Sub-Space Missing Feature Imputation and Environment Sniffing for Robust Speech Recognition

Yuan-Fu Liao

Abstract
Noise robustness is the most important issue for real-life speech recognition/spoken dialogue system. In this paper, a sub-space missing feature theory (S-MFT) front-end is proposed to alleviate the corruption of the background noise. S-MFT incorporates temporal information and applies principle component analysis (PCA) to find a noise-suppressed sub-space for more precise missing feature imputation. Moreover, two parameters including signal-to-noise ratio (SNR) and divergence-based reliability measure of an input utterance generated by S-MFT are employed to judge whether the utterance can be successfully recognized or not. Experiments on TIDIGIT and three kinds of noises from NOISEX-92 corpora have verified the benefits of the proposed S-MFT approach. The average recognition rate could be improved from 61.6% (cepstral mean and variance normalization, CN) to 77.1% and further to 89.9% while rejecting bad utterances.

1. Introduction
The sensitiveness of a speech recognizer to the corruption of background noise is one of the most important issues that prohibit the widespread use of speech recognition/spoken dialogue techniques. There are many successful approaches to make speech recognizers more robust to background noise. Most of them can be divided into three categories including feature-, model- and score-domain approaches. Among them, the feature-domain, especially, the feature enhancement/normalization-based approaches are attractive recently. Because they aim to directly enhance the input speech feature vectors in the front-end, the requirement to modify the backend recognizer is relaxed.

There are several feature enhancement/normalization-based noise robust approaches including cepstral mean and variance normalization (CN) [1], long-term average method (LTA) [2], Wiener filter (WF) [3] and missing feature theory (MFT) [4]. CN and LTA blindly compensate the corruption of background noise. On the other hand, WF and MFT rely on a voice activity detector (VAD) [4] to separate the feature vectors of an input utterance into speech and noise segments in order to capture the characteristics of the background noise. Among them, the MFT-based speech enhancement front-ends may have some advantages, since it not only collects the underlying noise characteristics online but also utilizes a speech model as the a priori knowledge. Therefore, it may more precisely enhance the input speech feature vectors.

However, in the unreliable part (usually lower SNR) of the input speech signal, background noise may dominate and mislead the imputation procedure of MFT approach. Therefore, in this paper, a sub-space-based MFT (S-MFT) front-end is proposed to improve the performance of the conventional MFT approach. In brief, S-MFT incorporates temporal information by exploring the relationship of neighboring frames and applies principle component analysis (PCA) [5] to generate a noise suppressed sub-space to filter out background noise before applying the Gaussian mixture model (GMM)-based missing feature imputation [6]. The probabilities of speech/non-speech frames as a byproduct of S-MFT are also employed to constrain the backend Viterbi search to improve the recognition performance.

Moreover, two extra measurements generated by S-MFT are also used to sniff the underlying recognition environment in order to find a suitable way to deal with different situations. Specifically, the utterance-level SNR and divergence-based reliability measures between the input feature vectors and the underlying GMMs are used to judge whether an input utterance can be successfully recognized or not. If it is not possible due to background noise or speaking variance, the user will be informed with reasons and will be asked to help to change the situation before invoking the computation-intensive backend recognizer.

It is worthy noting that S-MFT is totally operated in the front-end and is not computation intensive. On the other hand, conventional confidence measure-based approaches reject bad utterances without giving reasons and are calculated in the backend decoder with high computation complexity.

This paper is organized as follows. Section 2 introduces the proposed S-MFT approach. Bad utterance detection and modified Viterbi search are described in Section 3. Section 4 reports the experimental results on the well-known TIDIGIT [7] and NOISEX-92 [8] databases. Some conclusions and future works are given in the last section.

2. Sub-space missing feature imputation
Recently, missing feature theory has been proposed to help speech recognition while the incoming noisy speech is partially corrupted by the background noise [9]. The basic principle is to first estimate the characteristics of the background noise in order to classify the input speech features along the time and frequency-band into reliable and unreliable (missing) parts. Then spectral subtraction (SS) and GMM-based imputation are applied to the reliable and unreliable parts, respectively, to remove noise components or to recover the missing features.

In order to improve the performance of the GMM-based missing feature imputation, S-MFT, as shown in Fig. 1, is proposed. It incorporates temporal information and includes two PCA-based temporal filters for noise-suppressed sub-space
projection, a synthesized sub-space noisy speech GMM, a sub-

space imputation weight estimator and a GMM-based missing

feature imputation approach. The detail procedures of the S-

MFT are described in the following subsections.

2.1. GMM-based imputation

Once the unreliable parts of input feature vectors are detected and the background noise model is estimated, the GMM-based imputation approach is adopted to recover the missing features. Assume that filter-bank features (log-magnitude) are used; the procedures of the GMM-based imputation are described as follows:

(i) Transform the $d$-th components of mean and variance vectors $\{u(m,d), \sigma^2(m,d)\}$ of the $m$-th mixture of the pre-trained clean-speech GMM into linear domain by using the inverse log-normal approximation:

$$
\begin{align*}
\hat{u}_{in}(m,d) &= \exp\left(u(m,d) + \frac{\sigma^2(m,d)}{2}\right) \\
\hat{\sigma}^2_{in}(m,d) &= \hat{u}_{in}^2(m,d)\left(\exp(\sigma^2(m,d)) - 1\right)
\end{align*}
$$

(ii) Combine the clean-speech GMM and the noise model $\lambda_n = \{\hat{u}_{in}(d), \hat{\sigma}^2_{in}(d)\}$, which is estimated from the unreliable parts of input features, in the linear domain to become the noisy-speech GMM, $\hat{\lambda}_n = \{\hat{u}_{in}(m,d), \hat{\sigma}^2_{in}(m,d)\}$, by

$$
\begin{align*}
\hat{u}_{in}(m,d) &= u_{in}(m,d) + u_d(d) \\
\hat{\sigma}^2_{in}(m,d) &= \sigma^2_{in}(m,d) + \sigma^2_d(d)
\end{align*}
$$

(iii) Transform the noisy speech GMM back to log-spectral domain, i.e., $\hat{\lambda}_n = \{\hat{u}(m,d), \hat{\sigma}^2(m,d)\}$, by

$$
\hat{u}(m,d) = \log\left(\frac{\hat{u}_{in}^2(m,d)}{\hat{u}_{in}^2(m,d) + \hat{\sigma}^2_{in}(m,d)}\right)
$$

(iv) For a sequence of input speech feature vectors $X = \{x(t,d), t=0 \sim T-1, d=0 \sim D-1\}$, the imputation weights $\gamma(m|x(t), \hat{\lambda})$ associated with the $m$-th mixture is computed by

$$
\gamma(m|x(t), \hat{\lambda}) = \frac{p(m)\prod_{d=0}^{D-1} \Phi\{x(t,d), \hat{u}(m,d), \hat{\sigma}^2(m,d)\}}{p(x(t)|\hat{\lambda})}
$$

(v) The missing features $\hat{x}_n(t,d)$ are finally replaced by a weighted sum of the mean vectors of the corresponding clean-speech GMM by

$$
\hat{x}_n(t,d) = \sum_{m=0}^{M} \gamma(m|x(t), \hat{\lambda}) \cdot u(m,d)
$$

Therefore, how to precisely estimate the imputation weights, $\gamma(m|x(t), \hat{\lambda})$, are the most critical issue of the GMM-based imputation approach. In the following sub-sections, PCA temporal filtering and sub-space GMM imputation are employed to increase the accuracy.

2.2. PCA-based noise suppressed temporal filtering

To suppress the dominant background noise in the unreliable parts of feature vectors before estimating the imputation weights, PCA is used to first learn from the feature vectors of clean speech to generate a noise suppressed sub-space for each component of feature vectors. Then the input noisy speech feature vectors are projected into the sub-space by the PCA-generated temporal filters to suppress the background noise. Therefore, a set of $(2M+1)$-point FIR-filters have to be designed.

Given a sequence of $D$-dimensional noisy feature vectors with length $T$, $X = \{x(t,d), t=0 \sim T-1, d=0 \sim D-1\}$, the $d$-th dimension time trajectory of $X$ is defined as $X(d) = \{x(t,d), t=0 \sim T-1\}$. Then, an $(2M+1)$-point rectangular window is shifted along the $d$-th dimension time trajectory $X(d)$ to obtain a sequence of $(2M+1)$-dimensional vectors $Y(d)$:

$$
Y(d) = \{x(t-M,d),\ldots,x(t,d),\ldots,x(t+M,d)\}, t=0 \sim T-1
$$

The $(2M+1) \times (2M+1)$ dimensional covariance matrix of $Y(d)$, i.e., $R(d)$, is established and analyzed by PCA to find the corresponding eigen-values and vectors. Then the largest $K$ eigen-vectors are combined and used as the coefficients of the temporal filter of the $d$-th dimension time trajectory as follows:

$$
\begin{align*}
\Phi_d &= \sum_{\kappa=1}^{K} \lambda_{\kappa,d} \varphi_{\kappa,d} \\
\lambda_{\kappa,d} &= \frac{1}{\sqrt{\sum_{\kappa=1}^{K} \lambda^2_{\kappa,d}}} \Phi_d
\end{align*}
$$

where $\lambda_{\kappa,d}$ and $\varphi_{\kappa,d}$ are the $\kappa$-th eigen-value and eigen-vector of the $d$-th dimension time trajectory.
Finally, the noise suppressed noisy speech feature vectors become
\[
Z(d) = \left\{ z(t,d) = \sum_{i=0}^{M} w_i(t)x(t+i,d), t = 0 ~ T - 1 \right\}
\]
where \( \{w_{-M}(t), \ldots, w_{M}(t)\} \) are the coefficients of the combined PCA-driven temporal filter.

2.3. Sub-space GMM-based imputation

As shown in Fig. 1, three different GMMs are required for sub-space GMM-based imputation. They include a noise GMM, a synthesized sub-space noisy speech GMM and a clean-speech GMM.

By incorporating temporal information, the feature vectors from clean speech are extended to contain the current frame and several neighboring frames. In brief, given a sequence of D-dimensional clean-speech feature vectors with length \( T \), \( X_c = \{x_c(t,d), t = 0 ~ T - 1, d = 0 ~ D - 1\} \); a \((2M+1)\)-point rectangular window is shifted along the time trajectory of \( X_c \) to obtain a sequence of \((2M+1)\)-dimensional vectors \( \hat{Y}_c \) (similar to Eq. (7)). The extended clean-speech feature vectors \( \hat{Y}_c \) are used to train the clean-speech GMM \( \Lambda_c = [U_c, \Sigma_c] \). Besides, the feature vectors of the background noise are processed online in the same way to generate the noise model \( \Lambda_n = [U_n, \Sigma_n] \). These two GMMs are then combined using Eq. (11) to synthesize the extended noisy speech GMM \( \hat{\Lambda} = [U, \hat{\Sigma}] \).

The noisy speech GMM is then projected into the noise suppressed sub-space using the temporal filters previously generated by the PCA method. Finally, the projected extended noisy speech GMM \( \hat{\Lambda}_{PCA} = [U_{PCA}, \hat{\Sigma}_{PCA}] \) is used to compute the imputation weights for a filtered input noisy utterance \( Z(d) = \{ z(t,d), t = 0 ~ T - 1 \} \) by
\[
\gamma(m|z(t), \hat{\Lambda}_{PCA}) = \frac{p(m)\prod_{d=0}^{T-1}q(z(t,d), \hat{U}_{PCA}(m,d), \hat{\Sigma}_{PCA}(m,d))}{p(z(t)|\hat{\Lambda}_{PCA})}
\]
(10)

Since the one-to-one linkage between the mixtures of the sub-space noisy speech and the clean-speech GMM are carefully maintained, the following imputation formulation can be used to recover the missing features:
\[
\hat{x}_a(t,d) = \sum_{n=0}^{M-1} \gamma(m|z(t), \hat{\Lambda}_{PCA}) \cdot U(m,d)
\] (11)

3. S-MFT-based environment sniffing

The block diagram of the S-MFT-based environment sniffing is shown in Fig. 2. It includes a bad utterance detection front-end and a modified backend Viterbi decoder. The detection is relied on two features generated by MFT including the utterance-level SNR for detecting background noise and divergence-based reliability for measuring speaking variance between input feature vectors and the underlying clean speech GMMs.

3.1. Bad utterance detection

To judge whether an input utterance could be successfully recognized by the backend recognizer or not, two parameters, including the utterance-level SNR and a divergence-based reliability measure, are calculated using S-MFT using GMMs. The utterance-level SNR and the reliability measures are aimed to measure the level of background noise and the degree of speaking variance comparing with the underlying GMMs.

First, the frame-level SNR of an input utterance could be evaluated as
\[
SNR(t) = \frac{\sum_{d=0}^{D-1} \hat{x}(t,d)}{\sum_{d=0}^{D-1} u_a(d)}
\]
t where \( \hat{x}(t,d) \) is the enhanced filter-band magnitude feature vector, \( u_a(d) \) is the estimated noise spectrum.

Secondly, the divergence between the probability density function (PDF) of the input feature vectors \( p = \mathcal{N}(u, \Sigma) \) (assume Gaussian PDF) and the corresponding PDF of an acoustic model \( q = \mathcal{N}(u, \Sigma) \) is calculated and converted into a zero-one function as follows:
\[
Div(p \parallel q) = \frac{1}{2} \left( \frac{1}{\max((1 \div \alpha)Div + \beta)} \right)
\]
(13)

Lastly, to decide whether the input utterance should be accepted or rejected, the peak frame-level SNR and the reliability measure are linearly combined and compared with a threshold \( \theta \) by:
\[
a \cdot \arg\max_r SNR(t) + b \cdot \arg\max_r R(t) \geq \theta \rightarrow \text{accept}
\]
\[
a \cdot \arg\max_r SNR(t) + b \cdot \arg\max_r R(t) < \theta \rightarrow \text{reject}
\]
(15)

3.2. Modified Viterbi search

The frame-level SNR defined in Eq. (12) carries the speech/non-speech segmentation information. We therefore normalize and convert the frame-level SNR into the probability of speech, \( P_s(t) \), by
\[
P_s(t) = \frac{\max_r \{ SNR(t)-\min_r(SNR(t)) \}}{\max_r \{ SNR(t)-\min_r(SNR(t)) \}}
\]
(16)

This probability function is used to constrain the Viterbi search module of the backend speech recognizer by
\[
\begin{cases}
S_{sp}(t) + \alpha \cdot \log\{P_s(t)\} & \text{for speech models} \\
S_{sil}(t) + \alpha \cdot \log\{1 - P_s(t)\} & \text{for silence model}
\end{cases}
\]
(17)

where \( S_{sp}(t) \) and \( S_{sil}(t) \) are, respectively, the log-likelihoods of the underlying speech and silence models of the \( t \)-th frame,
4. Experiments

4.1. Databases and experiment conditions

In this study, 8-KHz down-sampled TIDIGIT and NOISEX-92 corpora were used to evaluate the proposed S-MFT approach. From TIDIGIT, 112/26 speakers, 8,700/2,000 utterances were chosen as the training/test set respectively. Moreover, three kinds of noise types including babble, factory and white from NOISEX-92 with different SNRs (0~20dB) were added to all test utterances to expand the test set to 30,000 utterances. There were in total 28,583 training (clean speech) and 98,625 test digits (noisy speech).

In all MFT experiments, 24 dimensional filterbank features and a 512-mixture GMM were used for feature enhancement. Beside, partitioned hidden Markov models (HMMs) were hired as the digit recognizer. 38-dimensional mel-scale cepstral coefficients (MFCC) feature vectors were used, including 12-dimensional MFCCs, their delta- and delta-delta-log-energies. Moreover, CN was applied to all approaches. Each digit HMM model had 12 4-mixture states. Besides, a 64-mixture silence model was also used in the system.

4.2. Experimental results

First of all, the conventional CN method was evaluated and treated as the baseline. The results by removing the MFCC biases and normalizing the MFCC variances are listed in Table 1. From Table 1, it can be seen that the recognition rates degraded quickly as the SNR decreased. Secondly, the popular LTA approach was applied to remove the long-term biases of filterbank (mel-scale) features before applying CN and the results are shown in Table 1. Can be seen from the table, although LTA is a blind compensation approach, it did greatly improve the average recognition rate from 61.6% to 72.3%.

The MFT and the proposed S-MFT were then evaluated. The length of the PCA-based temporal filter was empirically set to three. First, the temporal filters generated by PCA are shown in Fig. 3. From the figure, the temporal filters are all low-pass filter and their cut-off frequencies (-3dB) are around 18.6Hz. Secondary, the spectra (mel-scale filterbank) of an input noisy utterance (5dB SNR) and its corresponding LTA, MFT- and S-MFT-enhanced speech are plotted in Fig. 4. From the figure, it shows that LTA, MFT and S-MFT all could recover the speech signal to certain degree. However, MFT and S-MFT were preferred than LTA. It is also worthy noting that the recovered speech spectrum of the S-MFT is smoother than conventional MFT. Finally, the results in Table 1 show that MFT and S-MFT could further increase the average recognition rate from 72.3% to 75.2% and 76.2%, respectively. Therefore, both MFT and S-MFT are promising approaches.

Moreover, the probability of speech/non-speech in Eq. (16) was added to constrain the backend Viterbi search using Eq. (17). The weighting constant $\alpha$ was empirically set to 2.0 and an average recognition rate of 77.1% was achieved (see Table 1, S-MFT+MV). These results show the usefulness of the S-MFT generated speech/non-speech segmentation information.

Finally, the bad utterance detection scheme as shown in Fig. 2 was evaluated using the S-MFT+MV approach. According to Eq. (15), two criteria including utterance SNR and recognition rate were tested and different thresholds were set to detect bad utterances. In brief, the linear regression coefficients in Eq. (15) were empirically chosen and an input test utterance was classified as bad utterance if its estimate of SNR is lower than 10dB or its estimate of digit recognition rate is lower than 80%.

The equal error ratios (EERs) of bad utterance detection using different criteria and features were listed in Table 2, their corresponding receiver operating characteristic (ROC) figures were plotted in Fig. 5 and 6 according to the SNR (>10 dB) and the recognition rate criterion (>80%), respectively.

From Table 2 and Fig. 5, an equal error ratio (EER) of 8.1% could be achieved using the SNR criterion. On the other hand, from Table 2 and Fig. 6, an EER of 27.4% could be achieved according to the recognition rate criterion, respectively. Moreover, if those detected bad utterances were ignored, the average recognition rates could be raised to 88.7% and 89.9%, respectively. Therefore, Tables 1 and 2 show the effectiveness of the proposed S-MFT and bad utterance detection approach.

5. Conclusions

In this paper, a sub-space missing feature theory (S-MFT) front-end is proposed to alleviate the corruption of the background noise and to sniff the underlying recognition environment. Experiments on TIDIGIT and NOISEX-92 corpora have verified the significance of the proposed S-MFT approaches. The average recognition rate can be improved from 61.6% (CN) to 77.1%, and further to 89.9% while rejecting bad utterances. Therefore, the proposed S-MFT is a promising approach and is worthy of further studying.

6. Acknowledgements
This paper is a partial result of the project No. B34BXT5200 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan.

7. References


Table 1. The recognition rates (%) achieved by the CN, LTA, MFT, S-MFT and S-MFT+MV noise enhancement approaches on TIDIGIT and NOISEX-92 corpora, respectively

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>Babble</th>
<th>Factory</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.1</td>
<td>99.2</td>
<td>99.1</td>
<td>99.1</td>
</tr>
<tr>
<td>CN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTA</td>
<td>93.0</td>
<td>95.2</td>
<td>97.1</td>
<td>97.2</td>
</tr>
<tr>
<td>MFT</td>
<td>85.8</td>
<td>90.5</td>
<td>93.1</td>
<td>93.6</td>
</tr>
<tr>
<td>S-MFT</td>
<td>71.7</td>
<td>82.1</td>
<td>84.6</td>
<td>85.4</td>
</tr>
<tr>
<td>S-MFT+MV</td>
<td>47.8</td>
<td>64.0</td>
<td>67.4</td>
<td>69.3</td>
</tr>
<tr>
<td>0dB</td>
<td>19.6</td>
<td>37.6</td>
<td>38.2</td>
<td>40.9</td>
</tr>
<tr>
<td>Ave. of 0-20dB</td>
<td>63.6</td>
<td>73.9</td>
<td>76.1</td>
<td>77.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SNR</th>
<th>SNR+Rel.</th>
<th>Rec. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>Features</td>
<td>EER</td>
<td>Reject</td>
</tr>
<tr>
<td>SNR &gt;10dB</td>
<td>SNR</td>
<td>10.3</td>
<td>42.2</td>
</tr>
<tr>
<td>SNR &gt;10dB</td>
<td>SNR+Rel.</td>
<td>8.1</td>
<td>41.5</td>
</tr>
<tr>
<td>Rec. rate &gt;80%</td>
<td>SNR</td>
<td>27.8</td>
<td>44.1</td>
</tr>
<tr>
<td>Rec. rate &gt;80%</td>
<td>SNR+Rel.</td>
<td>27.4</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Table 2. The average recognition rates (%) achieved by the S-MFT+MV-based bad utterance detection scheme using the SNR or recognition rate criterion on TIDIGIT and NOISEX-92 corpora.

Figure 3. The frequency responses of the set of the 2-order PCA-generated temporal filters.

Figure 5. The ROC curve of the S-MFT+MV-based bad utterance detection scheme using the SNR criterion (> 10dB).
Figure 6. The ROC curve of the S-MFT+MV-based bad utterance detection scheme using the recognition rate criterion (>80%).
Figure 4. Comparison of the enhanced spectra (mel-scale) of an input noisy utterance (5dB SNR) between LTA, MFT and S-MFT-based feature enhancement approaches: (a) the reliable/unreliable masking of an input speech generated by MFT/S-MFT and the spectrum of (b) clean, (c) noisy, (d) LTA-enhanced, (e) MFT-enhanced and (f) S-MFT-enhanced speech, respectively.