COMPENSATION OF PARTLY RELIABLE COMPONENTS FOR BAND-LIMITED SPEECH RECOGNITION WITH MISSING DATA TECHNIQUES

Yongjun He, Jiqing Han, Tieran Zheng, Guibin Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

Mismatch in speech bandwidth between training and real operation greatly degrades the performance of automatic speech recognition (ASR) systems. Missing feature technique (MFT) is effective in handling bandwidth mismatch. However, current MFT-based methods ignore the mismatch in the filterbank channels which cover the upper and lower limit cutoff frequencies. To solve this problem, we propose to partition the feature into reliable, unreliable and partly reliable parts, and then modify the probability density functions (PDFs) of the partly reliable part to match band-limited features. Experiments showed that such compensation further improved the performances of MFT-based methods under band-limited conditions.

Index Terms— Probability density function, missing feature techniques, band-limited speech recognition

1. INTRODUCTION

An ASR system trained with full-bandwidth data suffers a dramatic accuracy degradation when recognizing band-limited data because of bandwidth mismatch. Such mismatch is widely existing in speech recognition under narrowband channel, Internet or distributing conditions. It is also a common thing in speech document retrieval, where audio recordings have different bandwidths [1].

Two category methods, namely feature enhancement and model adaptation, can be considered to deal with bandwidth mismatch. The former attempts to compensate features extracted from the distorted speech to obtain pseudo-undistorted ones (see [1-4]). The latter modifies a system’s models to fit new recognition environments (e.g. [5]).

MFT is reported to be effective in noisy speech recognition[6][7]. It deems low signal-to-noise ratio time-frequency components as unreliable, and performs recognition with the remaining reliable components [6]. MFT is also used for dealing with band-limited distortion by Kim, et al [3][4], who extends a cluster-based missing-feature reconstruction method [7] to make compensation in feature domain.

Current MFT-based methods always assume one component either completely missing or completely reliable in dealing with band-limited distortion. However, the components in the filterbank channels which cover the lower and upper limit cutoff frequencies are partly reliable (or partly missing). As seen from Fig.1, the filterbank distributes in the full-bandwidth $K_O \sim K_E$, where $K_O$ and $K_E$ are both frequency-bin indexes satisfying the constraint $K_O > 0$, $K_E < K$ ($K$ is a half of the number of discrete Fourier transform (DFT) points). The limited-bandwidth is $K_L \sim K_U$. $K_L$ is located in the $I$-th and ($I + 1$)-th filterbank channels; $K_U$ in the $J$-th and ($J + 1$)-th filterbank channels. The four components $x_I, x_{I+1}, x_J$ and $x_{J+1}$ are partly missing. Taking $x_I$ as an example, in the $i$-th filterbank channel, spectrums which frequency-bin indexes are smaller than $K_L$ are lost, while the remaining spectrums are reliable. If such a component is deemed as reliable, mismatch still exists; if it is deemed as unreliable and discarded, more available information will be lost. In real application, the number of filterbank channels is about 30, and no more than 20 components are reliable for band-limited speech. So it is highly necessary to compensate the four components.

In this paper, we firstly partitioned the feature into reliable, unreliable and partly reliable parts, and then derived the relationship between the PDFs of band-limited speech and that of full-bandwidth speech for partly reliable parts, finally resolved their relationship with Monte Carlo method. We integrated our compensation into two versions of MFT, namely, data imputation and marginalisation [6]. Experiments indicated that the proposed method achieve better performances than conventional MFT-based methods.

2. COMPENSATION OF PARTLY RELIABLE COMPONENTS

The speech recognizer is assumed to have an architecture of hidden Markov model (HMM) with Gaussian mixture acoustic models. In the front-end, a Mel-spectral representation is computed by a filterbank with $N$ channels (Fig. 1) through windowing, framing, DFT and filterbank integration. Compared with full-bandwidth speech, band-limited speech has the 1-th to ($I - 1$)-th and ($J + 2$)-th to $N$-th components completely lost, but the $I$-th, ($I + 1$)-th, $J$-th and ($J + 1$)-th components...
th ones partly lost. For a full-bandwidth speech frame which
filterbank magnitude vector is \( \mathbf{X} = [x_1, x_2, \ldots, x_N] \), the corresponding band-limited filterbank magnitude vector is:

\[
\mathbf{Y} = \text{diag}(s) \mathbf{X}
\]

(1)

where \( s \) is a \( N \)-dimensional vector with 0 or 1 indicating the corresponding components are completely missing or completely reliable, respectively, \( \text{diag}(.) \) stands for the diagonal matrix with its diagonal component value equal to the value of the vector in the argument. If the bandwidth is known (given \( I, J \)), \( s \) can be written as

\[
s = [0, \ldots, 0, s(I), s(I + 1), 1, \ldots, 1,
\text{s(J), s(J + 1), 0, \ldots, 0}]
\]

(2)

Current MFT-based methods assume that the four components on \( s(I), s(I + 1), s(J) \) and \( s(J + 1) \) are completely reliable or unreliable under band-limited conditions (namely with values 0 or 1 for the four elements). However, we use continuous values between 0 and 1 here for the four elements. In a partly missing filterbank channel, \( s(n) \) \( (n = I, I + 1, J, J + 1) \) is computed according to (1):

\[
s(n) = \frac{\sum_{k=K_u}^{K_v} u^{(n)}(k) Y(k)}{\sum_{k=K_o}^{K_e} u^{(n)}(k) X(k)}
\]

(3)

where \( u^{(n)}(k) \) is the \( k \)-th coefficient of the \( n \)-th filterbank channel, \( X(k) \) and \( Y(k), \( k = K_o, K_o + 1, \ldots, K_e \), are the power spectrums of a full-bandwidth frame and its corresponding band-limited one, respectively. Suppose a full-bandwidth data vector has been partitioned into reliable, unreliable and partly reliable parts, namely, \( \mathbf{X} = [\mathbf{x}_F^T, \mathbf{x}_U^T, \mathbf{x}_D^T]^T \), the corresponding band-limited data vector \( \mathbf{Y} \) is also partitioned similarly into \( [\mathbf{y}_F^T, \mathbf{y}_U^T, \mathbf{y}_D^T]^T \), where \( T \) is the transpose operator. In MFT, \( \mathbf{y}_U \) is discarded, \( \mathbf{y}_F \) and \( \mathbf{y}_D \) are used for recognition. According to (1), the relationship between \( \mathbf{x}_D \) and \( \mathbf{y}_D \) is written as

\[
\mathbf{y}_D = \text{diag}(\varepsilon) \mathbf{x}_D
\]

(4)

where

\[
\varepsilon = [s(I), s(I + 1), s(J), s(J + 1)]
\]

(5)

We also assume that all dimensions of feature vector are independent to each other [6], and use diagonal matrix for covariance. So the mean and covariance for mixture component \( m \) of state \( i \) are similarly partitioned into

\[
\mu_{m,i} = [\mu_{x,F,m,i}^T, \mu_{x,U,m,i}^T, \mu_{x,D,m,i}^T]^T
\]

(6)

\[
\Sigma_{m,i} = [\Sigma_{x,F,m,i}^T, \Sigma_{x,U,m,i}^T, \Sigma_{x,D,m,i}^T]^T
\]

(7)

where \( \Sigma_{m,i} \) is the vector formed by all diagonal component values of the covariance matrix. The PDFs of the reliable, unreliable and partly reliable parts can be written as:

\[
f(x_r | m, C_i) = N(x_r; \mu_{r,m,i}, \Sigma_{r,m,i})
\]

(8)

\[
f(x_u | m, C_i) = N(x_u; \mu_{u,m,i}, \Sigma_{u,m,i})
\]

(9)

\[
f(x_p | m, C_i) = N(x_p; \mu_{p,m,i}, \Sigma_{p,m,i})
\]

(10)

Taking the mean and covariance of both sides of (4) and assuming \( \mathbf{x}_D \) is independent to \( \text{diag}(\varepsilon) \) yields

\[
\hat{\mu}_{p,m,i} = \text{diag}(E(\varepsilon_{m,i})) \mu_{p,m,i}
\]

(11)

\[
\hat{\Sigma}_{p,m,i} = \text{diag}(E(\varepsilon_{m,i})^2) \Sigma_{p,m,i} + \text{diag}(\text{Var}(\varepsilon_{m,i})) \mu^2_{p,m,i}
\]

(12)

where \( E(.) \) and \( \text{Var}(.) \) are the expectation and variance operators, respectively, \( \varepsilon_{m,i} = [s_{m,i}(I), s_{m,i}(I + 1), s_{m,i}(J), s_{m,i}(J + 1)] \), \( \varepsilon_{m,i} = \frac{[s_{m,i}(I)]^2, [s_{m,i}(I + 1)]^2, [s_{m,i}(J)]^2, [s_{m,i}(J + 1)]^2} {4} \). Readers are referred to the Appendix for the detailed derivation of (12). Then the compensated PDF for partly reliable part is

\[
f(y_p | m, C_i) = N(y_p; \hat{\mu}_{p,m,i}, \hat{\Sigma}_{p,m,i})
\]

(13)

It is not a trivial thing to compute the expectation and covariance of \( \varepsilon_{m,i} \), we adopt Monte Carlo method to achieve this objective (Algorithm 1). The power spectrum of a generating sample is obtained by:

\[
X_t = W^{-1} X_t
\]

(14)

where \( W^{-1} \) is the Pseudo-inverse of Mel filterbank matrix. Although \( W \) is not squared and equation (14) is ill-posed, it can provide a rough estimation of a power spectrum. Experiments show that such processing is enough to obtain more matched PDFs for partly reliable components. In addition, \( L \) is set as 6000 in our experiments, and all \( \hat{\mu}_{p,m,i} \) and \( \hat{\Sigma}_{p,m,i} \) are computed in an off-line manner, which does not influence a real time application.

3. IMPLEMENTATION TO MARGINALISATION AND DATA IMPUTATION

3.1. Implementation to marginalisation

Marginalisation computes the HMM state output probabilities using a reduced distribution based solely on reliable components. It is extended here to address partly reliable components. Considering \( y_r \approx y_r [6] \), the state output probability
Algorithm I

for each $N(x; \mu_{m,i}, \Sigma_{m,i})$

draw $L$ samples $X_l(l = 1, \ldots, L)$ from $N(x; \mu_{m,i}, \Sigma_{m,i})$;

compute power spectrums $X_l(l = 1, \ldots, T)$ according to limited-bandwidth;

compute $\hat{\epsilon}_l$ for each $X_l$ with (3) (5);

compute $E(\hat{\epsilon}_{m,i}), E(\hat{\epsilon}_{m,i}^2)$ and $Var(\hat{\epsilon}_{m,i})$ with $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_L$;

compute $\hat{\beta}_{m,i}, \hat{\Sigma}_{m,i}$ with (11) (12);

endfor

of a state $i$ (denoted by $C_i$) is computed as

$$
\hat{x}_{u,i} = \sum_m P(k|x_r, x_p, C_i) \mu_{u,m,i}
$$

(16)

where $P(k|x_r, x_p, C_i) = \frac{P(k|x_r, x_p, C_i)f(x_r|k, C_i)f(x_p|k, C_i)}{\sum_n P(k|C_i)f(x_r|k, C_i)f(x_p|k, C_i)}$. In addition, the partly reliable parts are reconstructed by:

$$
x_p = diag(e^{-1})y_p
$$

(17)

In the experiment, bounded marginalisation and bounded data imputation were used. Readers are suggested to refer to [6] for more details.

### 3.2. Implementation to data imputation

Data imputation attempts to reconstruct the missing components using reliable components. In our methods, the matched PDFs of partly missing components are used. According to [6], the missing components can be reconstructed by:

$$
\hat{x}_{u,i} = \sum_m P(k|x_r, x_p, C_i) \mu_{u,m,i}
$$

(16)

where $P(k|x_r, x_p, C_i) = \frac{P(k|x_r, x_p, C_i)f(x_r|k, C_i)f(x_p|k, C_i)}{\sum_n P(k|C_i)f(x_r|k, C_i)f(x_p|k, C_i)}$. In addition, the partly reliable parts are reconstructed by:

$$
x_p = diag(e^{-1})y_p
$$

(17)

In the experiment, bounded marginalisation and bounded data imputation were used. Readers are suggested to refer to [6] for more details.

### 4. EXPERIMENTS AND ANALYSIS

TIMIT and NTIMIT databases were chosen to evaluate the proposed method. TIMIT contained a total of 6300 sentences, of which 4620 utterances spoken by 462 speakers were used for training, and 1680 utterances spoken by another 168 speakers were used for testing. The NTIMIT database was collected by transmitting all TIMIT utterances through various telephone channels and re-digitizing them. The bandlimited speech samples were generated by low-pass and bandpass filtering the original clean speech in the TIMIT database with 6.0 kHz as the full-bandwidth frequency. Six kinds of

of a state $i$ (denoted by $C_i$) is computed as

$$
f(y_r, y_p|C_i) \approx \int f(x_r|x_u, y_p|C_i)dx_u
$$

$$
= \sum_m P(m|C_i)f(x_r|m, C_i)f(y_p|m, C_i)
$$

$$
\times \int f(x_u|m, i)dx_u
$$

(15)

According to (8), (9) and (13), the output probability of the HMM state can be computed.

### 3.2. Implementation to data imputation

Data imputation attempts to reconstruct the missing components using reliable components. In our methods, the matched PDFs of partly missing components are used. According to [6], the missing components can be reconstructed by:

$$
\hat{x}_{u,i} = \sum_m P(k|x_r, x_p, C_i) \mu_{u,m,i}
$$

(16)

where $P(k|x_r, x_p, C_i) = \frac{P(k|x_r, x_p, C_i)f(x_r|k, C_i)f(x_p|k, C_i)}{\sum_n P(k|C_i)f(x_r|k, C_i)f(x_p|k, C_i)}$. In addition, the partly reliable parts are reconstructed by:

$$
x_p = diag(e^{-1})y_p
$$

(17)

In the experiment, bounded marginalisation and bounded data imputation were used. Readers are suggested to refer to [6] for more details.

### 4. EXPERIMENTS AND ANALYSIS

TIMIT and NTIMIT databases were chosen to evaluate the proposed method. TIMIT contained a total of 6300 sentences, of which 4620 utterances spoken by 462 speakers were used for training, and 1680 utterances spoken by another 168 speakers were used for testing. The NTIMIT database was collected by transmitting all TIMIT utterances through various telephone channels and re-digitizing them. The bandlimited speech samples were generated by low-pass and bandpass filtering the original clean speech in the TIMIT database with 6.0 kHz as the full-bandwidth frequency. Six kinds of

low-pass filters (35th-order ChebyshevTypeII filter), with 3.0, 3.5, 4.0, 4.5, 5.0 and 5.5 kHz as their cutoff frequencies, respectively, were used to generate low-pass databases. Another six kinds of band-pass filters with 0.5~6.0 kHz, 1.0~5.5 kHz, 1.5~5.0 kHz, 2.0~4.5 kHz, 2.0~4.0 kHz as their bandwidths, respectively, were used to generate band-pass databases. Therefore, the size of each train/test set was the same as that of the original clean train/test data.

Sphinx-3 toolkit from CMU was used for experiments. The Hamming window of 30-ms duration with 10-ms shift was used. For each frame, a 28 Mel-scaled filterbank distributed in the region $0$~6.0 kHz was applied. The Mel filterbank spectra (28-dimensional) was used as feature. Context-dependent HMMs with 1500 tied states were trained with TIMIT train set. The test sets of the filtered databases were used for testing. The pronunciation dictionary and vocabulary were downloaded from the CMU website. The trigram language model was adopted on the TIMIT database using a language model trained on RM database as an initial model. The following methods are used for comparison:

- **Baseline**: without any compensation;
- **BM**: bounded marginalisation;
- **BD**: bounded data imputation;
- **BMC**: bounded marginalisation with compensation of part missing components;
- **BDC**: bounded data imputation with compensation of part missing components.

In our experiments, Oracle mask is used for BM and BD. It is obvious that Oracle mask discards two partly missing components, namely, the $I$-th and $J + 1$-th ones. In addition, mismatch is still existing in the ($I + 1$)-th and $J$-th components. We treat partly missing components to be reliable since they are compensated in our methods. It is to be noted that $I$ and $J$ are decided by the bandwidth of band-limited speech.

The performance comparisons are presented in Table 1 and 2. The word accuracy rate (WAR) is 89.4% under matched condition (The bandwidth of test data is $0$~6.0 kHz). The WAR of Baseline decreases drastically as the available speech band decreases for the test data. This indicates that the mismatch becomes larger with the missing spectrum band increasing. When the bandwidth is 2.0~4.0 kHz, the WAR drops down to 28.8%. Both BM and BD obtains better performances than Baseline. When the cutoff frequencies are higher than 4.5 kHz, the WAR approaches to that of full-bandwidth test data. Under low-pass conditions, BMC performs better than BM, obtaining a relative improvement of more than 1.0%. BDC also achieves a better performance than BD. Under band-pass conditions, both low and high frequency components are missing. It is seen that BMC and BDC show much more compensation ability. Especially, when the lower limit cutoff frequency is higher than 1.0 kHz and the upper limit cutoff frequency is lower than 5.5 kHz, a relative improvement of more than 2.0% has been achieved. This may be because the four partly missing com-
ponents play a more important role when more components are missing.

Table 1. Comparison of the performance under low-pass conditions (WAR, %).

<table>
<thead>
<tr>
<th>Available bandwidth (kHz)</th>
<th>Methods</th>
<th>0~3.0</th>
<th>0~3.5</th>
<th>0~4.0</th>
<th>0~4.5</th>
<th>0~5.0</th>
<th>0~5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>52.8</td>
<td>68.2</td>
<td>70.9</td>
<td>80.4</td>
<td>85.8</td>
<td>87.0</td>
</tr>
<tr>
<td>BM</td>
<td></td>
<td>66.5</td>
<td>75.1</td>
<td>80.3</td>
<td>84.9</td>
<td>87.4</td>
<td>89.2</td>
</tr>
<tr>
<td>BD</td>
<td></td>
<td>65.4</td>
<td>74.5</td>
<td>80.0</td>
<td>85.2</td>
<td>87.7</td>
<td>89.2</td>
</tr>
<tr>
<td>BMC</td>
<td></td>
<td>68.0</td>
<td>76.2</td>
<td>82.4</td>
<td>85.7</td>
<td>88.0</td>
<td>89.2</td>
</tr>
<tr>
<td>BDC</td>
<td></td>
<td>66.3</td>
<td>75.0</td>
<td>81.6</td>
<td>86.2</td>
<td>88.3</td>
<td>89.2</td>
</tr>
</tbody>
</table>

Table 2. Comparison of the performance under band-pass conditions (WAR, %).

<table>
<thead>
<tr>
<th>Available bandwidth (kHz)</th>
<th>Methods</th>
<th>0.5~6.0</th>
<th>1.0~5.5</th>
<th>1.5~5.0</th>
<th>2.0~4.5</th>
<th>2.0~4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>89.2</td>
<td>85.0</td>
<td>62.3</td>
<td>48.1</td>
<td>28.8</td>
</tr>
<tr>
<td>BM</td>
<td></td>
<td>89.2</td>
<td>86.2</td>
<td>69.8</td>
<td>53.1</td>
<td>45.3</td>
</tr>
<tr>
<td>BD</td>
<td></td>
<td>89.2</td>
<td>86.8</td>
<td>70.4</td>
<td>51.6</td>
<td>43.8</td>
</tr>
<tr>
<td>BMC</td>
<td></td>
<td>89.2</td>
<td>87.4</td>
<td>72.9</td>
<td>57.6</td>
<td>50.1</td>
</tr>
<tr>
<td>BDC</td>
<td></td>
<td>89.2</td>
<td>87.5</td>
<td>72.9</td>
<td>53.6</td>
<td>48.3</td>
</tr>
</tbody>
</table>

Table 3 presents the experimental results on TIMIT/NTIMIT databases. TIMIT train set is used for training, and NTIMIT test set is used for testing. As expected from the results in Table 1 and 2, Baseline at 37.6% WAR provides only a poor performance, while BM and BD improves the WAR up to 57.2% and 59.6%, respectively. After compensating the partly missing components, the WAR further goes up to 63.5% and 64.6%.

5. CONCLUSION

Although MFT-based methods are effective in compensating band-limited distortion, mismatch still exists in partly missing components. To solve this problem, we have developed a new approach to compensate the distortion in these partly missing components. This approach uses Monte Carlo method to obtain the matched PDFs of the partly missing components, moreover, it can be integrated easily into conventional MFT-based methods. Experimental results showed that the proposed method achieved better performance, especially when more components were missing.

6. APPENDIX

Taking the covariance (be equal to variance if the covariance matrix is diagonal) of both sides of (4) and assuming \( x_p \) is independent to \( \hat{\epsilon} \) yields

\[
\hat{\Sigma}_p = \text{Var}(\hat{\epsilon} x_p) = E(\text{diag}(\epsilon) x_p^2) - E^2(\text{diag}(\epsilon))E^2(x_p)
\]

where \((.)^2\) is an element-wise square operator. Considering \( x_p \) is independent to \( \hat{\epsilon} \):

\[
\hat{\Sigma}_p = E(\text{diag}(\epsilon) x_p^2) - E^2(\text{diag}(\epsilon))E^2(x_p)
\]

\[
= E(\text{diag}(\epsilon)^2)\text{Var}(x_p) + \text{Var}(\text{diag}(\epsilon))E^2(x_p)
\]

\[
= \text{diag}(E(\epsilon^2))\Sigma_p + \text{diag}(\text{Var}(\epsilon))\mu_p^2
\]

After adding subscript \( m \) and \( i \), equation (12) is obtained.

7. REFERENCES


