Missing-Feature Reconstruction for Band-Limited Speech Recognition in Spoken Document Retrieval

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems, Dept. of Electrical Engineering
University of Texas at Dallas, Richardson, Texas, USA
{wikim,john.hansen}@utdallas.edu

Abstract

In spoken document retrieval, it is necessary to support a variety of audio corpora from sources that have a range of conditions (e.g., channels, microphones, noise conditions, recording media, etc.). Varying band-limited speech represents one of the most challenging factors for robust speech recognition. The missing-feature reconstruction method shows the effectiveness in recognition of the speech corrupted by additive noise. However, it has a problem when applied to the band-limited speech reconstruction, since it assumes that the observations in the unreliable regions are always greater than the latent original clean speech. In this study, we propose to modify the current way to calculate the marginal probability for reconstruction into the computation depending only on the reliable components. To detect the cut-off regions from incoming speech, the blind mask estimation scheme is proposed, which employs the synthesized band-limited speech model without training database. Experimental results on Aurora 2.0 and actual band-limited speech (NGSW corpus) indicate that the proposed method is effective in improving recognition accuracy of the band-limited speech. Through combining with an adaptation method, 22.17% of relative improvement is obtained on NGSW.

Index Terms: speech recongnition, missing-feature, mask, band-limited, spoken document retrieval.

1. Introduction

The mismatch between training conditions and the environment where an actual speech recognition system operates is one of the primary factors degrading recognition accuracy. This is especially true for speech document retrieval systems which face the problem of robust speech recognition in order to address the wide diversity of speech corpora having severe acoustic conditions and enormous mismatches from training conditions. Bandwidth-restricted speech is one common issue that makes speech recognition challenging not only in spoken document retrieval but also in real-life scenarios involving transmission via different bandwidth media.

To address band-limited speech recognition, CMN (Cepstral Mean Normalization) and various data-driven or adaptation techniques have been proposed [1][2]. Retraining an HMM using band-limited database is an alternative. However, data-driven methods and retraining HMM require a prior knowledge and availability of the band-limited speech.

In this study the missing-feature method is considered as a solution to address band-limited speech for speech recognition. Missing-feature method has been effective in improving speech recognition in additive background noise conditions. It depends mostly on the characteristics of speech that are resistant to noise, rather than on the characteristics of the noise itself. The missing-feature method consists of two steps. The first step is estimation of a "mask" which determines which spectral parts of the noisy input speech are unreliable [3]. The second step is to reconstruct the unreliable regions or bypass them for other processing.

The cluster-based reconstruction method is employed for missing-feature of band-limited speech in our work [4]. We propose the modified calculation of the posterior probability to decide the cluster while depending only on reliable components. In order to detect the cut-off region from incoming speech, the mask estimation method is also proposed using the synthesized band-limited speech model. The proposed methods will be evaluated on synthesized band-limited speech from Aurora 2.0 and the real-life NGSW [5] speech samples.

We first review a spoken document retrieval system and its speech corpus in Sec.2. In Sec.3, the missing-feature reconstruction method employed in our work is discussed followed by development of the proposed algorithms in Sec.4 and 5. Representative experimental procedures and their results are presented and discussed in Sec.6. Finally, in Sec.7, we conclude our work.

2. SpeechFind system and NGSW corpus

SpeechFind is a spoken document retrieval system currently serving as the search engine for the National Gallery of the Spoken Word (NGSW) [5]. SpeechFind consists of two main phases; (i) enrollment and (ii) online search retrieval. In the enrollment, the focus is on automatic transcription of the speech materials. This includes automatic audio segmentation and transcription by a large vocabulary continuous speech recognition (LVCSR) engine. The second phase deals with information retrieval of transcribed documents using the modified version of the MG system.

The speech corpus from NGSW covers one of the largest ranges of audio materials available today. The audio content includes a diverse range of audio formats, recording media, and diverse time periods including names, places, topics, and choice of vocabulary. Some of these include severe bandwidth restrictions, poor audio from aged recording media, differences in microphone type, reverberation at public places, recordings from telephone, broadcasts, background noise, a wide range of speaking styles and accents, and so on [5].

The spectrograms shown in Fig.1 indicate representative examples of the wide range of distortion present in NGSW recording conditions. The speeches are spoken by (a) Thomas Edi-

This work was funded by grants from U.S. Air Force (F30602-03-0110) and by University of Texas at Dallas under Project EMMITT.

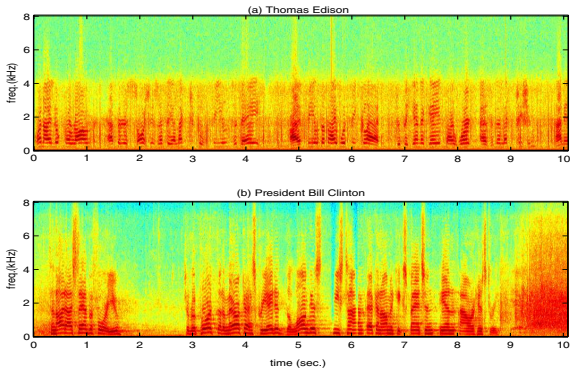

Figure 1: *Spectrograms of speech samples from NGSW.*

son(1907) and (b) President Bill Clinton(1999) respectively. Both are sampled at 16kHz, but (a) has a bandwidth restriction of about 1-2.5kHz due to the original recording media (i.e., Edison style cylinder disk). These kinds of severe conditions of speech increase the acoustic mismatch between training and testing conditions, and finally lead to degraded performance of speech recognition for automatic transcription. In this paper, we, especially, focus on distorted speech which is due to frequency bandwidth restriction.

3. Missing-feature reconstruction

The cluster-based reconstruction method has been proposed by Raj, et. al [4]. It restores the unreliable spectral parts of incoming speech using the known distributions of clean speech and the reliable regions determined by the masks. The distribution of the log-spectra of clean speech is modeled by Gaussian mixture with K clusters. Suppose that a noisy speech vector $S(t)$ has latent original components in an unreliable region $S_m(t)$ and reliable components $S_0(t)$. The cluster k of clean speech model is determined by the posterior probability. Since $S(t)$ has unreliable elements, the marginal computation is applied by integrating across them:

$$\begin{aligned}\hat{k}_{S(t)} &= \arg\max_k \{P(k)P(S(t) \mid k)\} \\ &= \arg\max_k \{P(k) \int_{-\infty}^{Y_m(t)} P(S(t) \mid k) dS_m(t)\} \quad (1)\end{aligned}$$

where $Y_m(t)$ represents the observed value of the unreliable parts and is assumed to be greater than $S_m(t)$. Finally, $S_m(t)$ is reconstructed using bounded MAP (Maximum A Posteriori) estimation based on the observations in the reliable regions with the Gaussian model of the cluster selected by Eq.(1), and an upper bound of $Y_m(t)$ as follows,

$$\widehat{S}_m(t) = \arg\max_{S_m(t)} \{P(S_m(t) \mid S_0(t), \mu_{\hat{k}_S(t)}, \Sigma_{\hat{k}_S(t)}, S_m(t) \le Y_m(t))\}. \quad (2)$$

4. Cluster determination for band-limited speech

The cluster-based reconstruction method described in Sec.3 assumes the case of missing speech which is corrupted by additive noise. The observation in the missed region $Y_m(t)$ is assumed to be greater than the latent clean component of the same region which will be estimated. The observation gives the upper bound of integration for the marginal probability to determine the cluster as shown in Eq.(1).

However, the situation is different in the case of channel-distorted band-limited speech which is the focus in this paper. The observations are not necessarily greater than the original clean spectral parts. This is especially the case for band-limited speech, where the observations of the cut-off frequency region generally have very low energy signals. Therefore, integration using the observation values as the upper bound no longer correctly reflects the marginal computation over the unreliable space where the original clean speech might exist. This leads to an erroneous calculation of the marginal probability and finally results in an incorrect reconstruction of the missing-feature.

Here, we propose to change the formulation of the marginal probability in Eq.(1) to a relation that only depends on the reliable observations $S_0(t)$ by integrating the unreliable elements over the entire feature space. This is approximated using the following equation,

$$\begin{aligned}\hat{k}_{S(t)} &\approx \arg\max_k \{P(k) \int_{-\infty}^{\infty} P(S(t) \mid k) dS_m(t)\} \\ &= \arg\max_k \{P(k)P(S_0(t) \mid k)\}. \quad (3)\end{aligned}$$

This is not a accurate calculation compared to the original form, and is especially the case since the estimated probability becomes more incorrect as the number of unreliable elements increases. However, it would mitigate the incorrectly computed marginal probability due to relying on the observations in the cut-off frequency region of the band-limited speech.

5. Blind mask estimation using band-limited speech model

As a preceding step for missing-feature reconstruction, it is required to determine which parts are the missed regions in the spectrum of the incoming speech. In real-world conditions, the information concerning band restriction is often unavailable, so it is necessary to detect this automatically from speech. Here, we propose a blind mask estimation method using the synthesized models to classify the unreliable regions from band-limited speech.

The band-limited speech we focus on is a special case in the fact that the reliable spectral information of the speech exists only from zero to a particular frequency range. Considering this, we can synthesize the band-limited speech model from the distribution of clean speech without a training database. For the missing-feature reconstruction as shown in Sec.3, we already have a K-mixture GMM of clean speech in the log-spectral domain,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}). \quad (4)$$

If the frequency region from the mth band to full range in clean speech $\mathbf{x}$ is cut-off by a band restriction, the band-limited speech $\mathbf{y}$ in the log-spectral domain can be presented as,

$$\mathbf{y} = [y_1, y_2, \ldots, y_M]^T = [x_1, x_2, \ldots, x_{m-1}, c_m, \ldots, c_M]^T. \quad (5)$$

Here, M refers to the number of coefficients which is identical to the size of the Mel-filter bank and c_m denotes the floor value which

has very low energy in the cut-off frequency region. If the band-limited speech $\mathbf{y}$ is also assumed to have a Gaussian distribution, its mean vector of kth mixture is given by

$$\boldsymbol{\mu}_{\mathbf{y},k} = [\mu_{\mathbf{x},k,1}, \mu_{\mathbf{x},k,2}, \ldots, \mu_{\mathbf{x},k,m-1}, c_m, \ldots, c_M]^T. \quad (6)$$

The GMM of the band-limited speech $\mathbf{y}$ which has the mth to full range cut-off can be defined as,

$$\boldsymbol{\lambda}_m = (\omega_k, \boldsymbol{\mu}_{m,k}, \boldsymbol{\Sigma}_{m,k}), 0 \leq m \leq M, 1 \leq k \leq K, \quad (7)$$

and its mean vector $\boldsymbol{\mu}_{m,k}$ is same as in Eq.(6). The mean of 0th model λ_0 becomes $[c_1, \ldots, c_M]^T$ which indicates the full-band cut-off speech, and the mean of the Mth model $\boldsymbol{\lambda}_M$ is $[\mu_{\mathbf{x},k,1}, \ldots, \mu_{\mathbf{x},k,M}]^T$ which implies the clean speech $\mathbf{x}$. Now, we have $(M+1)$ GMMs which represent the distribution of band-limited speech from the 0 to Mth band as the limited frequency regions. In our work, the prior probabilities ω_k and covariance matrices $\boldsymbol{\Sigma}_{m,k}$ are maintained the same as the GMM of the clean speech.

The obtained $(M+1)$ number of band-limited speech models can be converted into the cepstral domain.

$$\boldsymbol{\lambda}_m^{\{c\}} = (\omega_k, \mathbf{C}\boldsymbol{\mu}_{m,k}, \mathbf{C}\boldsymbol{\Sigma}_{m,k}\mathbf{C}^T) = (\omega_k, \boldsymbol{\mu}_{m,k}^{\{c\}}, \boldsymbol{\Sigma}_{m,k}^{\{c\}}) \quad (8)$$

where $\mathbf{C}$ refers to the DCT (Discrete Cosine Transform) matrix and $\{c\}$ represents the cepstral domain. The computational expense is reduced by decreasing the number of coefficients and avoiding the full-covariance matrix of the log-spectrum domain. Finally, a particular band-limited model is determined based on MAP estimation from the incoming speech, followed by selection of the binary mask $S[m]$ for the spectrogram as the number of cut-off frequency bands of the selected model as shown in Eq.(9) and (10),

$$\widehat{m} = \arg\max_m P(\boldsymbol{\lambda}_m^{\{c\}} \mid \mathbf{x}^{\{c\}}) = \arg\max_m \{P_m P(\mathbf{x}^{\{c\}} \mid \boldsymbol{\lambda}_m^{\{c\}})\}, \quad (9)$$

$$S[m] = \begin{cases} 1 \text{ (reliable)}, & \text{if } m < \widehat{m} \\ 0 \text{ (unreliable)}, & \text{otherwise} \end{cases} \quad 1 \leq m \leq M \quad (10)$$

where P_m denotes the prior information of the mth band-limited speech model.

6. Experimental results

6.1. Evaluation on sythesized speech: Aurora2.0

We evaluate the proposed methods following the procedures specified by Aurora 2.0 [6]. HMMs for speech recognition and GMMs for cluster-based reconstruction were trained using a clean training database that contains 8,440 utterances. The band-limited speech for testing was generated by low-pass filtering a set of clean speech in Aurora 2.0 which has 4kHz as a full-band frequency. Four kinds of low-pass filters were used for generating the test database including 1.5, 2, 2.5, and 3kHz respectively as the cut-off frequencies. A 32th-order Butterworth filter was used. Each test set has 1,001 samples. Fig.2 presents samples of the band-limited speech used in our experiments.

The performance of a baseline system and conventional method were examined as shown in Table 1. The recognition accuracies drastically decrease as the band-limited ranges of the test data shrink. This suggests that the difference between train and test conditions for speech recognition becomes larger as the cut-off region increases. When the HMM was trained on the identical band-limit condition as the test data, the performance was close to the baseline system on clean speech (Matched HMM). In order to compare our approach with existing methods for channel-distortion, we evaluated CMN and RATZ [1] which is one of several data-driven methods. For RATZ, 256-mixture GMM of the clean speech was used and its correction factors were obtained using the band-limited training database which has an identical condition to the test condition.

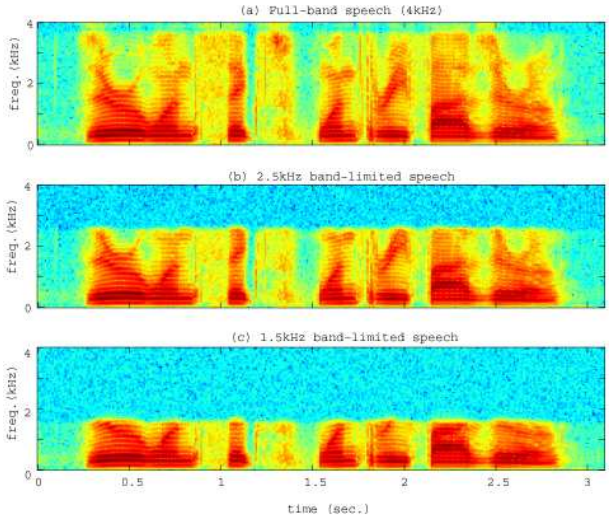

Figure 2: *Spectrograms of speech samples of Aurora2.0.*

Table 1: *Baseline performance on Aurora2.0. (word accuracy, %)*

	Clean	1.5kHz	2kHz	2.5kHz	3kHz
Baseline	98.87	22.77	26.81	50.34	90.37
Matched HMM	98.87	97.91	98.36	98.84	99.05
CMN	-	74.80	91.38	97.35	98.30
RATZ	-	73.37	89.23	96.00	98.03

Table 2 compares the recognition accuracy obtained using the missing-feature methods proposed in this paper to that obtained by the original missing-feature method. For cluster-based reconstruction, 32-mixture GMM was employed, which showed the best performance in our work. The first row presents the performance of the original missing-feature reconstruction with the masks derived from "Oracle" information which can be obtained by considering the cut-off range of the testing speech as shown in Table 3. Although the Oracle information concerning the band-restriction is known, the accuracies in cases of 1.5kHz and 2kHz are very low. This indicates that determining the cluster for missing-feature reconstruction relying on the observation values is not helpful in the case of band-limited speech as discussed in Sec.4.

The second row of Table 2 presents the performance also with the Oracle masks using the modified calculation of the posterior probability proposed in this paper which depends only on the reliable spectral components. Although the performance decreases as the cut-off frequency region becomes wider, there is significant improvement compared to the original reconstruction results in the first row. These results prove that the proposed modification for computing the posterior probability is very effective in missing-feature reconstruction of the band-limited speech.

The results in the third row of Table 2 indicate the recogni-

Table 2: *Performance of missing-feature methods. (word acc., %)*

	1.5kHz	2kHz	2.5kHz	3kHz
MF0+Oracle	47.72	87.09	97.49	98.60
MF+Oracle	**91.56**	**95.08**	**98.06**	**98.60**
MF+EstMask	**90.37**	**93.17**	**98.06**	**98.51**

Table 3: *Oracle masks used for missing-feature reconstruction.*

	Oracle masks	
1.5kHz	11111111111110000000000	(13/23)
2kHz	11111111111111110000000	(16/23)
2.5kHz	11111111111111111110000	(19/23)
3kHz	11111111111111111111000	(20/23)

tion accuracies obtained by employing the blind mask estimation proposed in Sec.5. Considering the number of log-spectral coefficients (=23) and the limited range of testing conditions, that is, 1.5 to 3kHz, twelve kinds of band-limited speech models were generated, which cover the cut-off frequencies from 1.0 to 3kHz. These are obtained by assigning the prior probabilities P_m of the band-limited models from 1.0 to 2kHz evenly (=$1/12$) while setting to zero for the other remaining twelve models from 0 to 1kHz in Eq.(9). The limited-band was determined once at every utterance by comparing the accumulated posterior probabilities. The results indicate that the proposed blind mask estimation using the synthesized band-limited model was considerably effective in detecting the reliable region from the spectrum of the band-limited speech.

6.2. Evaluation on actual band-limited speech: NGSW

The proposed missing-feature method was also evaluated on the band-limited speech obtained from the actual historical recordings. The testing samples are part of the NGSW corpus and they have 8kHz as their full band. About 3.8 hours of speech samples from six decades was transcribed by human experts for performance evaluation. Among them, three documents from 1950, 1960, and 1970 were identified as band-limited speech samples which are considered here.

The speech recognition engine used is SPHINX3 which was trained on 200-hour broadcast news [5]. The left-hand side of Table 4 shows the baseline performance of the recognition system on band-limited NGSW samples. If we compare the right part of Table 4, which presents the performance of the identical system for full-band speech, there is a degradation of performance by 11.7% in terms of WER, which could be considered due to missed information in the cut-off frequency regions of the band-limited speech.

From Table 5, RATZ did not improve performance and the missing-feature method (MF) did not show significant improvement compared to Aurora2.0 experiments even though there is improvement by 3.6% in average WER. In the missing-feature method (MF), the proposed scheme for determining the cluster was employed and Oracle masks were used. Applying the blind mask estimation failed to correctly detect the cut-off regions from the band-limited NGSW samples. The reason of difference from Aurora 2.0 in performance is that the mismatch between broadcast news used for training HMM and actual condition of NGSW is more severe than the mismatch between full band and band-limited data of NGSW.

The fourth and last rows are the results of performance by employing MLLR (Maximum Likelihood Linear Regression) adaptation. We obtained 18.01% relative improvement using the HMM adapted by MLLR. Through combining missing-feature method and MLLR, there was a relative improvement of 22.17%. The result implies that the mismatch of training and testing conditions affecting the recognition performance is more than the band-limited condition.

Table 4: *Baseline performance on NGSW corpus.*

	Band-limited (35 min.)		Full band (68 min.)	
	# words	WER(%)	# words	WER(%)
1950s (3-4.5kHz)	1,523	50.6	4,713	35.6
1960s (4.5kHz)	681	59.3	1,480	29.8
1970s (5.5kHz)	1,806	29.6	2,624	23.1
Total	4,010	**42.6**	8,817	**30.9**

Table 5: *Performance comparison on NGSW corpus. (WER, %)*

	1950s	1960s	1970s	Avg.(relative)
Baseline	50.6	59.3	29.6	42.6
RATZ	52.5	59.6	28.6	42.9 (-0.76%)
MF	**48.4**	**54.3**	**25.3**	**39.0 (8.50%)**
MLLR	46.7	45.1	21.2	34.9 (18.01%)
MF+MLLR	**44.1**	**42.6**	**20.4**	**33.2 (22.17%)**

7. Conclusions

In this study, we considered the problem of speech recognition of band-limited speech using missing-feature reconstruction. We proposed to modify the current calculation of the marginal probability for the reconstruction method to the computation depending only on the reliable components. To detect the cut-off regions from the incoming speech, the blind mask estimation scheme was also proposed, which employs the synthesized band-limited model without training database. Experimental results on Aurora 2.0 and NGSW demonstrate that the proposed method is effective in improving recognition accuracy of band-limited speech.

8. References

[1] Moreno, P. J., Raj, B., Stern, R. M., "Data-driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Comm.*, 24:267-85, 1998.

[2] Morales, N., Toledano, D. T., Hansen, J. H. L., Colas, J., Garrido, J., "Statistical Class-Based MFCC Enhancement of Filtered and Band-limited Speech for Robust ASR," *Interspeech2005*, Sep. 2005.

[3] Kim, W., Stern, R. M., "Band-Independent Mask Estimation for Missing-feature Reconstruction in the Presence of Unknown Background Noise," *ICASSP2006*, pp.305-308, 2006.

[4] Raj, B., Seltzer, M. L., Stern, R. M., "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Comm.*, 43(4): 275-296, 2004.

[5] Hansen, J. H. L., Huang, R., Zhou, B., Seadle, M., Deller, J. R. Jr., Gurijala, A. R., Kurimo, M., Angkititrakul, P., "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word," *IEEE Trans. Speech & Audio Proc.*, vol.13, no.5, Sep. 2005.

[6] Hirsch, G., Pearce, D., "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000*, Sep. 2000.