Missing-Feature Method for Speaker Recognition in Band-Restricted Conditions

Wool Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA

{wikim, John.Hansen}@utdallas.edu

Abstract

In this study, the missing-feature method is considered to address band-limited speech for speaker recognition. In an effort to mitigate possible degradation due to the general speaker independent model, a two-step reconstruction scheme is developed, where speaker class independent/dependent models are used separately. An advanced marginalization in the cepstral domain is proposed employing a high order extension method in order to address loss of model accuracy in the conventional method due to cepstrum truncation. To detect the cut-off regions from incoming speech, a blind mask estimation scheme is employed which uses a synthesized band-limited speech model. Experimental results on band-limited conditions indicate that our two-step reconstruction scheme with missing-feature processing is effective in improving in-set/out-of-set speaker recognition performance for band-limited speech, particularly in severely band-restricted conditions (i.e., 4.72% EER improvement in 2, 3, and 4kHz band-limited conditions over a conventional data-driven method). The improvement of the proposed marginalization method proves its effectiveness for acoustic model conversion by employing high order extension, showing 0.57% EER improvement over conventional marginalization.

Index Terms: speaker recognition, band-limited, missing-feature, marginalization, high order extension

1. Introduction

Bandwidth-restricted speech is one common issue that makes speaker recognition challenging in real-life scenarios involving transmission via different bandwidth media. An example where variable bandwidth restrictions occur happen in current broadcast news (e.g., CNN Headline News), where audio content from field correspondents will generally have restricted bandwidths versus studio anchors. These severe conditions on the audio stream increase acoustic mismatch among the available speech samples for extracting speaker traits, and finally lead to degraded performance of speaker recognition, which can be improved effectively for speaker clustering and adaptation to improve automatic speech recognition.

In general, band-limited speech, the focus in this paper, would severely degrade speaker recognition more than speech recognition. The limited spectral information of band-restricted speech prevents speaker recognition from utilizing the speaker traits latent in the entire frequency range especially at higher frequency. To address band-limited speaker recognition, CMN (Cepstral Mean Normalization), feature compensation/mapping schemes, various data-driven or adaptation techniques can be applied [1][2][3]. Retraining acoustic models using band-limited database is also an alternative.

In this study the missing-feature method [4][5] is considered as a solution to address band-limited speech for speaker recognition. We previously proposed a modified calculation of the posterior probability while depending only on reliable components for band-limited speech reconstruction [6]. To mitigate speaker characteristic degradation due to a general speech model, we develop a two-step feature reconstruction approach with speaker class independent/dependent models. The marginalization method is also considered for addressing band-limited speech. An advanced marginalization method employing high order extension in the cepstral domain is proposed to improve model accuracy which degrades in conventional marginalization due to truncation of high order coefficients. In order to detect the cut-off region from incoming speech, a blind mask estimation method is also employed.

2. Missing-Feature Reconstruction for Band-Limited Speech

2.1. Missing-Feature Reconstruction

The cluster-based reconstruction method proposed in [5], restores the unreliable spectral parts of input speech using the known distributions of clean speech and the reliable regions determined by the masks. In our earlier work, we changed the formulation of the marginal probability to a relation that only depends on the reliable observations $x_{r}^{(l)}$ by integrating the unreliable elements $x_{u}^{(l)}$ over the entire feature space for reconstruction of the band-limited speech [6]. Here, $\{l\}$ represents log-spectral domain. The equation for the posterior probability is approximated as follows,

$$P(k \mid x^{(l)}) \approx P(k) \int_{-\infty}^{\infty} P(x^{(l)} \mid k) dx_{u}^{(l)}$$

$$= P(k) P(x_{r}^{(l)} \mid k),$$

(1)

where $k$ is the cluster index among $K$ components in the clean speech model and $x^{(l)} = [x_{r}^{(l)} x_{u}^{(l)}]$. The final formulation is the posterior probability calculated using only observations which are determined to be reliable, that is, clean speech.

Missing-feature reconstruction is expected to be more effective particularly in speaker recognition of band-limited speech compared to other conventional feature compensation schemes such as RATZ, SPLICE, and VTS. In missing-feature reconstruction, only unreliable parts are reconstructed depending on the reliable spectral information which still maintains its
speaker characteristics. However, in other compensation methods, the entire feature vector is compensated or transformed indiscriminately resulting in a possible degradation of speaker traits due to the general speech model.

2.2. Two-Step Reconstruction Scheme

The missing-feature reconstruction method utilizes Gaussian Mixture Models (GMMs), which represents the feature space of general speech (i.e., speaker independent). If we rely on the general speech model, this would degrade speaker-dependent characteristics of the reconstructed speech, resulting in decreased speaker recognition performance. In this study, we employ a two-step feature reconstruction scheme where speaker class independent and dependent models are used. Fig.1 shows our two-step feature reconstruction framework.

Detection of the band-limited range of the input speech and reconstruction in the first step are accomplished by a speaker class independent model which is generated using the entire training speaker set. The feature obtained by the class-independent model is submitted to the speaker class detector and a second reconstruction procedure employing the class-dependent model. The missing-feature reconstruction and the data-driven feature compensation methods examined in this paper are implemented within the presented two-step framework.

3. Marginalization Method for Band-Limited Speech

3.1. Marginalization in Cepstral Domain

In this study, we also consider marginalization method for speaker recognition in band-limited condition. Instead of reconstructing missing parts, a marginalization method is applied to the recognition procedure, where the likelihood score is calculated only on reliable spectral component in log-spectral domain [4]. For more effectiveness, the marginalization in the cepstral domain has been proposed [7]. This method can be simply implemented by inserting a weighting matrix $W$ during the likelihood calculation of incoming cepstral feature vector $x$ with a model $\{\mu_k, \Sigma_k\}$ given as follows,

$$ (x - \mu_k)^T C^{-T} W^T C^{T} \Sigma_k^{-1} C WC^{-1} (x - \mu_k), \quad (2) $$

where $C$ is DCT (Discrete Cosine Transform) matrix. Here, diagonal elements of $W$ are corresponding to the mask information which is determined based on the range of band-restriction.

In our experiment, the marginalization is implemented in the following two steps:

(i) Marginal Model Construction: $\mu_k^\prime = C W C^{-1} \mu_k$, and

(ii) Marginal Feature Generation: $x^\prime = C W C^{-1} x$.

In the marginal model construction step, each mean vector $\mu_k$ in cepstral domain is converted into the log-spectral domain by inverse DCT, and unreliable components are weighted with 0 by matrix $W$, followed by returning to the cepstral domain. In the feature generation step, we can obtain the feature vector for marginalization in the cepstral domain in a manner similar to the model generation step.

3.2. Model Improvement Employing High Order Extension

The existing marginalization in the cepstral domain discussed in Sec.3.1 must have performance degradation due to employing a smoothed version of inverse DCT, since cepstral coefficients are generally obtained by truncation of higher orders after DCT transform of log-spectral coefficients. Here, we propose an advanced method using high order extension for the marginal model generation in order to mitigate the performance degradation due to coefficient truncation.

We first obtain an extended GMM model of the clean speech in the cepstral domain which has the same number of coefficients with log-spectral coefficients (i.e., no truncation),

$$ \lambda^e = \{\mu^e_i, \mu^i, \Sigma^e_i\}, \quad 1 \leq i \leq K^e. \quad (3) $$

where $\mu^e$ and $\Sigma^e$ have $M$ dimension which is greater than the size of original cepstral coefficients $N$. Here, we propose that the extended components of the mean vector are estimated by the following equation,

$$ \tilde{\mu}_{k,n} = \sum_{i=1}^{K^e} p(i|\mu_k) \mu_{i,n}, \quad N + 1 \leq n \leq M, \quad (4) $$

where,

$$ p(i|\mu_k) = \frac{\omega^e_i \mu_{i,n} \Sigma^e_i}{\sum_{j=1}^{K^e} \omega^e_j \mu_{j,n} \Sigma^e_j}. \quad (5) $$

In Eq.(5), $\tilde{\mu}_{k,n}$ and $\Sigma^e_i$ are the truncated versions of $\mu^e$ and $\Sigma^e_i$, which consist of $[\mu^e_1, \ldots, \mu^e_N]$ and $[\sigma^e_1, \ldots, \sigma^e_N]$ respectively.

Finally, the mean vector of the marginal model $\mu_k^m$ can be obtained as follows,

$$ \mu_k^m = C W C^{-1} \tilde{\mu}_{k}, \quad (6) $$

where $\tilde{\mu}_{k}$ is the extended mean vector consisting of $[\mu^e_k, \tilde{\mu}_{k,N+1}, \ldots, \tilde{\mu}_{k,M}]$, including the extended components obtained by Eq.(4). The marginal feature is generated using the extended feature vector $x^e$ which can be extracted with $M$ dimension as follows,

$$ x^m = C W C^{-1} x^e. \quad (7) $$

The resulting $\mu_k^m$ and $x^m$ have the same dimension $N$ as the conventional method, since high order components need to be truncated when returning to the cepstral domain. However, in the proposed method, the model/feature vector is converted into the log-spectral domain from the high order augmented cepstral domain, therefore it will be effective in increasing the accuracy of model/feature generation for marginalization.

4. Blind Band-Restriction Detection

As a preceding step for the missing-feature method, it is required to determine the “mask” which classifies the spectrum of the incoming speech into reliable and unreliable regions. Here, we describe our work on the blind mask estimation method using the synthesized band-limited speech models to determine unreliable regions from band-limited speech [6]. The band-limited speech model is synthesized from $K$-mixture GMM of
clean speech $x$ in the cepstral domain as $\{\omega_k, \mu_k, \Sigma_k\}$. The synthesized mean vector with cut-off frequency $b$ is obtained in a similar manner as the marginal model construction in Sec.3,
\[
\mu_{b,k} = C(WC^{-1}\mu_k + \alpha_k),
\]
where $W$ is a same weighting matrix in Eq.(2) and $\alpha_k$ consist of $[0, \ldots, 0, \alpha_{k,b}, \ldots, 0]$. Here, $\alpha_n$ denotes the floor value which is determined in this study as,
\[
\alpha_n = \min_k \{\mu_{b,n}^{(1)} - \beta \cdot \sigma_{b,n}^{(1)}\}, \quad b \leq n \leq M,
\]
where parameters in log-spectral domain $\mu_{b,n}^{(1)}$ and $\sigma_{b,n}^{(1)}$ can be obtained using inverse DCT of $\mu_k$ and $\Sigma_k$. Here, $\beta$ is set to 2.0 in our evaluation. Now, we have a total of $(M + 1)$ GMMs $\lambda_b = \{\omega_k, \mu_{b,k}, \Sigma_{b,k}\}$ which represent the distribution of the band-limited speech with $b$th band as the beginning of the cut-off frequency regions. Finally, a particular band-limited model is determined based on MAP estimation for the incoming speech $y$, followed by selection of the binary mask $S[n]$ as the cut-off frequency bands of the selected model as follows,
\[
S[n] = \begin{cases} 
1 \text{ (reliable)}, & \text{if } n < \hat{b} \\
0 \text{ (unreliable)}, & \text{otherwise}
\end{cases}, \quad 1 \leq n \leq M,
\]
where $\hat{b} = \arg \max_b P(\lambda_b | y)$.

5. Experimental Results

5.1. Experimental Setup

The presented methods are evaluated using an in-set/out-of-set speaker recognition framework. In-set/out-of-set is similar to open speaker recognition, except for our case we want to recognize only the presence of an in-set speaker, not which in-set speaker. The in-set/out-of-set system consists of two stages: i) classification and ii) verification [8]. In the classification step, the test speech is classified as the most likely speaker among a set of in-set speakers based on likelihood scores for the speaker dependent models. Next, we verify if the test speech truly belongs to the selected speaker (in-set) or not (out-of-set) in the verification stage. In our framework, a Universal Background Model (UBM) is used for the out-of-set model which is obtained as a GMM trained from the development speaker set. For each target in-set speaker, a speaker dependent GMM is created by MAP adaptation of the UBM models.

A set of 60 male speakers from the TIMIT database were randomly selected as the speaker sample space. These 60 speakers serve both as 30 in-set speakers and 30 out-of-set speakers (impostors) depending on the experimental set. For example, 30 speakers were randomly selected from the speaker samples space as the in-set speakers, with the remaining 30 speakers taking the role of impostors. In our experiment, 1,000 different combinations of in-set and out-of-set speakers were created, and all EERs (Equal Error Rate) were calculated by averaging the results from 1,000 different combinations to observe consistency of performance on different combination of in/out sets.

The training data was limited to approximately 5sec worth of speech, while test data was created as 8sec length. The remaining 378 male speakers, each having about 30sec of data, were used as development data. The band-limited speech for testing was generated by low-pass filtering the test speech data which has 8kHz as full-band frequency. Four kinds of low-pass filters were used for generating the test data including 2, 3, 4, and 5kHz respectively as the cut-off frequencies. A 32th-order Butterworth filter was used.

Table 1: In-set/out-set speaker recognition baseline performance on band-limited speech (EER, %).

<table>
<thead>
<tr>
<th>Available Bandwidth Speech</th>
<th>0-2kHz</th>
<th>0-3kHz</th>
<th>0-4kHz</th>
<th>0-5kHz</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>49.05</td>
<td>30.14</td>
<td>47.34</td>
<td>47.70</td>
<td>48.56</td>
</tr>
<tr>
<td>Matched</td>
<td>30.74</td>
<td>19.83</td>
<td>12.47</td>
<td>9.01</td>
<td>18.01</td>
</tr>
<tr>
<td>CMN</td>
<td>49.17</td>
<td>49.17</td>
<td>38.33</td>
<td>30.00</td>
<td>41.67</td>
</tr>
<tr>
<td>RATZ</td>
<td>49.17</td>
<td>37.50</td>
<td>22.50</td>
<td>12.50</td>
<td>30.42</td>
</tr>
</tbody>
</table>

A conventional MFCC feature extraction algorithm was employed in our system. A 32 Mel-scaled filter-bank is used, and the filter-bank outputs are transformed into 20 cepstral coefficients (c0-c19). A UBM is trained using development data with 32-component GMM. The number of Gaussian components is fixed to be 32 for all speakers across all experiments.

5.2. Baseline Performance

Table 1 shows baseline system performance (EER) as cut-off frequency varies in the test speech. The baseline performance (no processing) was compared to matched condition, and conventional methods which are Cepstral Mean Normalization (CMN) and a data-driven feature compensation (i.e., RATZ [1]). EER in clean condition (i.e., without band-restriction) is 3.12%.\(^1\) RATZ in Table 1 is implemented with “Oracle” information on speaker class and range of band-limitation.

Results show that the band-limited test condition severely degrades speaker recognition performance. Our earlier work on band-limited speech recognition showed that the performance decreases proportionally as the band-limited range of test speech increases [6]. The performance of matched condition is also considerably different from the trend of the band-limited speech recognition. In our experiment here, the EERs for the matched condition increase proportionally as the band-limited range of test speech expands, still resulting in low performance in cases of 3kHz and 2kHz band-limited matched conditions. The speech recognition in the matched condition was comparable to the clean 4kHz full-band speech condition even in 1.5kHz and 2kHz band-limitation [6]. This indicates that band-limited condition has a significant effect on speaker recognition which depends on speaker traits that are latent in not only the low but also high frequency regions.

5.3. Missing-Feature Method

Table 2 shows performance of missing-feature reconstruction (MFR) and three marginalization (MFM) methods with “Oracle” mask information. Missing-feature reconstruction is effective in decreasing EERs for speaker recognition, showing better performance in severely band-restricted speech (2, 3kHz) compared to RATZ in Table 1. This suggests that missing-feature reconstruction is effective in classifying speaker characteristics from severely restricted speech by reconstructing the cut-off region while only employing the reliable spectral parts where the speaker traits are included.

The marginalization methods show comparable performance to the matched condition in Table 1. MFM-Ideal indicates an ideal case of the marginalization method which is implemented without truncation during DCT conversion by employing 32 dimension MFCC feature vector. MFM-Base is conventional marginalization [7] and MFM-HE is the proposed method in this paper, which employs the high order exten-

\(^1\)The higher EER for clean data is due to the severe data limitation of 5sec train, 8sec test.
Table 2: Missing-feature methods with “Oracle” mask (EER, %).

<table>
<thead>
<tr>
<th>Available Bandwidth Speech</th>
<th>0-2kHz</th>
<th>0-3kHz</th>
<th>0-4kHz</th>
<th>0-5kHz</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFR</td>
<td>39.17</td>
<td>33.33</td>
<td>22.50</td>
<td>19.17</td>
<td>28.54</td>
</tr>
<tr>
<td>MFM-Ideal</td>
<td>36.05</td>
<td>26.86</td>
<td>16.54</td>
<td>8.98</td>
<td>22.11</td>
</tr>
<tr>
<td>MFM-Base</td>
<td>38.59</td>
<td>29.10</td>
<td>20.03</td>
<td>13.41</td>
<td>25.28</td>
</tr>
<tr>
<td>MFM-HE</td>
<td>37.98</td>
<td>28.10</td>
<td>19.21</td>
<td>13.53</td>
<td>24.71</td>
</tr>
</tbody>
</table>

Table 3: Performance within “Blind” framework (EER, %).

<table>
<thead>
<tr>
<th>Available Bandwidth Speech</th>
<th>0-2kHz</th>
<th>0-3kHz</th>
<th>0-4kHz</th>
<th>0-5kHz</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATZ</td>
<td>52.50</td>
<td>40.00</td>
<td>30.83</td>
<td>16.67</td>
<td>35.00</td>
</tr>
<tr>
<td>MFR</td>
<td>47.50</td>
<td>38.33</td>
<td>27.50</td>
<td>26.67</td>
<td>35.00</td>
</tr>
<tr>
<td>MFM-HE</td>
<td>38.36</td>
<td>28.28</td>
<td>19.21</td>
<td>14.49</td>
<td>25.09</td>
</tr>
</tbody>
</table>

5.4. Blind System Performance

Table 3 shows performance of the “Blind” system where no prior band-limited region and speaker class information are known. The blind band-restriction estimation method described in Sec. 4 was employed to detect band-limited range for the experiments. Knowledge on detected cut-off frequency is used for model selection for data-driven feature compensation (RATZ) and mask information for missing-feature methods (MFR & MFM). Both RATZ and MFR were implemented within the two-step feature reconstruction framework described in Sec. 2.2.

Even though performance is lower than the Oracle system due to unknown band-restriction conditions, the methods examined show effectiveness in increasing speaker recognition accuracy in band-limited conditions. In particular, the results show that missing-feature reconstruction outperforms the data-driven method in 2, 3, and 4kHz band-limited conditions. Our proposed marginalization method (MFM-HE) shows comparable performance to the Oracle condition. The larger drop in performance of RATZ and MFR in the blind system is considered mostly due to incorrect detection of speaker class for the two-step processing.

5.5. Evaluation on FISHER Corpus

Table 4 shows the performance evaluation using FISHER corpus [9] which presents a very challenging environment for speaker recognition due to session and channel variations. Here, we used a similar experimental setup with TIMIT corpus, employing 60 male speakers for in-set/out-of-set (15sec for test, 30sec for train) and 300 speakers for development (30sec for each speaker). 3kHz band-limited speech for recognition test was generated from 4kHz full band speech. The in-set/out-of-set test for the full band speech is 20.63% EER. Table 4 demonstrates that the proposed marginalization method (MFM-HE) outperforms the conventional method (MFM-base) in an adverse environment with speaker session/channel variations.

6. Conclusions

In this study, we considered the problem of speaker recognition of band-limited speech using missing-feature reconstruction and marginalization methods. We formulated a two-step feature reconstruction scheme to address the degradation due to the general speech model. An advanced marginalization method employing high order extension was proposed to address modeling accuracy loss of conventional marginalization in cepstral domain due to truncation of high order cepstrum. Experimental results on band-limited conditions demonstrate that our two-step scheme with missing-feature reconstruction is effective in improving EERs of in-set/out-of-set speaker recognition, and outperforms existing data-driven method in severe band-limited conditions. We note this is achieved with severely reduced data (5sec train, 8sec test). The proposed marginalization method shows better performance compared to conventional marginalization. The results demonstrate that the proposed high order extension technique for model conversion will be useful in other applications where acoustic model conversion is required such as PMC [10] and other model combination methods. While the EERs here are higher than the original full band performance, with bandwidth as low as 2kHz (i.e., 2-8kHz removed), the goal of in-set/out-of-set speaker recognition must shift from accurate recognition to knowledge discovery.

7. References