Time–Frequency Correlation-Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions

Wooil Kim, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Band-limited speech represents one of the most challenging factors for robust speech recognition. This is especially true in supporting audio corpora from sources that have a range of conditions in spoken document retrieval requiring effective automatic speech recognition. The missing-feature reconstruction method has a problem when applied to band-limited speech reconstruction, since it assumes the observations in the unreliable regions are always greater than the latent original clean speech. The approach developed here depends only on reliable components to calculate the posterior probability to mitigate the problem. This study proposes an advanced method to effectively utilize the correlation information of the spectral components across time and frequency axes in an effort to increase the performance of missing-feature reconstruction in band-limited conditions. We employ an *F1 Area Window* and *Cutoff Border Window* in order to include more knowledge on reliable components which are highly correlated with the cutoff frequency band. To detect the cutoff regions for missing-feature reconstruction, blind mask estimation is also presented, which employs the synthesized band-limited speech model without secondary training data. Experiments to evaluate the performance of the proposed methods are accomplished using the SPHINX3 speech recognition engine and the TIMIT corpus. Experimental results demonstrate that the proposed time–frequency (TF) correlation based missing-feature reconstruction method is significantly more effective in improving band-limited speech recognition accuracy. By employing the proposed TF-missing feature reconstruction method, we obtain up to 14.61% of average relative improvement in word error rate (WER) for four available bandwidths with cutoff frequencies 1.0, 1.5, 2.0, and 2.5 kHz, respectively, compared to earlier formulated methods. Experimental results on the National Gallery of the Spoken Word (NGSW) corpus also show the proposed method is effective in improving band-limited speech recognition in real-life spoken document conditions.**

***Index Terms*—Band-limited speech, correlation, missing-feature, speech recognition, time–frequency (TF).**

I. Introduction

BANDWIDTH-RESTRICTED speech is one common issue that makes speech recognition challenging for scenarios involving transmission via different bandwidth media [1], [2]. This is especially true for spoken document retrieval (SDR) where transcript generation needs to be accomplished using automatic speech recognition (ASR) and input environment conditions are generally unknown. Historical data such as the National Gallery of the Spoken Word (NGSW) requires SDR, which consists of audio recordings from the last 110 years [3], [4]. Such recordings contain media with different bandwidths due to limitations in the recording devices in past times, and consequently the speech bandwidth restriction is a significant issue for the ASR-based SDR.

To address band-limited speech recognition, general approaches to channel-distorted speech recognition can be considered. Conventional cepstral mean subtraction (CMS), various types of feature compensation methods, and hidden Markov model (HMM) adaptation represent a number of previous employed methods [5]–[7]. Retraining an HMM on the band-limited speech database is also an alternative. In our previous work, a data-driven based feature compensation method was proposed for band-limited speech recognition [1]. However, retraining HMM and data-driven methods require prior knowledge and availability of the band-limited speech. Several bandwidth extension (BWE) techniques employing either a codebook or Gaussian mixture model (GMM) can be considered as a variation of data-driven compensation method [2], [8], [9]. Recently, a BWE technique based on HMM has also been proposed [10].

In this paper, the missing-feature method is considered as a solution to address band-limited speech for speech recognition. Missing-feature processing has been known to be effective in improving speech recognition in additive background noise conditions [11]–[13]. This method 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 [14], [15]. The second step is to reconstruct the unreliable regions or bypass them for other processing. A prior effort has also attempted to address convolutional distortion using the missing-feature method, but such an approach is not for band-limited speech [16].

A cluster-based reconstruction method [13] is considered for missing-feature processing of band-limited speech in our study. This method restores unreliable parts of speech representations using the known distributions of speech sounds and the reliable region as indicated by mask information. The existing cluster-based method [13] is designed to compute the *a posteriori* probability of the incoming speech, employing marginal computation for the missing spectral regions which are assumed to be corrupted by only additive background noise. Therefore, how

Manuscript received May 19, 2008; revised December 27, 2008. Current version published July 17, 2009. This work was supported in part by the USAF under a subcontract to RADC, Inc., Contract FA8750-05-C-0029 (Approved for public release. Distribution unlimited.) and in part by the University of Texas at Dallas under Project EMMITT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: wikim@utdallas.edu; john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/TASL.2009.2015080

to reliably compute the *a posteriori* probability is a significant issue for band-limited speech in this study. Our earlier study proposed a modified approach for calculation of the *a posteriori* probability which depends only on reliable components for band-limited speech [17]. However, the performance of that method degrades as the cutoff region expands, due to the reduction in reliable information.

This present study represents a new effort to improve band-limited speech recognition performance within the missing-feature reconstruction framework. In this paper, we utilize more knowledge on the correlation information of reliable spectral components for the missing components in the cutoff frequency region. This approach aims at incorporating a correlation-based method [13] into the cluster-based method for missing-feature reconstruction by including an improved number of reliable components which are highly correlated across the time and frequency axes to the missing frequency region. In order to detect the cutoff region within the incoming speech, a mask estimation method is also presented, which employs a synthesized band-limited speech model.

This paper is organized as follows. We first look at an example of band-limited speech which can be observed in real world in Section II, and review the missing-feature reconstruction method in Section III. In Section IV, the issue of developing a missing-feature approach applied to band-limited speech is discussed in detail and our earlier work is presented. Section V represents the core novel algorithm aspects, where we investigate the prospects of utilizing further knowledge of the correlation with content in the cutoff band, followed by the proposed time–frequency (TF)-based method. Blind mask estimation details are described next in Section VI. Representative experimental procedures and their results are presented and discussed in Section VII. Finally, in Section VIII we state the main conclusions of our work.

II. Band-Limited Speech in Real-Life Scenario

As available online digital collections drastically increase, the need for automatic and efficient information retrieval continues to expand, placing demands on advances in technology including computational power and storage capacity. Recently, there has been growing interest in retrieving information, especially for online multimedia data consisting of rich information such as audio, video, and speech. Today, multimedia information collections include radio/television broadcast news, interviews, entertainment content, user generated content (UGC) such as YouTube, and others. This increasing demand has drawn remarkable attention to expanding research on SDR [3], [18]–[21].

SpeechFind [3] is an SDR system serving as the platform for several programs across the U.S. for audio indexing and retrieval including the NGSW and the Collaborative Digitization Program (CDP) [4], [22]. The system consists of two main phases: 1) enrollment and 2) online search retrieval. In the enrollment phase, 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 a modified version of the MG system [3], [23].

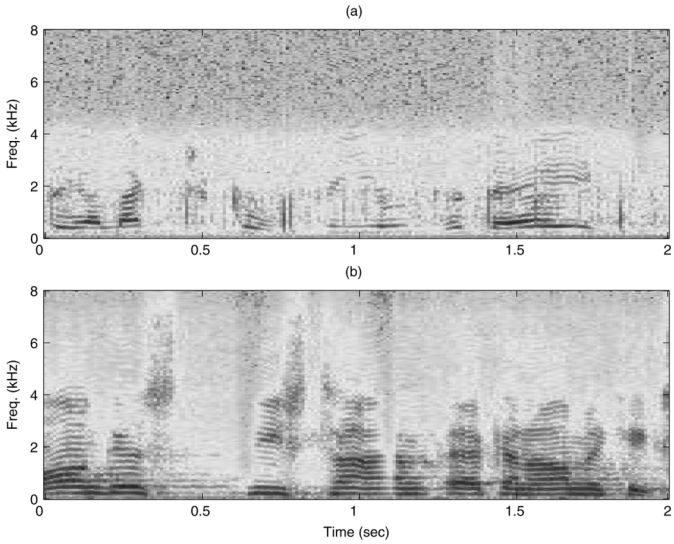


Fig. 1. Spectrograms of speech samples from NGSW. (a) Thomas Edison. (b) President Bill Clinton.

The speech corpora from NGSW and CDP cover a wide range of audio materials. The audio content includes a diverse range of audio formats, recording media, and diverse time periods including names, places, topics, and choice of vocabulary from the last 110 years. 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 others [3], [24].

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 Edison (1907) and (b) former President Bill Clinton (1999), respectively. Both are sampled at 16 kHz, but (a) has an available bandwidth of about 1.5–2.5 kHz due to the original recording media (i.e., Edison style cylinder disk). Other examples 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 while studio anchors' are typically full bandwidth. These severe conditions on the audio stream 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 focus on a feature reconstruction scheme to improve recognition performance of speech distorted by bandwidth limitations.

III. Missing-Feature Reconstruction

A cluster-based missing-feature reconstruction method has been proposed by Raj *et al.* [13]. It restores the unreliable spectral parts of input 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 $X(t)$ is modeled by a Gaussian mixture with K clusters

$$p(X(t)) = \sum_{k=1}^{K} \omega_k \mathcal{N}(X(t); \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}). \tag{1}$$

Suppose that a clean speech vector $X(t)$ has reliable components $X_r(t)$ with the latent original components in an unreliable (i.e., *missing*) region $X_u(t)$. That is, $X(t) = [X_r(t)X_u(t)]$. The reliable component $X_r(t)$ is identical to the corresponding observation $Y_r(t)$. The cluster k of the clean speech model is determined by the posterior probability [5], [25]. Since $X(t)$ contains unreliable elements, the marginal computation is applied by integrating out their dependency

$$\hat{k} = \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_u(t)} P(X(t) \,|\, k) dX_u(t) \right\} \quad (2)$$

where $Y_u(t)$ represents the observed value of the unreliable parts and is assumed to be greater than $X_u(t)$ because it is corrupted by additive background noise. Finally, the unreliable part $X_u(t)$ is reconstructed using bounded maximum *a posteriori* (MAP) estimation based on the observations in the reliable regions $X_r(t)$ with the model parameters of the cluster $\hat{k}$ selected by (2), and an upper bound $Y_u(t)$ as follows [13]:

$$\hat{X}_u(t) = \arg\max_{X_u(t)} \{ P(X_u(t) \,|\, X_r(t), \boldsymbol{\mu}_{X,\hat{k}}, \boldsymbol{\Sigma}_{X,\hat{k}}, X_u(t) \le Y_u(t)) \}. \quad (3)$$

Equation (3) can be simplified into the following equation [26]:

$$\hat{X}_u(t) = \boldsymbol{\mu}_{\hat{k},u} + \boldsymbol{C}_{\hat{k},ru} \cdot \boldsymbol{C}_{\hat{k},rr}^{-1} \cdot [Y_r(t) - \boldsymbol{\mu}_{\hat{k},r}] \quad (4)$$

where $\boldsymbol{C}_{\hat{k},ru}$ and $\boldsymbol{C}_{\hat{k},rr}$ are the covariance and cross-covariance matrices which are defined as follows:

$$\boldsymbol{C}_{\hat{k},rr} = E\{(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})^T\} \quad (5)$$

$$\boldsymbol{C}_{\hat{k},ru} = E\{(X_r(t) - \boldsymbol{\mu}_{\hat{k},r})(X_u(t) - \boldsymbol{\mu}_{\hat{k},u})^T\} \quad (6)$$

where $\boldsymbol{\mu}_{\hat{k},r}$ and $\boldsymbol{\mu}_{\hat{k},u}$ are mean vectors of the $\hat{k}$th cluster of the reliable component $X_r(t)$ and unreliable component $X_u(t)$ of the clean speech, respectively.

IV. Calculation of Posterior Probability for Band-Limited Speech

The cluster-based reconstruction method described so far assumes the case of missing speech which is corrupted by additive background noise. In this assumption, the observation in the missing region $Y_u(t)$ should be greater than the latent clean component of the same region $X_u(t)$ which will be estimated. The observation $Y_u(t)$ provides the upper bound of integration for the marginal probability to determine the cluster as shown in (2). In other words, it plays a role on the upper bound of the range where the original clean speech would be distributed.

However, the situation is different in the case of channel-distorted band-limited speech which is the focus of this study. The observations are not necessarily greater than the original clean spectral parts. The observations of the cutoff frequency region generally have very low levels of signal energy in case of band-limited speech. Therefore, integration using the observation values as the upper bound as in (2) 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.

In our earlier work [17], we have proposed to change the formulation of the marginal probability used for determining the most likely cluster in (2) to a relation that only depends on the reliable observations $X_r(t)$ by integrating the unreliable elements over the entire feature space for reconstruction of band-limited speech. The equation for the posterior probability is approximated using the following equation:

$$P(k \,|\, X(t)) \approx P(k) \int_{-\infty}^{\infty} P(X(t) \,|\, k) dX_u(t) = P(k) P(X_r(t) \,|\, k). \quad (7)$$

The final formulation is the posterior probability calculated using only the observations which are determined to be reliable, that is, the clean speech components. This might not be an accurate calculation for the posterior probability, which is especially the case since the estimated probability becomes less reliable as the number of unreliable elements increases. However, this formulation is expected to mitigate the incorrectly computed marginal probability which is obtained by the original equation relying on the observations in the cutoff frequency region of the band-limited speech.

Initial experiments employing this proposed prior method showed significant improvement compared to the original reconstruction method over band-restricted speech recognition [17]. However the performance, as expected, decreases as the cutoff frequency region becomes wider, which means that the estimate of the posterior probability based on the modified calculation method depending only on the reliable components becomes less effective as the number of reliable observations decreases. Raj also noted that the estimation of the cluster membership (i.e., $\hat{k}$ in (2)) would degrade when it depends only on reliable components [26]. Therefore, an alternative method is necessary to achieve improved performance for band-limited speech recognition.

V. TF-Based Missing Feature Reconstruction for Band-Limited Speech

In this section, we propose a novel approach to increase performance of the missing-feature reconstruction method for band-limited speech. The proposed method utilizes the correlation of the unreliable components in the cutoff frequency region with the reliable components from other neighboring frames as well as the current frame which conventional methods address. Raj *et al.* previously proposed a correlation-based reconstruction method in [13] and [26]. Their proposed method employs a relative covariance value to determine a neighborhood vector which is more correlated to missing components and used for reconstruction. In their method, however, the spectrogram of the clean speech signal is considered to be a wide-sense stationary random process, so the distribution of the clean speech is estimated simply using a single Gaussian probability density function (pdf). Such a simplification results in inferior performance compared to the cluster-based method which is the baseline scheme for our work in this paper [26]. Afify *et al.*

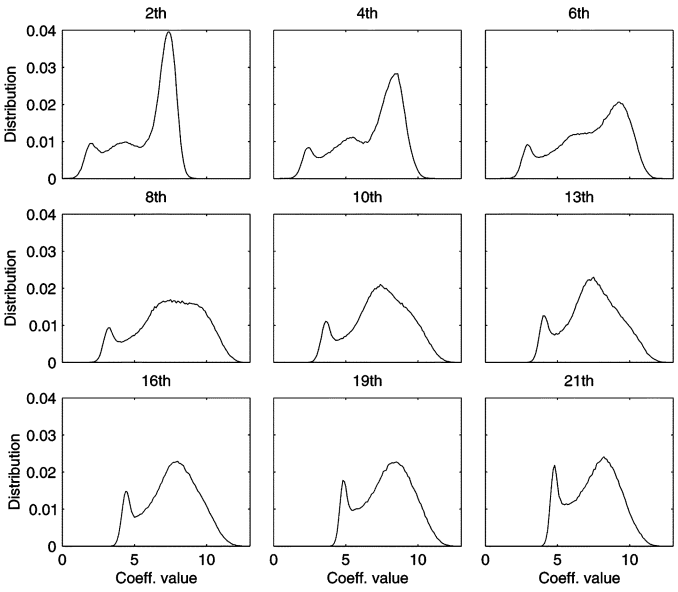


Fig. 2. Distributions of Mel filterbank output coefficients.

TABLE I
INDEX OF MEL FILTERBANKS AND CORRESPONDING CENTER FREQUENCY (HZ)

1	58	**9**	730	**17**	1,997
2	120	**10**	848	**18**	2,220
3	188	**11**	976	**19**	2,461
4	262	**12**	1,114	**20**	2,722
5	341	**13**	1,264	**21**	3,004
6	427	**14**	1,426	**22**	3,310
7	520	**15**	1,601	**23**	3,642
8	621	**16**	1,791		

recently proposed a method to exploit correlation for feature reconstruction of noisy speech in the cepstral domain [27].

As initial knowledge for our discussion, Table I and Fig. 2 show center frequency of each mel filterbank and the distributions of Mel filterbank coefficients employed in our study respectively. Our proposed method incorporates the concept of a correlation-based method into the cluster-based method by effectively utilizing the particular situation of incoming speech (i.e., band-limitation). Fig. 3[1] illustrates the correlation coefficients between a spectral component of a particular Mel filterbank index and the other components across 21 neighbor frames using clean speech from the TIMIT database (i.e., speaker and gender independent). The top 30% highly correlated components are highlighted with dark color in this figure. For example, the first plot shows the correlation coefficients between the tenth component of the Mel filterbank output and other adjacent components. It is not surprising that the correlation decreases as we increase the distance in time and index (i.e., frequency based dimension). However, it is quite interesting to note that the correlation values create a second peak near the sixth and seventh Mel filterbank indices, resulting in a bimodal distribution of the correlation coefficients. The sixth and seventh Mel filterbank indices approximately correspond to 500–600 Hz, where most of

[1]Similar plots have been presented in [26]. Here, we regenerated this figure to help illustrate the novel aspects of our proposed method.

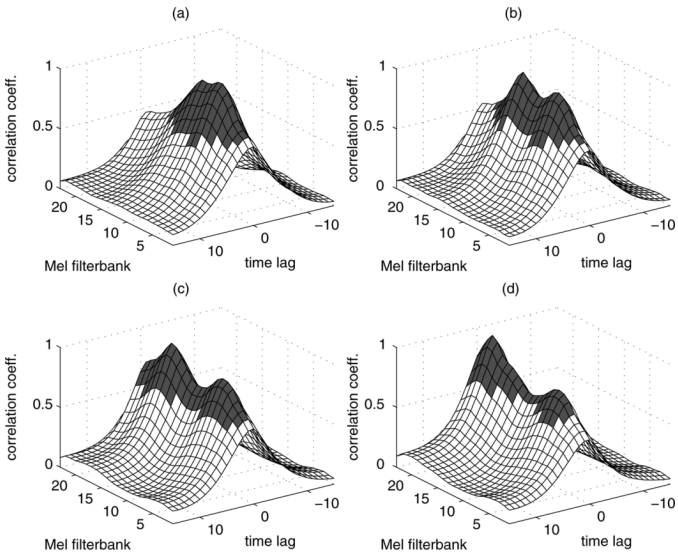


Fig. 3. Correlation coefficients with adjacent spectral components of clean TIMIT speech. (a) Tenth component. (b) 13th component. (c) 16th component. (d) 19th component.

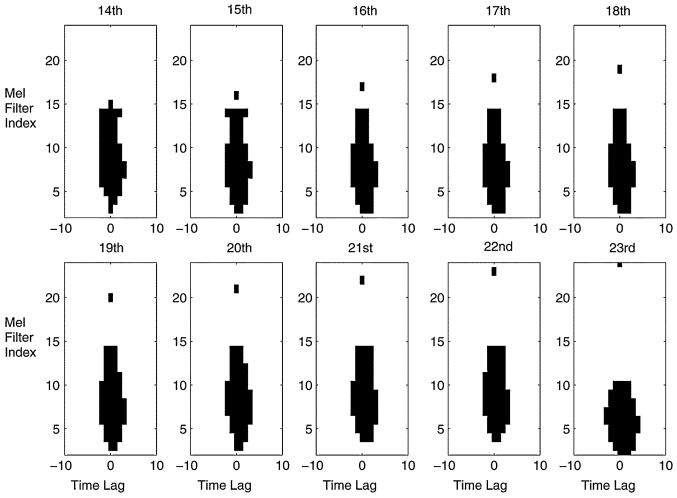


Fig. 4. Top 50 highly correlated spectral components for 0–1.5 kHz band-limited speech. Here, Mel filterbank indexes 14 to 23 correspond to the missing components for 1.5–4.0 kHz speech content.

the first formants for vowels are realized [28]. From this illustration, we can see that a particular spectral component is correlated highly not only with the adjacent components but also for spectral information around the first formant frequency.

Here, we shift our discussion of correlation characteristics of spectral components to the domain of interest, which is the frequency band-limited scenario. Figs. 4 and 5 show the 50 most highly correlated spectral components in the reliable frequency band with a particular Mel filterbank index in the cutoff frequency region in cases of 1.5 kHz (Mel filterbank index 14 to 23) and 2.0 kHz (Mel filterbank index 17 to 23) band-restrictions. For example, the first plot in Fig. 4 shows the top 50 spectral components (i.e., Mel filterbank outputs) in the reliable region (from 0 to 1.5 kHz) which are highly correlated with the 14th Mel filterbank index. The 14th index is the first component in the cutoff region (from 1.5 to 4 kHz) for the 1.5 kHz band-restricted speech. These figures were obtained by projecting the plots of the reliable frequency region in Fig. 3 onto the plane created

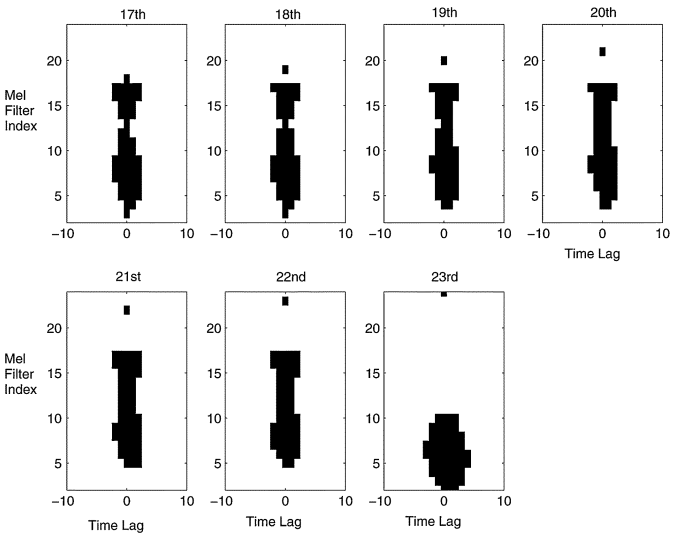


Fig. 5. Top 50 highly correlated spectral components for 0–2.0 kHz band-limited speech.

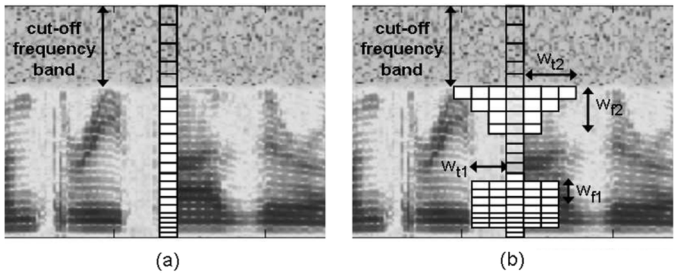


Fig. 6. Illustration of F1 Area Window and Cutoff Border Window for reliable spectral components. (a) Previous method [17]. (b) Proposed TF-MFR method.

by "Time Lag" and "Mel filterbank Index" axes. Therefore, the dark regions represent the 50 time-lag verse Mel-filterbank index entries with the highest correlation. A different trend of 23rd index from other indexes in Figs. 4 and 5 would be considered as lack of spectral characteristics due to out of the formant frequency range (i.e., over 3.5 kHz).

From these figures, we can see that the spectral components around the Mel filterbank index 3 to 9 are primarily correlated with each particular component in the 1.5–4.0 kHz cutoff frequency region. Indexes 3 to 9 correspond to 300–800 Hz, where we expect the first formant frequencies for vowels are distributed as discussed with Fig. 3. The spectral components at the boundary of the cutoff frequency region also show high correlation with the missing components in the cutoff region. Considering the trend in correlation coefficients in the band-limited condition, here we intend to increase the performance of missing feature reconstruction of band-limited speech by increasing the number of highly correlated components from reliable regions (i.e., the first formant and cutoff boundary areas) as well as the current frame components. Fig. 6 shows an illustration of how the window of reliable spectral components is constructed for incorporating additional content, which consists of two parts: 1) the first formant frequency area (*F1 Area Window*), and 2) cutoff frequency boundary area (*Cutoff Border Window*), in comparison to the previous method [17]. The *F1 Area Window* consists of two symmetric rectangular windows which are defined by w_{t1} and w_{f1}, having the sixth Mel filterbank index as a center frequency. The *Cutoff Border Window* also consists of two symmetric triangular windows which are determined by w_{t2} and w_{f2}, having the cutoff frequency as the upper limitation. Therefore, the proposed reconstruction method utilizes the correlation of the unreliable components to the spectral components across both time and frequency axes. The proposed F1 Area Window and Cutoff Border Window shapes would be different according to the range of unreliable components as shown in Figs. 4 and 5; however, we keep same shapes for the windows for simple implementation and performance evaluation in this study.

Since there is an increase in the number of reliable components from the *F1 Area* and *Cutoff Border* windows defined in Fig. 6, it is required that we modify the existing equations for reconstruction of the missing-feature components, previously discussed in Section III. Suppose n_B is the Mel-filterbank index corresponding to the cutoff frequency. The original clean speech at time t in the log-spectral domain can be represented by

$$X(t) = [X_r(t), X_u(t)]^T, \tag{8}$$

$$X_r(t) = [x(t,1), x(t,2), \ldots, x(t,n_B-1)]^T \tag{9}$$

and

$$X_u(t) = [x(t,n_B), \ldots, x(t,N)]^T \tag{10}$$

where N denotes the number of log-spectral components which is identical to the number of Mel filterbanks.

Here, we define the spectral components correlated across time and frequency axes $X^{\{tf\}}(t)$ by including the components in the additional window area $X_w(t)$ to $X(t)$ as follows:

$$X^{\{tf\}}(t) = [X(t), X_w(t)]^T. \tag{11}$$

Here, $X_w(t)$ consists of the F1 Area Window and Cutoff Border Window which are determined by $w_{t1} \times w_{f1}$ and $w_{t2} \times w_{f2}$.

The reliable components of $X^{\{tf\}}(t)$ is denoted as $X_r^{\{tf\}}(t)$, which represents $[X_r(t), X_w(t)]$. The distribution of the expanded clean speech $X^{\{tf\}}(t)$ is also assumed to be modeled by a Gaussian mixture with K components as follows:

$$p(X^{\{tf\}}) = \sum_{k=1}^{K} \omega_k \mathcal{N}\left(X^{\{tf\}}; \boldsymbol{\mu}_{X,k}^{\{tf\}}, \boldsymbol{\Sigma}_{X,k}^{\{tf\}}\right). \tag{12}$$

Equation (4) also needs to be rewritten to include $X_w(t)$ as follows:

$$\hat{X}_u(t) = \boldsymbol{\mu}_{\hat{k},u} + \boldsymbol{C}_{\hat{k},ru}^{\{tf\}} \cdot \boldsymbol{C}_{\hat{k},rr}^{\{tf\}^{-1}} \cdot \left[Y_r^{\{tf\}}(t) - \boldsymbol{\mu}_{\hat{k},r}^{\{tf\}}\right] \tag{13}$$

where $\hat{k}$ is determined by using (7), which shows more reliable performance for the GMM cluster decision rather than employing (12) in our experience. In a manner similar to that for (5) and (6), $\boldsymbol{C}_{\hat{k},rr}^{\{tf\}}$, and $\boldsymbol{C}_{\hat{k},uu}^{\{tf\}}$ are defined as follows:

$$\boldsymbol{C}_{\hat{k},rr}^{\{tf\}} = E\left\{\left(X_r^{\{tf\}}(t) - \boldsymbol{\mu}_{\hat{k},r}^{\{tf\}}\right)\left(X_r^{\{tf\}}(t) - \boldsymbol{\mu}_{\hat{k},r}^{\{tf\}}\right)^T\right\} \tag{14}$$

$$\boldsymbol{C}_{\hat{k},ru}^{\{tf\}} = E\left\{\left(X_r^{\{tf\}}(t) - \boldsymbol{\mu}_{\hat{k},r}^{\{tf\}}\right)\left(X_u(t) - \boldsymbol{\mu}_{\hat{k},u}\right)^T\right\} \tag{15}$$

where $\boldsymbol{\mu}_{\hat{k},r}^{\{tf\}}$ is the mean vector of $\hat{k}$th cluster of reliable component $X_r^{\{tf\}}(t)$ of clean speech, respectively.

As presented in (12), we employ K Gaussian distribution components for the expanded spectral components $X^{\{tf\}}(t)$ which includes $X_w(t)$. The expanded components $X^{\{tf\}}(t)$ include a greater number of components, which are expected to be more highly correlated to the missing spectral region. The obtained GMM for $X^{\{tf\}}(t)$ can represent a more diverse spectral distribution, which would address the problem of the existing correlation-based method where the distribution is too simplified and performance is low. In this paper, we name our proposed method as the Time–Frequency Correlation-based Missing-Feature Reconstruction (TF-MFR) method.

VI. Blind Mask Estimation Using Band-Limited Speech Model

As a preceding step for missing-feature reconstruction, it is required to determine the "mask" which classifies the spectrum of the incoming speech into reliable and unreliable ("missing") regions. In real-world conditions, the information concerning band restriction of the speech is often unavailable, so it is necessary to detect this automatically from the input speech. Here, we describe our approach to blind mask estimation using synthesized band-limited speech models to determine the unreliable regions from band-limited speech [17].

The band-limited speech we focus on is a special case where the reliable speech spectral information exists only from zero to a particular frequency range. With this condition, we can generate the band-limited speech models from the Gaussian mixture model of the clean speech without a training database. For missing-feature reconstruction as shown in Section III, we already have a K-mixture GMM of clean speech in the log-spectral domain as shown in (1).

If the frequency region from the n_Bth band to full range N in clean speech $X(t)$ is cutoff by a band restriction, the band-limited speech $Y(t)$ in the log-spectral domain (i.e., observation) can be written as

$$\begin{aligned} Y(t) &= [y(t,1),\ldots,y(t,N)]^T \\ &= [x(t,1),\ldots,x(t,n_B-1),y(t,n_B),\ldots,y(t,N)]^T. \end{aligned} \tag{16}$$

Here, the observations $y(t,n_B),\ldots,y(t,N)$ in the cutoff frequency region are assumed to have very low energy. If the band-limited speech $Y(t)$ with cutoff frequency n_B is also assumed to have a Gaussian distribution, its mean vector for the kth mixture is given by

$$\boldsymbol{\mu}_{n_B,k} = [\mu_{x_1,k},\ldots,\mu_{x_{n_B-1},k},c_{n_B},\ldots,c_N]^T \tag{17}$$

where c_n denotes the floor value which has a small value, and is determined in this study as

$$c_n = \min_t y(t,n), \quad 0 \le t \le T \tag{18}$$

which is a minimum value of observation at each nth band over T length of input speech $Y(t)$.

The GMM of the band-limited speech $Y(t)$, which has the n_Bth to full range cutoff N, can then be defined as

$$\boldsymbol{\lambda}_{n_B} = (\omega_k, \boldsymbol{\mu}_{n_B,k}, \boldsymbol{\Sigma}_{n_B,k}), \quad 0 \le n_B \le N, 1 \le k \le K. \tag{19}$$

Therefore, the mean of the 0th model $\boldsymbol{\lambda}_0$ becomes $[c_1,c_2,\ldots,c_N]^T$ which indicates the full-band cutoff speech, and the mean of the Nth model $\boldsymbol{\lambda}_N$ is $[\mu_{x_1,k},\mu_{x_2,k},\ldots,\mu_{x_N,k}]^T$ which is identical to the parameters in (1), implying the clean (i.e., full-band) speech $X(t)$. Now, we have a total of $(N+1)$ GMMs which represent the distribution of the band-limited speech from the 0 to the Nth band as the beginning of the cutoff frequency regions. In our work, the prior probabilities ω_k and covariance matrices $\boldsymbol{\Sigma}_{n_B,k}$ are maintained as the same values of the parameters of the GMM for the clean speech in (1).

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

$$\begin{aligned} \boldsymbol{\lambda}_{n_B}^{\{c\}} &= \left(\omega_k, \boldsymbol{C}\boldsymbol{\mu}_{n_B,k}, \boldsymbol{C}\boldsymbol{\Sigma}_{n_B,k}\boldsymbol{C}^T\right) \\ &= \left(\omega_k, \boldsymbol{\mu}_{n_B,k}^{\{c\}}, \boldsymbol{\Sigma}_{n_B,k}^{\{c\}}\right) \end{aligned} \tag{20}$$

where $\boldsymbol{C}$ refers to the discrete cosine transform (DCT) matrix and $\{c\}$ represents the cepstral domain. By converting to the cepstral domain, the computational expense is reduced by decreasing the number of coefficients and avoiding the full-covariance matrix needed in the log-spectrum domain. Finally, a particular band-limited model is determined based on MAP estimation using the incoming speech $Y^{\{c\}}(t)$, followed by selection of the binary mask $S[n]$ for the spectrogram as the number of cutoff frequency bands of the selected model as shown in (21) as

$$\begin{aligned} \hat{n}_B &= \arg\max_{n_B} P\left(\boldsymbol{\lambda}_{n_B}^{\{c\}} \middle| Y^{\{c\}}(t)\right) \\ &= \arg\max_{n_B} \left\{ P_{n_B} P(Y^{\{c\}}(t) \middle| \boldsymbol{\lambda}_{n_B}^{\{c\}}) \right\} \end{aligned} \tag{21}$$

$$S[n] = \begin{cases} 1\ (\text{reliable}), & \text{if } n < \hat{n}_B \\ 0\ (\text{unreliable}), & \text{otherwise} \end{cases} \quad 1 \le n \le N \tag{22}$$

where P_{n_B} denotes the prior probability of the n_Bth band-limited speech model. In this paper, for more reliable performance, a single decision is made for each utterance using accumulated posterior probabilities over the entire duration of the utterance under an assumption of stationarity of the given frequency band-restriction. The proposed blind mask estimation method using the synthesized band-limited speech model has an advantage requiring no training procedure, compared to a conventional data-driven method which needs training on band-limited speech database.

VII. Experimental Results

In our evaluations, the TIMIT database was used for the proposed method in this paper. A total of 4.1 h of speech (462

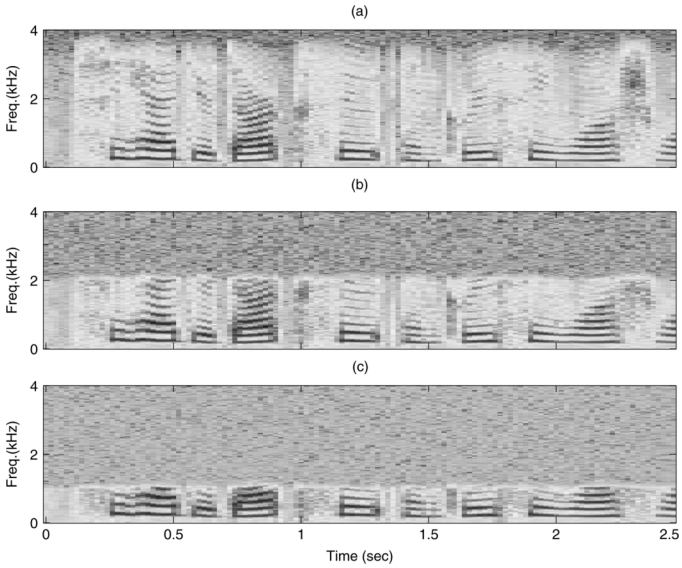


Fig. 7. Spectrograms of speech samples used in experiments. (a) Full-band speech. (b) 2.0-kHz band-limited speech. (c) 1.0-kHz band-limited speech.

speakers, 4620 utterances) were used for training, and 1.5 h of data (168 speakers, 1680 utterances) were used for test. Training and test sets do not overlap each other. Data was down-sampled to 8 kHz, so that each speech sample contains 4-kHz full-band frequency.

We employed SPHINX3 [29] to train the HMM for speech recognition and test recognition accuracy on band-limited speech. Each HMM represents a tri-phone which consists of three-state with an eight-component GMM per state, which is tied with 1138 states. The task has 6233 words as the vocabulary, and the trigram language model is adapted on the TIMIT database using a Broadcast News language model as an initial model.

A conventional Mel-frequency cepstral coefficient (MFCC) feature front-end is employed in this experiment, which uses a 23 of Mel-scaled filterbank. An analysis window of 25-ms duration is used with a 10-ms skip rate for 8-kHz speech data. The computed 23 Mel-filterbank outputs are transformed to 13 cepstrum coefficients including c0 (i.e., c0–c12). The first- and second-order time derivatives are also included, so the feature vector is 39-dimensional.

The band-limited speech samples for testing were generated by low-pass filtering the original clean speech for test in the TIMIT database with 4-kHz as full-band frequency. Four kinds of low-pass filters were used for generating the test database including 1.0, 1.5, 2.0, and 2.5 kHz, respectively, as the cutoff frequencies. A 32nd-order Butterworth filter was used to generate each set of band-limited speech database. Therefore, each test set consists of 1680 samples which is the same size as original clean test data. Fig. 7 presents samples of the band-limited speech used in our experiments.

A. Performance of Baseline and Previous Work

The performance of a baseline system (no compensation) and conventional methods (CMS and RATZ) were examined and summarized in Table II. Word error rates (WERs) drastically increase as the cutoff range increases for the test data. For the 0–2 kHz cutoff, the speech signal has generally lost the third formant information for almost all vowels, and for speech recognition with 0–1 kHz band-limited condition relies mostly on the first formant for vowels. This result also suggests that the difference between train and test conditions for speech recognition becomes larger as the cutoff region increases. When the HMM was trained on identical band-limit conditions as the test data (Matched HMM), the performance improved to the point of being comparable to the baseline system for clean (full-bandwidth) speech. In order to compare our approaches with existing methods for compensating channel-distortion, we evaluated cepstral mean subtraction (CMS) and RATZ (Multivariate Gaussian-based Cepstral Normalization) [5] which is one of several data-driven methods for feature compensation. For RATZ, a 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 (i.e., nonstereo training data).

TABLE II
BASELINE PERFORMANCE (WER, %)

	Available Bandwidth Speech				
	0-1kHz	0-1.5kHz	0-2kHz	0-2.5kHz	0-4kHz (clean)
Baseline	98.80	98.80	98.80	84.39	8.05
Matched HMM	20.61	15.78	12.24	9.46	8.05
CMS	91.84	49.78	20.78	11.70	8.05
RATZ	99.84	84.51	49.10	14.58	-

TABLE III
BAND-LIMITED SPEECH RECOGNITION EMPLOYING MISSING-FEATURE METHODS WITH ORACLE MASK (WER, %)

	Available Bandwidth Speech			
	0-1kHz	0-1.5kHz	0-2kHz	0-2.5kHz
MF0+Oracle	97.17	86.22	31.31	14.66
MF+Oracle	65.75	31.06	16.62	10.51

Table III shows the recognition performance obtained using the original missing-feature reconstruction method (MF0) [13] and our earlier proposed method (MF) [17] for band-limited speech. For cluster-based reconstruction, a 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 masks derived from "Oracle" information which can be simply obtained by considering the cutoff frequency of the testing speech as shown in Table IV, where 1 means reliable component and 0 indicates missing component in cutoff region. Although the Oracle information concerning the band-restriction is known, the recognition performance for cases of 0–1 kHz and 0–1.5 kHz 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 Section IV.

The second row of Table III presents performance also with the Oracle masks using the modified calculation of the posterior probability discussed in Section IV which depends only on the reliable spectral components. Although the performance degrades as the cutoff frequency region becomes wider, there is

TABLE IV
ORACLE MASKS USED FOR MISSING-FEATURE RECONSTRUCTION

	Oracle masks	
0-1kHz	11111111110000000000000	(10/23)
0-1.5kHz	11111111111110000000000	(13/23)
0-2kHz	11111111111111110000000	(16/23)
0-2.5kHz	11111111111111111100000	(18/23)

TABLE V
RECOGNITION PERFORMANCE AS CHANGE OF F1 AREA WINDOW SIZE ON 1.0-KHZ BAND-LIMITED SPEECH WITH ORACLE MASK (WER, %)

Baseline (MF+Oracle): 65.75% $w_{t1}\times w_{f1}$=0×0				
w_{f1}	w_{t1}			
	1	2	3	4
0	58.71	57.50	58.09	58.72
1	58.02	58.66	58.74	58.19
2	57.16	56.93	56.45	57.60
3	57.21	**56.13**	**55.98**	57.22
4	57.06	**54.47**	**54.73**	**55.35**

TABLE VI
RECOGNITION PERFORMANCE AS CHANGE OF F1 AREA WINDOW SIZE ON 2.0-KHZ BAND-LIMITED SPEECH WITH ORACLE MASK (WER, %)

Baseline (MF+Oracle): 16.62% $w_{t1}\times w_{f1}$=0×0				
w_{f1}	w_{t1}			
	1	2	3	4
0	15.61	14.84	14.79	15.37
1	14.69	14.92	14.64	14.68
2	14.86	15.11	**14.42**	**14.20**
3	14.77	14.84	**14.22**	14.55
4	15.10	14.82	**14.38**	**14.29**

significant improvement compared to the original reconstruction results in the first row. These results prove that the modified method for computing the posterior probability is very effective in missing-feature reconstruction of the band-limited speech.

B. Performance of TF-MFR Method: F1 Area Window

In this section, we present performance evaluation of the proposed TF-based missing-feature reconstruction (TF-MFR) method for band-limited speech. In order to find suitable sizes of the F1 Area Window and Cutoff Border Window for TF-MFR method (i.e., w_{t1}, w_{f1}, w_{t2}, and w_{f2}), the performance was examined for a changing window size along the time and frequency axes. First, to determine the size of the F1 Area Window (i.e., $w_{t1} \times w_{f1}$), the performance of the proposed TF-MFR method was investigated for a changing window size of w_{t1} and w_{f1} while setting both w_{t2} and w_{f2} to zero, which are shown in Table V to VII.

As illustrated in Fig. 6, the F1 Area Window (i.e., determined by $w_{t1} \times w_{f1}$) for the TF-MFR method consists of a two symmetric rectangle with having the current frame and the sixth Mel filterbank index as the center in time and Mel frequency bin respectively. In the experiments (Tables V to VII), we varied w_{t1} from 1 to 4, and w_{f1} from 0 to 4 to observe the performance trend. Considering the window size (25 ms) and skip rate (10 ms) for feature extraction (i.e., MFCC), changing the horizontal size of the F1 Area Window w_{t1} from 1 to 4 corresponds to a change in time from 45 ms ($=2\times 10\,\text{ms}+25\,\text{ms}$) to 105 ms ($=8\times 10\,\text{ms}+25\,\text{ms}$) including the current frame. The actual vertical size of the window w_{f1} from 0 to 4 corresponds to a change in frequency bin content from 90 Hz ($=430-340$) to 730 Hz ($=850-120$) having 384 Hz (i.e., sixth Mel filterbank index) as a center frequency.

TABLE VII
AVERAGE WER AND RELATIVE IMPROVEMENT AS CHANGE OF F1 AREA WINDOW SIZE ON FOUR BAND-RESTRICTIONS (1.0, 1.5, 2.0, AND 2.5 KHZ) WITH ORACLE MASK (%)

Baseline (MF+Oracle): 30.99% $w_{t1}\times w_{f1}$=0×0				
w_{f1}	w_{t1}			
	1	2	3	4
0	27.95 (7.75)	27.24 (10.00)	27.41 (9.60)	27.73 (8.42)
1	27.20 (10.87)	27.49 (9.22)	27.43 (9.92)	27.02 (11.05)
2	27.11 (10.42)	27.07 (9.99)	**26.40 (12.85)**	**26.68 (12.39)**
3	27.14 (9.94)	26.77 (10.90)	**26.30 (12.72)**	26.62 (12.29)
4	27.14 (10.00)	26.22 (12.31)	**25.96 (13.54)**	**26.10 (13.02)**

In Tables V–VII, the horizontal axis shows the change of w_{t1} and the vertical axis shows the change of w_{f1} for the TF-MFR method. Tables V and VI are the performance results for 0–1 kHz and 0–2 kHz band-restrictions, respectively. The average WERs for the four band-restrictions (0–1, 0–1.5, 0–2, and 0–2.5 kHz) are shown in Table VII and the average values of their relative improvements[2] are also presented, which are computed based on comparison to our previous work (MF + Oracle in Table III). From the tables, we can see there were significant improvements by incorporating the F1 Area Window for missing-feature reconstruction. The top five cases in performance are emphasized in bold font, and up to a 13.54% relative improvement in average WER was obtained for the case of $w_{t1} \times w_{f1} = 3 \times 4$.

The experimental results in Table VII are also presented as plots in Fig. 8. The upper panel shows average WERs at each window size w_{t1} using dash lines, and the lower panels are for each w_{f1}. The solid lines are the average performance for all cases in each panel. From the figure, we see that there is a consistent trend in performance versus a change of F1 Area Window size, which shows that WER improves as the window size in time and frequency axis increases, noting some exceptional outlier cases. It is not beyond our expectation that the more knowledge we have on the correlation information with the missing spectral components in cutoff regions, the more it would be helpful for their reconstruction.

In the case of 3×4 ($=w_{t1} \times w_{f1}$), which shows the best performance in Table VII, the actual size of F1 Area Window is 85 ms ($=6 \times 10$ ms $+ 25$ ms) and 730 Hz ($=850 - 120$) in time and frequency axes, and it uses 77 coefficients including the original 23 Mel filterbank outputs for the current frame. In other words, 54 spectral components are additionally employed by including the TF-based F1 Area Window. The case of $w_{t1} \times w_{f1} = 3 \times 3$ uses 65 spectral components which showed consistency in performance also for the Cutoff Border Window. These results (i.e., our selection is 3×4 or 3×3 for F1 Area Window) are considered to be well-matched to our findings presented with Figs. 4 and 5 in Section V. These figures show the highly correlated components located with the frequency region

[2]The average relative improvement is obtained by averaging the relative improvements of the four band-limited conditions, which is not calculated using the average WER.

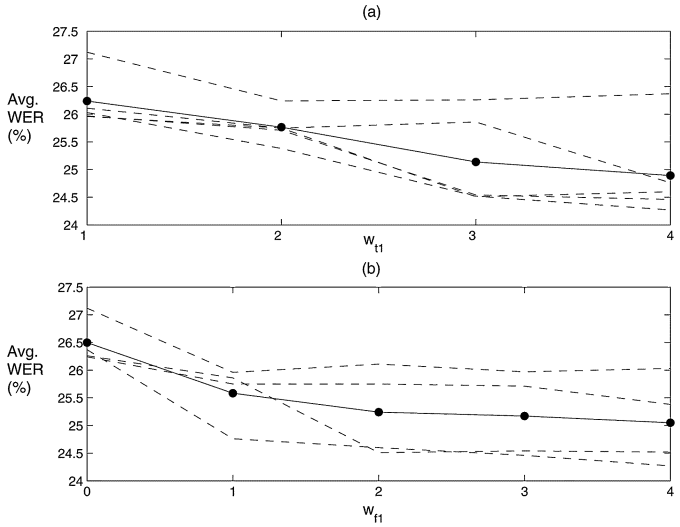


Fig. 8. Average WER (a) as change of w_{t1} and (b) as change of w_{f1} for F1 Area Window on four band-restrictions (1.0, 1.5, 2.0, and 2.5 kHz) with Oracle mask (%). (a) Average WER versus a change in w_{t1} from Table VII. (b) Average WER versus a change of w_{f1} from Table VII.

300–800 Hz and temporal space ± 30 ms time lag window. This selection also considered the computational expense and trainability of the acoustic model. Increasing the size of the F1 Area Window in time and frequency might improve overall performance; however, it requires increased size of the feature vector resulting in high computational expense for both model training and feature reconstruction. The computational expense is spent mostly on likelihood calculation with full covariance matrices and it primarily depends on the feature vector size. For example, the case of $w_{t1} \times w_{f1} = 4 \times 4$ needs a total of 95 components for the feature vector, which is greater than four times the feature vector used for the basic MFR. A large sized feature vector will also produce a "sparse data" problem under a restricted training data condition, leading to an unreliable acoustic model for missing-feature reconstruction. In the next section, the suitable size of Cutoff Border Window will be investigated based on the F1 Area Window determined in this section.

C. Performance of TF-MFR Method: Cutoff Border Window

In this section, we investigate the effect of the Cutoff Border Window on the performance of the proposed TF-MFR method. To find a suitable size of the Cutoff Border Window, the reconstruction performance was examined versus a changing size of Cutoff Border Window $w_{t2} \times w_{f2}$ with a fixed F1 Area Window. Tables VIII and IX show the average WERs as we change the Cutoff Border Window size, having $w_{t1} \times w_{f1} = 3 \times 3$ and $w_{t1} \times w_{f1} = 3 \times 4$, respectively, for the F1 Area Window.[3] The top 5 average relative improvements are highlighted in bold font. For the case of $w_{t1} \times w_{f1} = 3 \times 3, w_{t2} \times w_{f2} = 1 \times 3$ shows the best performance with a 1.60% ($=14.32 - 12.72$) increase in relative improvement compared to $w_{t2} \times w_{f2} = 0 \times 0$. We obtained the best performance with $w_{t2} \times w_{f2} = 2 \times 4$ in case of $w_{t1} \times w_{f1} = 3 \times 4$ having 1.07% improvement ($=14.61 - 13.54$).[4] For the selection of the Cutoff Border Window size (i.e., $w_{t2} \times w_{f2} = 1 \times 3$ or $w_{t2} \times w_{f2} = 2 \times 4$), we considered not only performance but also computational complexity and trainability which mostly depends on the feature vector size as discussed in the previous section for selection of the F1 Area Window size. The case of $w_{t2} \times w_{f2} = 2 \times 4$ of Table IX needs 87 spectral components for feature vector, while $w_{t2} \times w_{f2} = 4 \times 4$ requires 97 components.

The increase 1.60% and 1.07% in relative improvement might not appear to be significant. However, it should be noted that there is consistent WER improvement in most of the cases by employing the Cutoff Border Window, which outperforms the cases $w_{t1} \times w_{f1} = 4 \times 3$ and $w_{t1} \times w_{f1} = 4 \times 4$ in Table VII that have a larger sized F1 Area Window. This means that the spectral components in the Cutoff Border Window provide more effective information for correlation with the missing components in the cutoff frequency band, rather than excessively increasing the size of the F1 Area Window. In particular, note the case of $w_{t1} \times w_{f1} = 3 \times 3, w_{t2} \times w_{f2} = 1 \times 3$ which uses a total of 71 spectral components for TF-MFR method, while 77, 79, and 95 components are used in $w_{t1} \times w_{f1} = 3 \times 4, 4 \times 3$, and 4×4, respectively, with $w_{t2} \times w_{f2} = 0 \times 0$. This result proves that the combination of a suitable size F1 Area Window and Cutoff Border Window is effective in reconstructing the missing spectral components for band-restricted condition.

TABLE VIII
AVERAGE WER AND RELATIVE IMPROVEMENT AS CHANGE OF CUTOFF BORDER WINDOW SIZE WITH FIXED F1 WINDOW SIZE AS $w_{t1} \times w_{f1} = 3 \times 3$ (%)

Baseline1 (MF+Oracle): 30.99% $w_{t1} \times w_{f1}$=0×0				
Baseline2 (TF-MF+Oracle): 26.30% (12.72%) $w_{t1} \times w_{f1}$=3×3, $w_{t2} \times w_{f2}$=0×0				
w_{f2}	w_{t2}			
	1	2	3	4
1	26.00 (13.53)	25.86 (13.75)	**25.78 (14.21)**	**25.80 (14.14)**
2	25.94 (13.83)	25.94 (13.14)	25.85 (13.71)	25.80 (13.90)
3	**25.86 (14.32)**	**25.77 (14.15)**	25.93 (13.28)	25.82 (13.86)
4	**25.95 (13.99)**	25.84 (13.84)	25.84 (13.93)	25.81 (13.84)

TABLE IX
AVERAGE WER AND RELATIVE IMPROVEMENT AS CHANGE OF CUTOFF BORDER WINDOW SIZE WITH FIXED F1 WINDOW SIZE AS $w_{t1} \times w_{f1} = 3 \times 4$ (%)

Baseline1 (MF+Oracle): 30.99% $w_{t1} \times w_{f1}$=0×0				
Baseline2 (TF-MF+Oracle): 25.96% (13.54%) $w_{t1} \times w_{f1}$=3×4, $w_{t2} \times w_{f2}$=0×0				
w_{f2}	w_{t2}			
	1	2	3	4
1	25.88 (13.49)	25.74 (13.92)	**25.63 (14.27)**	25.77 (13.81)
2	25.82 (13.48)	25.67 (13.99)	**25.61 (14.25)**	25.72 (14.21)
3	25.74 (14.21)	25.69 (14.01)	25.78 (13.48)	25.73 (14.10)
4	25.75 (14.15)	**25.56 (14.61)**	**25.61 (14.39)**	**25.65 (14.70)**

[3]The case of 3×2 showed better performance in Table VII, however 3×3 showed more consistent performance in the following experiments.

[4]Even though the case of $w_{t2} \times w_{f2} = 4 \times 4$ showed better relative improvement in Table IX, we nevertheless select 2×4 as the best due to the number of coefficients employed.

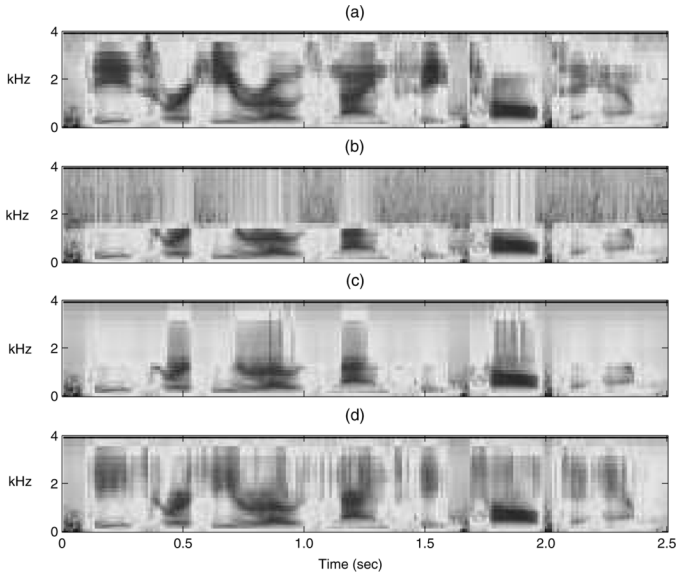


Fig. 9. Spectrograms of (a) full-band speech, (b) 1.5 kHz band-limited speech, (c) estimated by earlier work [17], (d) estimated by TF-MFR method; which are rebuilt from log-spectral coefficients. (a) Full-band speech. (b) 1.5-kHz band-limited speech. (c) Reconstructed by previous work [16]. (d) Reconstructed by proposed TF-MFR method.

TABLE X
SUMMARY OF AVERAGE WER AND RELATIVE IMPROVEMENT EMPLOYING F1 AREA WINDOW AND CUTOFF BORDER WINDOW WITH ORACLE MASK (%)

TF-MF Window Size		Available Bandwidth Speech				Avg.
$w_{t1}\times w_{f1}$	$w_{t2}\times w_{f2}$	0-1kHz	0-1.5kHz	0-2kHz	0-2.5kHz	
0×0	0×0	65.75	31.06	16.62	10.51	30.99
3×3	0×0	55.98 (14.86)	24.54 (20.99)	14.22 (14.44)	10.45 (0.57)	26.30 (12.72)
3×4	0×0	54.73 (16.76)	24.52 (21.06)	14.38 (13.48)	10.21 (2.85)	25.96 (13.54)
3×3	**1×3**	**55.20 (16.05)**	**24.04 (22.67)**	**13.96 (16.00)**	**10.24 (2.57)**	**25.86 (14.32)**
3×4	**2×4**	**54.24 (17.51)**	**23.62 (23.95)**	**13.97 (15.94)**	**10.40 (1.05)**	**25.56 (14.61)**

The performance evaluation results employing the F1 Area Window and Cutoff Border Window combined together that showed the best performance for experiments in Tables VIII to IX are summarized in Table X. This proves that our proposed method is effective in reconstructing the missing spectral components in the cutoff region by utilizing more effective knowledge from the adjacent spectral components which would be highly correlated with the missing components. Fig. 9 presents example spectrograms of (a) original (clean), (b) band-limited, (c) reconstructed from our previous work [17], and (d) reconstructed by the proposed TF-MFR method in this study. The example speech is band-limited at 1.5 kHz and the spectrograms are visually regenerated from the actual log-spectral coefficients for illustration. The spectrograms clearly show that the proposed method is very effective in restoring the missing spectral components in the cutoff frequency regions.

D. Performance of TF-MFR Method With Blind Mask Estimation

In this section, the proposed TF-MFR method is evaluated with the proposed blind mask estimation combined, which was presented in Section VI. First, Table XI shows the performance comparison of our blind mask estimation method to the oracle knowledge and conventional estimation method. Considering the number of log-spectral coefficients $(=23)$ and the range of test conditions (e.g., frequency bandwidth of 0–1.0 kHz and up to 0–2.5 kHz), 15 band-limited speech models were generated, which cover the cutoff frequencies from 0.7 to 4.0 kHz. These are obtained by assigning uniform prior probabilities $P_{n_B}(=1/15)$ of the band-limited models from 0.7 to 4.0 kHz, with P_{n_B} set to zero for the remaining eight models from 0.0 to 0.7 kHz in (21).[5] The cutoff band limit was determined once for each utterance by comparing the accumulated posterior probabilities over the entire duration of the utterance.

TABLE XI
CLASSIFICATION ACCURACY OF MASK ESTIMATION AND MFR PERFORMANCE (WER) BASED ON THE ESTIMATED MASK (%)

	Available Bandwidth Speech				Avg.
	0-1kHz	0-1.5kHz	0-2kHz	0-2.5kHz	
Oracle Mask	- (65.75)	- (31.06)	- (16.62)	- (10.51)	- (30.99)
Data Driven	90.12 (69.04)	54.58 (43.91)	100.00 (16.62)	100.00 (10.51)	86.18 (35.02)
Blind Estimation	**99.70 (65.99)**	**96.01 (42.91)**	**100.00 (16.62)**	**100.00 (10.51)**	**98.93 (34.01)**

For the conventional method, a data-driven method is employed, where acoustic models (i.e., GMMs) for different band-restrictions were obtained via training on band-limited speech. Fifteen GMMs are constructed to span 0.7 to 4.0 kHz cutoff frequencies, which represent identical conditions to the proposed blind mask estimator. Mask classification also employs the same procedure as the blind estimator, where a single decision is made employing MAP estimation for the entire duration of the input speech. Table XI demonstrates that the proposed blind mask estimation brings high classification accuracy which is comparable to the data-driven method. The achieved high performance is believed to be due to a well-matched model with band-limited test speech by using the artificially generated data. Taking minimum values from cutoff region in the proposed method would be effective to estimate the spectral characteristics matched with input speech, resulting in consistently better performance compared to the data-driven method which would be less reliable in modeling the poor spectra in cutoff region.

Classification performance improvement does not directly reflect a change in speech recognition performance when the estimated mask information is used for the missing-feature method. WER in the parenthesis in Table XI presents the recognition performance when the cutoff regions are reconstructed using the estimated mask information. The results show that blind mask estimation using the synthesized band-limited model was effective in detecting the reliable spectral region from the input band-limited speech. There is no significant loss in performance when employing the blind mask estimation method compared to the case of Oracle knowledge on the band-limited test speech, with the exception of 0–1.5 kHz case. The blind mask estimation is less correct for the 1.5 kHz case, resulting in degraded performance compared to the Oracle mask (42.91% versus 31.06%

[5]We assume that with speech at such low frequency content, it is not feasible to effectively estimate a pdf of speech where only up to 0.7 kHz of data is present.

TABLE XII
WER AND RELATIVE IMPROVEMENT WITH F1 AREA WINDOW AND CUTOFF BORDER WINDOW EMPLOYING BLIND MASK ESTIMATION (%)

TF-MF Window Size		Available Bandwidth Speech				Avg.
$w_{t1}\times w_{f1}$	$w_{t2}\times w_{f2}$	0-1kHz	0-1.5kHz	0-2kHz	0-2.5kHz	
0×0	0×0	65.99	42.91	16.62	10.51	34.01
3×3	**1×3**	**55.93 (15.24)**	**24.67 (42.51)**	**13.96 (16.00)**	**10.24 (2.57)**	**26.20 (19.08)**
3×4	**2×4**	**54.43 (17.52)**	**24.19 (43.63)**	**13.97 (15.94)**	**10.40 (1.05)**	**25.75 (19.53)**

WERs in Table XI). It is believed that band-restriction at 1.5 kHz smears the spectral characteristics of the second formant which generally exists around 1.5 kHz in frequency, leading to a degradation in discrimination among the models.

The proposed TF-MFR method employing the blind mask estimation is also consistently effective at improving the performance as shown in Table XII. It is worth noting that the performance for the case of 0–1.5 kHz is significantly improved, which had a 42.91% WER in the previous method with blind mask estimation. Here, we obtained WERs 24.67% and 24.19% for the 0–1.5 kHz case, which are comparable to the Oracle mask case in Table XII. Overall relative performance gains in WER range from 1.05 to 43.63% compared to the previous missing-feature reconstruction with blind mask estimation. This tells us that the proposed TF-MFR method is robust to the estimation of the band-limitation which might be incorrect depending on the particular band-limited conditions.

E. Performance of TF-MFR Method on Real-Life Conditions

The proposed missing-feature method was also evaluated on band-limited speech obtained from actual historical recordings within the NGSW corpus. The testing samples (191 utterances, 35 min) were found to be band-restricted to about 3–5 kHz, having 8 kHz as their full bandwidth. The speech recognition engine used for evaluation is SPHINX3 which was trained on 200 h of Broadcast News [3]. The baseline recognition system shows 30%–40% WER for other full band speech samples in the NGSW corpus, which indicates that the NGSW corpus contains challenging conditions for speech recognition even at full bandwidth speech.

From Table XIII, we can see the proposed missing-feature reconstruction method TF-MFR[6] consistently outperforms our previous proposed MFR method in both cases of Oracle mask and blind mask estimation. The relative low performance compared to results from the TIMIT corpus is believed to be due to mismatch between the acoustic model for missing-feature and test speech conditions. A clean TIMIT corpus was employed for training the model for missing-feature method, which is highly different from the acoustics of full bandwidth speech in the NGSW corpus. We obtained improved results by combining maximum likelihood linear regression (MLLR) adaptation on the reconstructed speech for the HMM. We can also see that the missing-feature method employing the blind mask estimation shows consistent improvement for the NGSW corpus, even

[6]We obtained the best result with $w_{t1} \times w_{f1} = 3 \times 3$ without the Cutoff Border Window in this experiment. Due to the restricted size of test data in the NGSW corpus, we did not find statistical significance of the performance change when the Cutoff Border Window is applied.

TABLE XIII
WER AND RELATIVE IMPROVEMENT WITH THE PROPOSED METHOD ON NGSW CORPUS (%)

Baseline		52.10	(-)
Oracle Mask	MFR	50.33	(3.39)
	MLLR	45.23	(13.18)
	MFR+MLLR	43.62	(16.29)
	TF-MFR	**49.80**	**(4.42)**
	TF-MFR+MLLR	**43.49**	**(16.52)**
Mask Estimation	MFR	51.54	(1.07)
	MFR+MLLR	46.01	(11.69)
	TF-MFR	**51.47**	**(1.20)**
	TF-MFR+MLLR	**45.19**	**(13.27)**

though the performance is low compared versus the Oracle mask solution.

These results suggest that the proposed missing-feature reconstruction and blind mask estimation methods are applicable to real-life band-limited conditions. In particular, spoken document retrieval system, which is of interest in this paper, must address a wide range of corpora including band-restricted conditions. A multiple-conditioned HMM system could be an alternative method, even though it is beyond the scope of this study, since our focus is on formulating effective "feature compensation" approaches in this paper. However, the HMM for ASR employed by SDR system is obtained through an elaborate and complicated procedure requiring a large training database, so it would not be practical to estimate multiple HMMs with a large number of parameters for different band-limited conditions. Considerable computational expense would be also required to apply a multiple-HMM system to the time-varying band-limited condition which often occurs in real-life spoken documents such as the NGSW corpus.

The ability to select a suitable size for the F1 Area and Cutoff Border windows for the proposed missing-feature method is a practical issue in real-life conditions. We believe that the correlation relationship presented in Section V must be valid in other types of speech corpora as well; however, the correlation range needs to be trimmed according to the sample rate, specification of the employed feature extraction method, and others. Therefore, the presented window size for the TF-MFR method here could be a guideline for suitable selection of the window size for another speech corpus, even it needs to be determined empirically using available pilot test data. It is noted that additive background noise would further degrade performance of our missing-feature reconstruction for band-limited speech, since the acoustic model used for reconstruction is assumed to be trained for a clean condition. Therefore, background noise needs to be addressed using proper schemes such as speech enhancement and/or feature compensation, prior to applying missing-feature reconstruction. Acoustic model adaptation combined with missing-feature reconstruction has been useful for increasing speech recognition performance, as shown in the experiments employing MLLR for the NGSW corpus in this section.

VIII. CONCLUSION

In this paper, we considered the problem of speech recognition of band-limited speech based on missing-feature reconstruction. We proposed a technique to compute the posterior

probability depending only on the reliable components for band-limited condition from our earlier work. This paper proposed an advanced method to utilize the correlation information on the spectral components across both time and frequency axis which are highly correlated with the missing components in the cutoff frequency region. To find the suitable spectral components to be employed, we investigated the correlation characteristics of the spectral components in the reliable frequency band with the cutoff band using a series of preliminary experiments. The experiments showed two parts in the reliable band are mainly correlated with missing components in the cutoff region, which are the area of first formant (F1) and the boundary of the cutoff frequency. Based on our findings, we employed the "F1 Area Window" and "Cutoff Border Window" to incorporate an increased number of reliable components that are determined to be highly correlated with the cutoff frequency band content. To detect the cutoff regions from the incoming speech, our approach to blind mask estimation was also presented, which employs a synthesized band-limited model which does not require a secondary training database.

The experiment to evaluate the performance of the presented methods was accomplished using the SPHINX3 recognizer and TIMIT corpus. We determined the suitable size of F1 Area Window and Cutoff Border Window through a combination of different sizes in the performance evaluation. Experimental results demonstrated that the proposed TF-based missing-feature reconstruction method is significantly effective in improving band-limited speech recognition accuracy. We obtained up to 14.61% average relative improvement in WER on four types of band-restrictions (1.0, 1.5, 2.0, and 2.5 kHz) by employing the proposed TF-MFR method compared to our earlier work [17]. The proposed method employing the blind mask estimation also showed consistent improvement in performance. This consistency in recognition performance proved that our effort to incorporate more substantial correlation information from the spectral components across time and frequency axes is effective in reconstructing the missing spectral components in band-limited speech. The results on actual conditions such as the NGSW corpus also showed the advantage of the proposed time–frequency correlation based method applied to band-restricted speech. Such bandwidth restrictions can be found in a wide range of acoustic conditions within real-life spoken documents that make speech recognition highly challenging.

REFERENCES

[1] N. Morales, D. T. Toledano, J. H. L. Hansen, and J. Colas, "Blind feature compensation for time-variant band-limited speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 1, pp. 70–73, Jan. 2007.

[2] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?," *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 106–111, May 2006.

[3] J. H. L. Hansen, R. Huang, B. Chou, M. Seadle, J. R. Deller, Jr., A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, Sep. 2005.

[4] [Online]. Available: http://www.ngsw.org.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "Data-driven environmental compensation for speech recognition: A unified approach," *Speech Commun.*, vol. 24, pp. 267–285, 1998.

[6] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[7] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[8] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. EUSIPCO*, Sep. 1994, pp. 1178–1181.

[9] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. ICASSP'00*, Jun. 2000, pp. 1847–1850.

[10] P. Jax, "Bandwidth extension for speech," in *Audio Bandwidth Extension*. New York: Larsen and Aarts, Wiley, 2004, ch. 6.

[11] J. Barker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech'01*, 2001, pp. 213–216.

[12] M. Cook, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2001.

[13] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, 2004.

[14] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, Sep. 2004.

[15] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. ICASSP'06*, May 2006, pp. 305–308.

[16] K. J. Palomaki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with missing data automatic speech recognition," *Speech Commun.*, vol. 43, pp. 123–142, 2004.

[17] W. Kim and J. H. L. Hansen, "Missing-feature reconstruction for band-limited speech recognition in spoken document retrieval," in *Proc. Interspeech.06*, Sep. 2006, pp. 2306–2309.

[18] S. Wegmann, P. Zhan, and L. Gillick, "Progress in broadcast news transcription at dragon systems," in *Proc. ICASSP'99*, Mar. 1999, pp. 33–36.

[19] S. S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements To IBM's speech recognition system for automatic transcription of broadcast news," in *Proc. ICASSP'99*, Mar. 1999, pp. 37–40.

[20] S. E. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. C. Woodland, "The Cambridge University Spoken Document Retrieval System," in *Proc. ICASSP'99*, Mar. 1999, pp. 49–52.

[21] B. Ramabhadran, J. Huang, and M. Picheny, "Towards automatic transcription of large spoken archives—English ASR for the MALACH project," in *Proc. ICASSP'03*, 2003, pp. 216–219.

[22] [Online]. Available: http://cdpheritage.org.

[23] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 1999.

[24] W. Kim and J. H. L. Hansen, "Advances in spoken document retrieval for the U.S. collaborative digitization program," in *Proc. IEEE ASRU'07*, Dec. 2007, pp. 687–692.

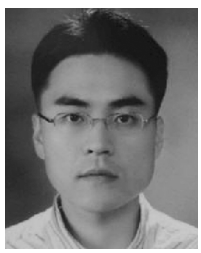
[25] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[26] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.

[27] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," in *Proc. ICASSP'07*, Apr. 2007, pp. 377–380.

[28] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

[29] [Online]. Available: http://cmusphinx.sourceforge.net



Wooil Kim received the B.S., M.S., and Ph.D. degrees in electronics engineering from Korea University, Seoul, Korea, in 1996, 1998, and 2003, respectively.

He has been a Research Assistant Professor in the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, since September 2007. He is also a member of the Center for Robust Speech Systems (CRSS) at UTD. Previously, he was a Research Associate at UTD (2005–2007) and a Postdoctoral Researcher in the electrical and computer engineering, Carnegie Mellon University, Pittsburgh, PA (2004–2005), and Korea University (2003–2004), respectively. His research interests are robust speech recognition in adverse environments, acoustic modeling for large vocabulary continuous speech recognition, and spoken document retrieval.

John H. L. Hansen (S'81–M'82–SM'93-F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is a Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 43 (18 Ph.D., 25 M.S.) thesis candidates, is author/coauthor of 294 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior Modeling* (Springer, 2008), and lead author of the report "The impact of speech under 'stress' on military speech technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and will serve as Technical Program Chair for IEEE ICASSP-2010, Dallas, TX. In 2007, he was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005–2006), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council.