</td>
<td>93</td>
<td>93.33</td>
</tr>
<tr>
<td>MMSE-FW-LA</td>
<td>81</td>
<td>86</td>
<td>89</td>
<td>94.33</td>
<td>94</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

In this paper we first propose a novel feature weight algorithm and a special approach to calculate feature weight, which makes use of the gain coefficients obtained as by-products from front-end speech enhancement. This algorithm has been shown to afford a significant performance improvement over the front-end MMSE-based speech enhancement and LA model compensation algorithm at low SNRs when tested on a connected-digit recognition task.

Then the low algorithm complexity MMSE-FW-LA scheme for robust ASR is presented, which reduces the mismatches in the signal, feature and model domain simultaneously and achieve significant recognition improvement across a wide range of SNRs greatly, especially at low SNRs.

7.REFERENCE