Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 63–87
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A schema-based model for phonemic restoration

Soundararajan Srinivasan [a,*], DeLiang Wang [b]

[a] Biomedical Engineering Center, The Ohio State University, 395 Oreese Laboratories, 2015 Neil Avenue, Columbus, OH 43210, USA
[b] Department of Computer and Engineering Science and Center for Cognitive Science, The Ohio State University, 395 Oreese Laboratories, 2015 Neil Avenue, Columbus, OH 43210, USA

Received 20 January 2004; received in revised form 17 June 2004; accepted 6 September 2004

Abstract

Phonemic restoration is the perceptual synthesis of phonemes when masked by appropriate replacement sounds by utilizing linguistic context. Current models attempting to accomplish acoustic restoration of phonemes, however, use only temporal continuity and produce poor restoration of unvoiced phonemes, and are also limited in their ability to restore voiced phonemes. We present a schema-based model for phonemic restoration. The model employs a missing data speech recognition system to decode speech based on intact portions and activates word templates corresponding to the words containing the masked phonemes. An activated template is dynamically time warped to the noisy word and is then used to restore the speech frames corresponding to the masked phoneme, thereby synthesizing it. The model is able to restore both voiced and unvoiced phonemes with a high degree of naturalness. Systematic testing shows that this model outperforms a Kalman-filter based model.

Keywords: Phonemic restoration; Top-down model; Speech schemas; Computational auditory scene analysis; Prediction; Missing data ASR; Dynamic time warping

1. Introduction

Listening in everyday acoustic environments is subject to various noise interference and other distortions. The human auditory system is largely robust to these effects. According to Bregman (1990), this is accomplished via a process termed auditory scene analysis (ASA). ASA involves two types of organization, primitive and schema-driven. Primitive ASA is considered to be an innate mechanism based on bottom-up cues such as pitch, and spatial location of a sound source. Schema-based ASA use stored knowledge about auditory inputs, e.g. speech patterns, to supplement primitive analysis and sometimes provides the dominant basis for auditory organization. This frequently occurs when parts of speech are severely corrupted by other sound sources.

* Corresponding author. Tel.: +1 614 292 7402.
E-mail addresses: srinivasan.36@osu.edu (S. Srinivasan), dwang@cse.ohio-state.edu (D. Wang).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.09.002

Phonemic restoration refers to the perceptual synthesis of missing phonemes in speech when masked by appropriate intruding sounds on the basis of contextual knowledge about word sequences. In 1970, Warren discovered that when a masker (cough) fully replaced the first "s" of the word "legislatures" in the sentence, "The state governors met with their respective legislatures convening in the capital city," listeners reported the hearing of the masked phoneme (Warren, 1970). When phonemic restoration happens, subjects are unable to localize the masking sound within a sentence accurately; that is, they cannot identify the position of the masking sound in the sentence. When "s" was replaced with silence instead, phonemic restoration was not observed. Subsequent studies have shown that phonemic restoration is dependent on the linguistic skills of the listeners, the characteristics of the masking sound and temporal continuity of speech (Bashford et al., 1992; Samuel, 1981, 1997; Warren and Sherman, 1974).

Fig. 1 depicts a visual analogue of phonemic restoration. Fig. 1(a) shows the fragments of multiple images of the letter 'B' (Bregman, 1981). We cannot perceive 'B' patterns from these fragments. Fig. 1(b) shows the fragments in the presence of an irregularly shaped occluding pattern. We are now able to organize the fragments as parts of the letter 'B'. The standard explanation of this visual phenomenon is in terms of amodal completion—the process of perceptually completing occluded visual surfaces (Nakayama et al., 1995). In this case, it is suggested that the organization is triggered by a top-down representation or a schema for 'B' with the occluder providing bottom-up evidence for the organization (Bregman, 1990).

Auditory organization can also be classified as simultaneous and sequential (Bregman, 1990). Simultaneous organization involves grouping of acoustic components that belong to a sound source at a particular time. Sequential organization refers to grouping of acoustic components of a sound source across time. Phonemic restoration may be viewed as a sequential integration process involving top-down (schema-based) and bottom-up (primitive) continuity. Monaural computational auditory scene analysis (CASA) systems employ harmonicity as the primary cue for simultaneous grouping of acoustic components corresponding to the respective sound sources (Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004). These systems do not perform well in those time–frequency regions that are dominated by aperiodic components of noise. Phonemic restoration is therefore a natural way to introduce other sequential integration cues. Monaural CASA systems currently also lack an effective cue for grouping unvoiced speech. Schema-based grouping in particular, may provide a strong grouping cue for integration across unvoiced consonants. Schemas can be used to gener-

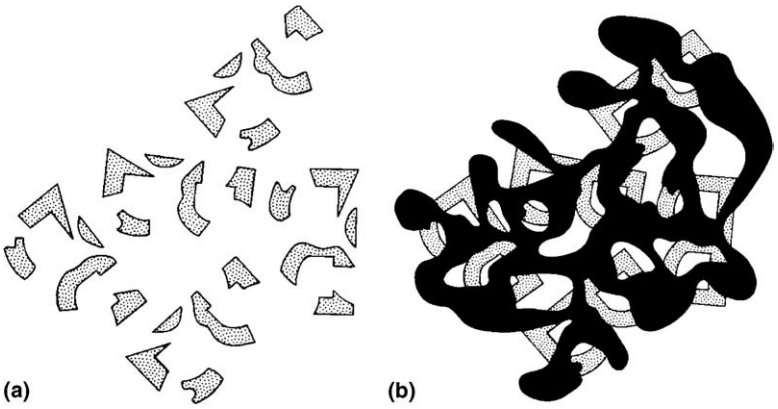

Fig. 1. Visual analogue of phonemic restoration (from Bregman, 1981). (a) Fragments of multiple instances of the letter 'B'. (b) The same fragments of (a) together with an irregularly shaped occluding pattern.

ate expectations for verification by existing bottom-up grouping algorithms and may provide a cue for resolving competition among different primitive organization principles. Schema-based features also inherently bring to the fore top-down aspects like memory and attention into CASA. Additionally, phonemic restoration helps to restore lost packets in speech transmission systems (Perkins et al., 1998; Hassan et al., 2000) and increase the performance of speech enhancement (Nakatani and Okuno, 1999).

Previous attempts to model phonemic restoration have been only partly successful. Cooke and Brown (1993) use a weighted linear interpolation of the harmonics preceding and succeeding the masker for restoration. The later work of Masuda-Katsuse and Kawahara (1999) uses Kalman filtering to predict and track spectral trajectories in those time–frequency regions that are dominated by noise. In its use of temporal continuity for restoration, the Masuda-Katsuse and Kawahara model is similar to that of Cooke and Brown. Note that we use temporal continuity to refer to continuity of individual spectral components. The biggest problem for a filtering/interpolation system for predicting missing speech segments is that temporal continuity of speech frames can be weak or even absent. This typically occurs with unvoiced speech. In the absence of co-articulation cues, it is impossible to restore the missing portions by temporal continuity; in such cases it seems that lexical knowledge must be employed.

Fig. 2 depicts one such situation. In Fig. 2(a), the phoneme /t/ in the coda position of the word 'Eight' possesses no temporal continuity with the preceding phoneme. Thus, when white noise masks the final stop (Fig. 2(b)), this phoneme cannot be recovered by extrapolating the spectrum at the end of the preceding phoneme, /eI/. An automatic speech recognizer (ASR) though could be used to hypothesize the noisy word based on its vocabulary. This hypothesis could then be used to predict the masked phoneme. Ellis (1999) proposes a prediction-driven architecture to hypothesize the information in the missing regions using an ASR. Though the direction is promising, the proposed system is incomplete with few results obtained; in particular, recognition of corrupted speech and resynthesis of speech from the ASR output are largely unaddressed.

In this paper, we present a predominantly top-down model for phonemic restoration, which employs lexical knowledge in the form of a speech recognizer and a sub-lexical representation in word templates realizing the role of speech schemas. The main purpose of our model is speech enhancement. In the first stage of the model, reliable regions of the corrupted speech are identified using a perceptron classifier and a spectral continuity tracker. A missing data speech recognizer (Cooke et al., 2001) is used to recognize the input sounds as words based on the reliable portions of the speech signal. The word template corresponding to the recognized word is then used to "induce" relevant acoustic signal in the spectro-temporal regions

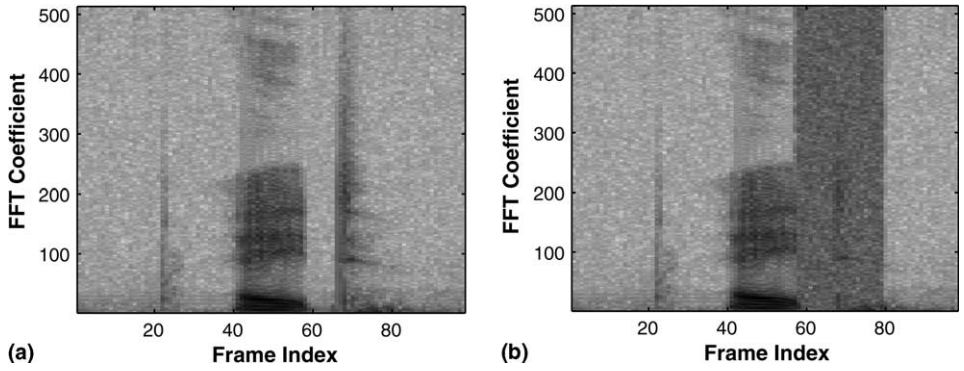

Fig. 2. (a) The spectrogram of the word 'Eight'. (b) The spectrogram obtained from (a) when the stop /t/ is masked by white noise.

previously occupied by noise. Phonemic restoration is typically interpreted as induction based on intact portions of the speech signal and then followed by synthesis of masked phonemes. This synthesis is based on bottom-up confirmation of top-down induction (Warren, 1999). Our approach is consistent with this understanding. The templates are formed by averaging tokens of each word with sufficient spectral detail to permit phonemic synthesis. Finally the induced information is pitch synchronized with the rest of the utterance to maintain the naturalness of restored speech.

The rest of the paper is organized as follows. Section 2 outlines our model. We then describe the details of feature extraction and identification of corrupted regions of the speech signal in Section 3. Section 4 describes the core of our model: The missing data recognizer, word templates and pitch synchronization. The model has been tested on both voiced and unvoiced phonemes and the test results are presented in Section 5. In Section 6, we compare the performance of our model with the Kalman filter based model of Masuda-Katsuse and Kawahara (1999). Finally, conclusion and future work are given in Section 7.

2. Model overview

Our model for phonemic restoration is a multistage system as shown in Fig. 3. The input to the model is utterances with words containing masked phonemes. The maskers used in our experiments are broadband sound sources. Phonemes are masked by adding a noise source to the signal waveform. In the first stage, input waveform with masked phonemes, sampled at 20 kHz with 16 bit resolution, is converted into a spectrogram. A binary mask for the spectrogram is generated in this stage to identify reliable and unreliable parts. If a time–frequency unit in the spectrogram contains predominantly speech energy, it is labeled reliable; it is labeled unreliable otherwise.

The second stage is the missing data ASR (Cooke et al., 2001) based on hidden Markov model (HMM), which provides word level recognition of the input signal by utilizing only the reliable spectro-temporal regions. Thus, the input to the missing data ASR is the spectrogram of the input signal along with a corresponding binary mask. Raj et al. (2000) restore the unreliable units prior to recognition by the missing data ASR. Hence their restoration does not utilize lexical information. Cooke et al. (2001) suggest that for restoration, one can use the maximum likelihood estimate of the output distribution of the winning states. Winning states are obtained during recognition by Viterbi decoding in an HMM-based speech recognizer. We find that such a restoration does not work well and degrades with increasing number of frames that need to be restored. This is not surprising as the missing data ASR has only 10 states to model each word (Section 4.1) and hence state-based imputation is an ill-posed one-to-many projection problem.

On the other hand, template-based speech recognizers use spectral templates to model each word. These templates could be used as a base for restoration. We train a word-level template corresponding to each HMM model in the missing data ASR. Two sets of templates are considered, speaker-independent and speaker-dependent. The speaker-independent template is derived from utterances of speakers different from the test speaker. The speaker-dependent template is derived from those utterances of the test speaker which are not utilized for testing. In certain applications, e.g. tracking of a known speaker, speaker identity may be known. In such applications, these speaker-dependent templates could be applied. Evidence from psychophysical studies suggests that speaker-specific details are utilized by listeners in improving their word identification performance (Goldinger, 1996; Goldinger and Azuma, 2003; Nygaard and Pisoni, 1998).

Based on the results of recognition, word templates corresponding to the noisy words are selected. The time–frequency units of the template corresponding to the unreliable time–frequency units then replace the unreliable units of the noisy word.

A template is an average representation of each word. Thus, the restored phoneme may not be in consonance with the speaking style of the remaining utterance. In order to maintain the overall naturalness of the utterance, we perform pitch based

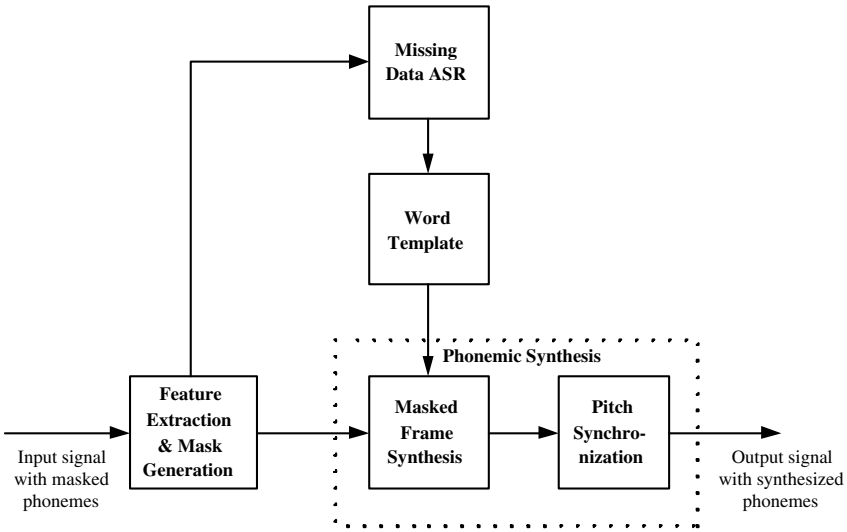

Fig. 3. Block diagram of the proposed system. The input signal with masked phonemes is converted into a spectrogram. A binary mask is generated to partition the spectrogram into its clean and noisy parts. The spectrogram and the mask are fed to the missing data ASR. Based on recognition results, trained word templates are activated corresponding to the words whose phonemes are masked. The masked frames are synthesized by dynamically time warping the templates to the noisy words. These frames are then pitch synchronized with the rest of the utterance. Notice that the information flows bottom-up leading to recognition and then top-down leading to restoration.

smoothing. The last stage of the model is the overlap and add method of resynthesis. Resynthesized waveforms are used for informal listening and performance evaluation.

3. Feature extraction and mask generation

The first stage of our model extracts spectral and cepstral features from the signal and also generates a binary time–frequency (T–F) mask.

3.1. Feature extraction

The acoustic input is analyzed by the feature extraction stage which generates 512 DFT coefficients every frame. Each frame is 20 ms long with 10 ms frame shift. Frames are extracted by applying a running Hamming window to the signal. Finally, log compression is applied to the power spectrum. Thus the input signal is converted into a time–frequency representation, suitable for use by the missing data ASR and subsequent restoration by the synthesis stage. Additionally, the spectral coefficients are converted to cepstral coefficients via the discrete cosine transform (Oppenheim et al., 1999). The cepstral coefficients are sent to the mask generation stage and also to the masked frame synthesis stage.

3.2. Missing data mask generation

The missing data recognizer and the phonemic synthesis stage, both require information about which T–F regions are reliable and which are unreliable. Thus a binary mask, corresponding to the spectrogram, needs to be generated. A T–F unit is deemed reliable and labeled 1 if in this unit, the speech energy is greater than noise energy and otherwise deemed unreliable and labeled 0. Spectral subtraction is frequently used to generate such binary masks in missing data applications (Cooke et al., 2001; Drygajlo and El-Maliki, 1998). Noise is assumed to be long-term stationary and its spectrum estimated from frames that do not contain speech (silent frames containing background noise). In phonemic restoration, noise is usually short-term stationary at best and masks

frames containing speech (corresponding to one or more phonemes). Hence, for phonemic restoration, estimation of noise spectrum followed by spectral subtraction cannot be used to generate the binary mask.

We propose a two-step process for generation of the mask. In all our experiments, we use broadband noise sources as maskers (see Section 5). Hence, as a first step, only a frame-level decision of reliability is made. A frame is labeled 1 if it is dominated by speech, else labeled 0. The individual T–F units of a frame labeled 0 are further analyzed in the second step. The spectral trajectory of the noisy speech signal is tracked using a Kalman filter. We compare the spectral coefficients of the noisy and the filtered signals. If the difference between them is small, we treat these coefficients as reliable and label them 1 and 0 otherwise. Fig. 4(a) shows the spectrogram of the word 'Five'. White noise is used to mask the approximant /j/ in the diphthong /aj/ and the resulting spectrogram is shown in Fig. 4(b). From the figure, we can see that there is a strong spectral continuity (especially for the formants) from the /a/ part to the /j/ part. We seek to recover these regions of spectral continuity and label them 1. Accurate estimation of pitch is difficult, if not impossible, due to the low SNR in the masked frames. Under these conditions, the harmonics of speech in the masked frames may not be reliably recovered through pitch based simultaneous grouping. Hence, the spectral continuity cue is needed to recover the harmonics. Spectral continuity can be tracked and recovered using a Kalman filter (Masuda-Katsuse and Kawahara, 1999).

As the first step, at each frame, we generate two features for classification by assuming noise to be broadband and short-term stationary. The first feature is a spectral flatness measure (SFM) (Jayant and Noll, 1984), defined as the ratio of geometric mean to arithmetic mean of the power spectral density (PSD) coefficients:

$$\mathrm{SFM} = \frac{\left[\prod_{k=1}^{N} S_{xx}(k,n)\right]^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=1}^{N} S_{xx}(k,n)}, \tag{1}$$

where $S_{xx}(k,n)$ is the k^{th} power spectral density coefficient of the noisy speech signal in a frame 'n'. Consistent with the feature extraction stage, N is set to 512. This measure is known to provide good discrimination between voiced frames and other frames (unvoiced and silent) across various speakers in clean speech (Yantorno et al., 2001). Additionally, SFM is related to predictability of speech (Herre et al., 2001; Jayant and Noll, 1984). Specifically, low values of SFM imply high predictability. This property is indirectly used in the second step to refine the mask generated at the frame-level.

The second feature used is the normalized energy (NE). It is defined as

$$\mathrm{NE} = 10\log\left(\frac{\sum_{k=1}^{N} S_{xx}(k,n)}{\max_{n}\sum_{k=1}^{N} S_{xx}(k,n)}\right). \tag{2}$$

As in (1), N is set to 512. Normalization is done to make the energy value independent of the overall signal energy. The log operation is used to expand the range of NE to provide better discriminability amongst frames. Unvoiced, silent and masked frames have high values of SFM but unvoiced and silent frames have low values of NE. Thus, SFM and NE are sufficient to classify a frame as being masked or clean. We use two tokens of isolated word utterances from each of the 50 randomly chosen speakers in the training portion of the TIDigits corpus (Leonard, 1984) to train a perceptron classifier. One phoneme in each utterance is masked by mixing with white noise to yield a local SNR of -1 dB on average. Higher SNR values are not used as the intrusion needs to be strong enough to mask the phoneme (Warren, 1999). Lower SNR values will not affect the model performance. Due to the large variability in the values of SFM and NE in clean speech and distortion in noise, the two classes are found to be linearly inseparable. Hence, we train a one-hidden-layer (2-2-1) perceptron classifier (Principe et al., 2000). The inputs are SFM and NE and outputs are two class labels: 1 (reliable) and 0 (unreliable). The ratio in (1), by definition, is constrained as $0 \leqslant \mathrm{SFM} \leqslant 1$. NE, as defined in (2), is

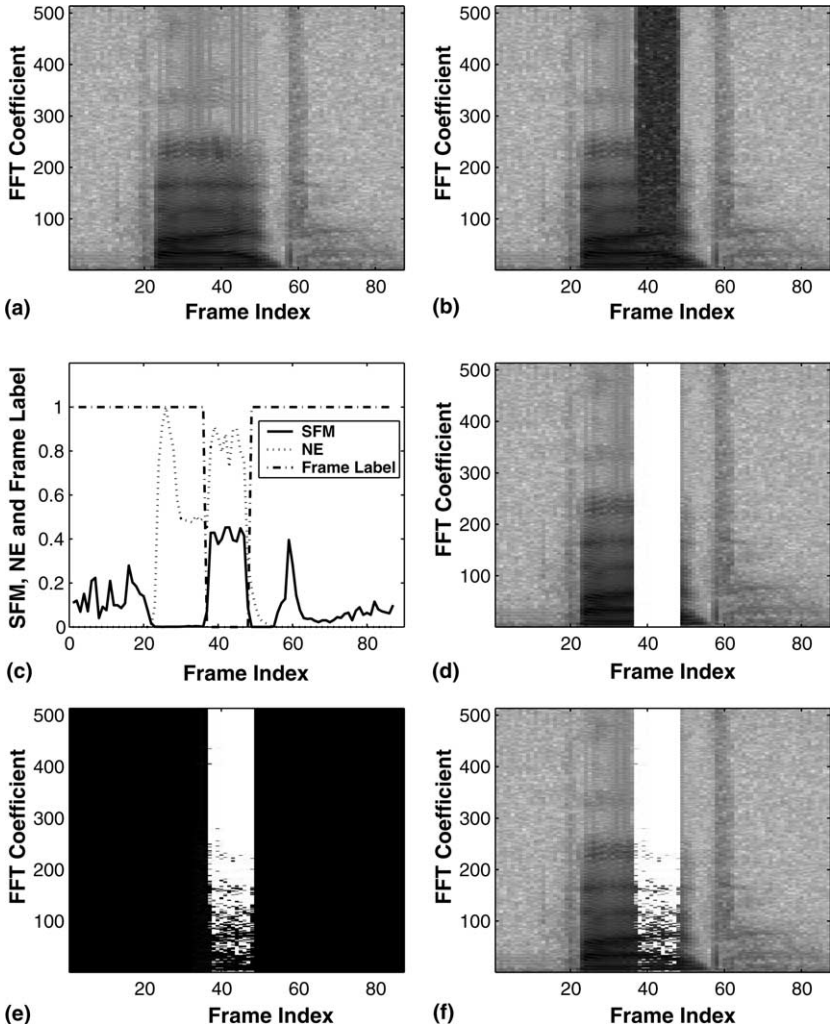

Fig. 4. (a) The spectrogram of the word 'Five'. (b) The spectrogram obtained from (a) when white noise masks the approximant part /j/ in the diphthong /aj/. (c) The distribution of frame-level features and frame-level labels for this utterance (1-reliable and 0-unreliable). Spectral flatness measure (SFM) and normalized energy (NE) are used to generate the frame-level labels. (d) The spectrogram obtained from (b) with only reliable frames. (e) The labels of each T–F unit in the spectrogram. (f) The spectrogram with only reliable T–F units. Unreliable units are marked white.

in the range −80 to 0 dB. The transfer functions of all the neurons are log-sigmoid. The network is trained using backpropagation, optimized by the Levenberg–Marquardt algorithm (Principe et al., 2000). The network is trained for 1000 epochs. Fig. 4(c) shows how the two features, SFM and NE, are distributed for the utterance 'Five' with the masked phoneme /aj/. For the purpose of comparison with SFM, NE is shown without the application of the log operation and the multiplication factor. The spectral flatness measure is high for masked frames, silent frames and frames corresponding to the fricatives /f/ and /v/. The normalized energy though is high only for frames corresponding to the masked phoneme /aj/. Since the masked phoneme is a vowel, the energy in the masked frames is reliably high and we get a perfect frame-level labels of reliability. The resulting spectrogram with only reliable frames is shown in Fig. 4(d).

As the second step, we use Kalman filtering to further analyze the spectral regions in frames labeled 0 by the first stage. For this we adapt the Kalman filter model of Masuda-Katsuse and Kawahara (1999). Kalman filtering is used to predict the spectral coefficients in the unreliable frames from the spectral trajectories of the reliable frames. In the frames labeled as 0 by the first step, we compare the spectral values of the filtered and original noisy signal. If there is true spectral continuity, the magnitude of the difference between the spectral values of the filtered and original signal will be small. This can be restated as a local SNR criterion. Let $S_{ff}(k,n)$ denote the k^{th} power spectral density coefficient of the filtered signal in a frame 'n'. Then each T–F region can be labeled using a threshold δ as

$$\text{label} = \begin{cases} 1 & \text{if } 10\log\frac{S_{ff}(k,n)}{S_{xx}(k,n)-S_{ff}(k,n)} \geqslant \delta \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The choice of δ represents a trade-off between providing more T–F units with reliable labels to the missing data ASR (Section 4.1) and preventing wrong labeling of T–F units (Renevey and Drygajlo, 2001). The optimal value of δ is also dependent on the local SNR (Renevey and Drygajlo, 2001; Seltzer et al., 2003). For simplicity we set δ to be a constant. The value of $\delta = 5$ dB is found to give the best recognition performance on the training data and is used for all the data during testing.

Cepstral coefficients in each order are regarded as a time series and are modeled as a second order auto-regressive (AR) process as suggested by Masuda-Katsuse and Kawahara (1999). This process is predicted and tracked by a Kalman filter and thus used to interpolate the cepstral coefficients in the masked frames from clean frames. The state space model of this system is

$$x(n) = F(n)x(n-1) + Gv(n), \tag{4}$$

$$y(n) = Hx(n) + w(n). \tag{5}$$

In the equations above, $y(n)$ is the observed cepstral coefficient at time-frame n and the filtering problem is to find the information about the state of the system, $x(n)$ (the true value of the cepstral coefficient), at this time. Since the cepstral coefficients follow a second order AR model,

$$F(n) = \begin{bmatrix} a_1(n) & a_2(n) \\ 1 & 0 \end{bmatrix}, \tag{6}$$

where $a_1(n)$ and $a_2(n)$ are the first and second order AR coefficients at time-frame n. We let $G = [1 \quad 0]^{\mathrm{T}}$ and $H = [1 \quad 0]$ as suggested by Masuda-Katsuse and Kawahara (1999). The system white noise $v(n)$ is zero mean with covariance $Q(n)$. The observation white noise $w(n)$ is zero mean with covariance $R(n)$. Hence, the model in (4)–(6), has four unknown parameters that need to be estimated at each frame, $a_1(n)$, $a_2(n)$, $Q(n)$ and $R(n)$.

Let $\theta = (a_1(n), a_2(n), Q(n))$. The log likelihood of the model given θ and initial state mean vector $\overline{x(0)}$ is as follows:

$$l(\theta, \overline{x(0)}) = \sum_{n=1}^{N} \log f(y(n) \mid Y(n-1), \theta, \overline{x(0)}). \tag{7}$$

$$f(y(n) \mid Y(n-1), \theta, \overline{x(0)}) = \mathcal{N}(Hx(n \mid n-1), HV(n \mid n-1)H^{\mathrm{T}} + R(n)),$$

where $Y(n-1) = (y(1), y(2), \ldots, y(n-1))$ (Kato and Kawahara, 1998). The conditional state mean $x(n|n-1)$ and the error covariance $V(n|n-1)$ are estimated by the Kalman predictor:

$$x(n \mid n-1) = F(n-1)x(n-1 \mid n-1),$$

$$V(n \mid n-1) = F(n-1)V(n-1 \mid n-1)F^{\mathrm{T}}(n-1) + GQ(n-1)G^{\mathrm{T}}.$$

The filtered estimates are computed by the Kalman filter.

$$x(n \mid n) = x(n \mid n-1) + K(n)(y(n) - Hx(n \mid n-1)),$$

$$V(n \mid n) = (I - K(n)H)V(n \mid n-1),$$

where $K(n)$ is the Kalman gain computed as

$$K(n) = V(n \mid n-1)H^{\mathrm{T}}(HV(n \mid n-1)H^{\mathrm{T}} + R(n))^{-1}.$$

The parameter θ is updated at each frame by the maximum likelihood estimate conditioned on the present and past observed cepstral values. We

use a numerical subroutine, DALL (Ishiguro and Akaike, 1999), to estimate θ by maximizing (7). The variance of the noise in the observation model, $R(n)$, is set to 1.0 if the cepstral coefficients belong to a frame previously labeled 1. It is set to a high value otherwise. Hence, $R(n)$ acts as a factor that balances the tracking and predicting roles of the Kalman filter. The discrete change in the value of $R(n)$ causes the Kalman filter to switch from a predominantly tracking phase to a predominantly predicting one.

Since processing is performed off-line, the cepstral coefficients at all times are available for processing, enabling a smoothing operation. To mitigate the effects of binary transition in the variance of the observation noise, we perform one step backward Kalman smoothing (Anderson and Moore, 1979). As smoothing additionally uses the cepstral coefficients of the reliable frames, available after the masked frames, it results in a more accurate estimation of the coefficients in the masked frames. Finally, the cepstral coefficients are converted back to spectral coefficients $S_{ff}(k,n)$ via inverse discrete cosine transform (Oppenheim et al., 1999) and exponentiation. The spectral coefficients are used in (3) to generate the labels for each frequency unit. Fig. 4(e) shows the labels generated for the noisy utterance 'Five'. The spectrogram with only the reliable T–F units is shown in Fig. 4(f). It is seen that using Kalman filtering, most formant regions corresponding to the masked part of the diphthong /aj/ are recovered and labeled 1. The regions exhibiting no strong spectral continuity are labeled 0.

4. Recognition and synthesis of masked phonemes

A missing data speech recognizer is used to recognize the input utterance as words based on the T–F units labeled 1. The word template corresponding to the noisy word in the input is then warped to the noisy word. The T–F units of the noisy signal labeled 0 (previously corrupted by noise) are then replaced by the corresponding T–F units of this template. The restored frames are then pitch synchronized with rest of the utterance.

4.1. The missing data speech recognizer

The performance of conventional ASR systems in the presence of acoustic interference is very poor. The missing data ASR (Cooke et al., 2001) makes use of the spectro-temporal redundancy in speech to make optimal decisions about lexical output units. Given a speech observation vector x, the problem of word recognition is to maximize the posterior $P(\omega_i|x)$, where ω_i is a valid word sequence. When parts of x are masked by noise or other distortions, x can be partitioned into its reliable and unreliable constituents as x_r and x_u, where $x = x_r \cup x_u$. The missing data ASR treats the T–F regions labeled 0 as unreliable data during recognition. One can then seek a Bayesian decision given the reliable features. In the marginalization method, the posterior probability using only the reliable features is computed by integrating over the unreliable constituents. Furthermore, if the range for the true value of the unreliable feature is known, it provides bounds (limits) over which the unreliable feature is integrated. This bounded marginalization method is shown to have a better recognition score than the regular marginalization method (Cooke et al., 2001), and is hence used in all our experiments. In missing data methods, recognition is typically performed using spectral energy as feature vectors. If x represents spectral magnitude and sound sources being additive, the unreliable parts can be constrained as $0 \leqslant x_u^2 \leqslant x^2$. This bound provides some additional information about the unreliable features. For example, a low value of x^2 would provide evidence against the high energetic states (e.g. states corresponding to vowels).

We use the 10-state continuous density HMM as suggested by Cooke et al. (2001). The task domain is recognition of connected digits. Thirteen (1–9, a silence, very short pause between words, zero and oh) word level models are trained. All except the short pause model have 10 states. The short pause model has only three states. The emission probability in each state is modeled as a mixture of 10 gaussians with a diagonal covariance structure. Training and testing are performed on the male speaker dataset in the TIDigits database. Note that recognition is performed in the spectral

domain. A HMM toolkit, HTK (Young et al., 2000) is used for training. During testing, the decoder is modified to use the missing data mask for marginalizing the unreliable spectrograhic features. The decoded output from ASR represents the lexical knowledge in our model.

4.2. Word template training by dynamic time warping

A template corresponding to each of the HMMs is trained using DTW. From the training portion of the TIDigits corpus, we randomly select 50 speakers (Section 3.2). Two tokens of isolated word utterances from each of the speakers are used to train each speaker-independent (SI) word template. Assuming all tokens are consistent, we find their warped cepstral average. For this purpose, these tokens are time normalized by DTW. The distortion measure used in the dynamic programming cost function is the cepstral distance. The local continuity constraint used is the Itakura constraint (Rabiner and Juang, 1993). Isolated word utterances corresponding to one test speaker in the test database are used to train a speaker-dependent (SD) template. Utterances of this speaker can then be used for testing. We include SD templates to test whether the use of such templates can further enhance the performance. Together the two sets of templates form word schemas. Fig. 5 shows the SI and SD templates for two words in the lexicon, 'Five' and 'Eight'. The templates in Fig 5(a) and (c) show good representation of formants and frication, including formant transitions into the fricatives. In addition to the formants, the onset and spectra of the burst (corresponding to the stop, /t/) are also adequately represented (Figs. 5(b) and (d)). Also note that the SI templates possess substantial details for use in restoring phoneme spectra, though not as detailed and clean as the SD templates.

4.3. Phonemic synthesis

A maximum of two phonemes are masked in each utterance by mixing with noise to yield a local SNR of -1 dB on average. We use three broadband noise sources: white noise, clicks and coughs. Consistent with experiments on phonemic restoration, all transitions into and out of the phoneme

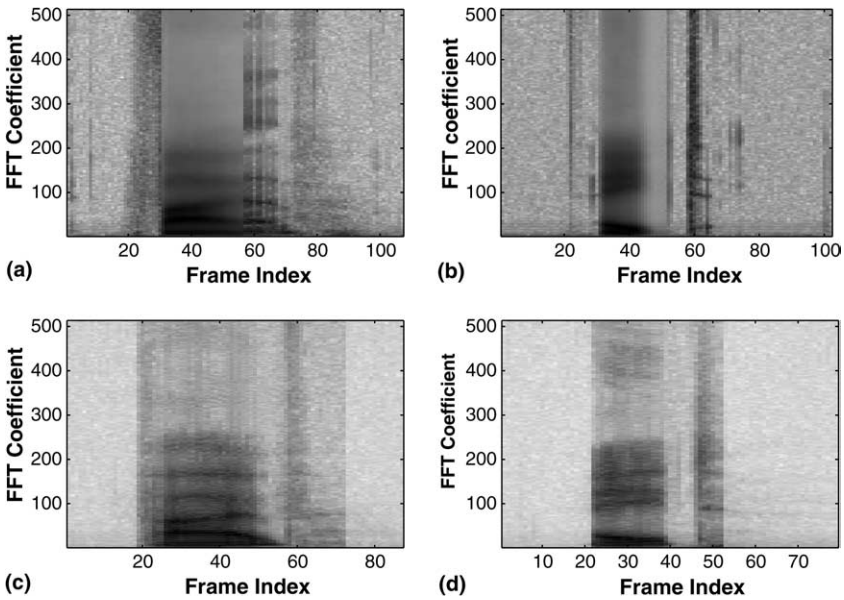

Fig. 5. (a) and (b) The speaker-independent templates of the words 'Five' and 'Eight', respectively. (c) and (d) the corresponding speaker-dependent templates.

are masked too. The signal and the mask are sent to the missing data recognizer which provides the most likely word sequence. Additionally the recognizer provides time end points of the recognized words in the signal. We then choose the word templates corresponding to the noisy word and warp them to the noisy word segment in the input signal by DTW. Specifically, the word template is normalized to span the time end points of the noisy word. The T–F units of the template corresponding to the masked T–F units (with label 0) then replace the masked units. Our restoration in this stage is thus a top-down schema-based process. Recall that some T–F units which exhibit good (bottom-up) spectro-temporal continuity have already been recovered during the mask generation process (Section 3.2). Figs. 6(a) and (b) show the restoration of the masked phoneme /t/ using SI and SD templates respectively. The phoneme is clearly seen to be restored with good spectral quality. Notice that the lack of spectral continuity of the masked phoneme /t/ with the preceding phoneme, has not prevented its effective restoration.

After spectral restoration, the utterance is resynthesized from the spectral coefficients using the overlap and add method (Oppenheim et al., 1999). Since we used a Hamming window during the analysis stage (Section 3.1), we use a rectangular window during the synthesis stage. Also note that the spectral restoration is performed only in the power or magnitude domain. The phase information in the corrupted frames is not restored. Hence, we use noisy phase information during resynthesis.

The word templates are average representation of each word. Hence, the restored information is generally not attuned to the speaking style and the speaking rate of the test utterance. The use of DTW for restoration helps to prevent any significant change in the speaking rate after restoration. To explicitly compensate for co-articulation, the restored frames are manipulated by pitch synchronous overlap and add (PSOLA) techniques, which use interpolated pitch information. In particular, we consider PSOLA (Moulines and Charpentier, 1990) and linear prediction coding (LPC) PSOLA (Moulines and Charpentier, 1988), which are speech synthesis techniques that modify the prosody by manipulating the pitch of the speech signal as required. The former works directly on the speech waveform while the latter on excitation signal of the linear prediction analysis. Praat (Boersma and Weenink, 2002) and a local spectral smoother are used for synchronization.

Fig. 7 shows the pitch track formed by the resynthesized utterance 'Five' after two different stages of restoration. The pitch track formed by the SI restoration after the use of PSOLA is continuous and relatively smooth, indicating the naturalness of the restored phoneme. Restoration without the use of PSOLA yields only a discontinuous pitch track. For comparison, the pitch track of the clean speech signal (before masking of the vowel /aj/) is also shown. We can see that the pitch track after the use of PSOLA is close to the pitch track of the clean speech signal. The LPC-PSOLA technique improves the listening experience compared to PSOLA, but is not better than PSOLA

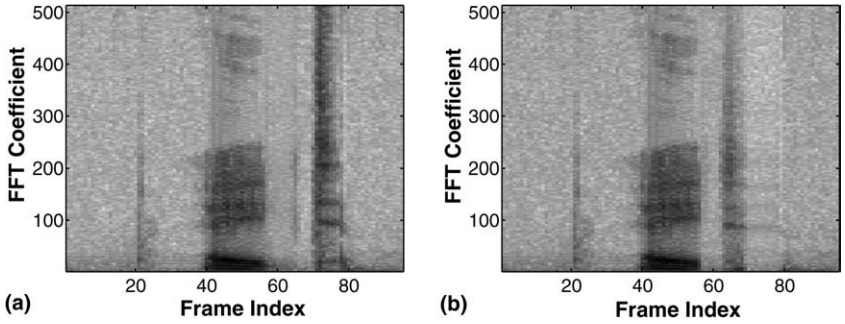

Fig. 6. (a) The restoration of the masked phoneme /t/ in the word 'Eight' using the speaker-independent template. (b) The restoration using the speaker-dependent template.

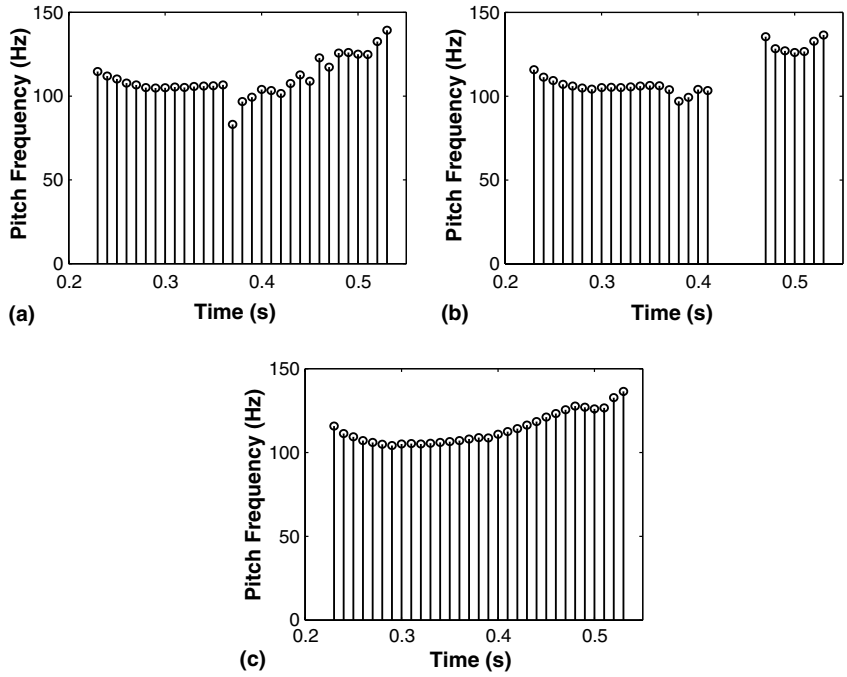

Fig. 7. Comparison of pitch information under various methods of restoration of the diphthong /aj/ in the word 'Five'. (a) and (b) The pitch information extracted from the resynthesized signal using speaker-independent restoration, with and without pitch synchronization respectively. For comparison, the pitch information corresponding to the original clean speech utterance is also shown in (c).

as measured by the objective criteria discussed in Section 5. Consequently only the results of synchronization using the PSOLA technique are used in the assessment of the results. The pitch synchronized utterances are used for informal listening tests and in measuring the performance using the objective criteria outlined in Section 5.

5. Evaluation results

Informal listening by two listeners to the restored signals show that masked voiced and unvoiced phonemes are clearly restored. The listeners also indicate that the restored signal without the use of PSOLA is slightly "hoarse" or "raspy", and the use of PSOLA alleviates this problem greatly. Indeed, they report that the restored signal after the use of PSOLA sounds very natural.

To evaluate the performance of the proposed model objectively, we use two measures: Cepstral and COSH distances. The rms log spectra model the speech spectra very well, but are hard to compute because of the problem of estimating the power spectral density accurately (Stoica and Moses, 1997). The related cepstral and the COSH distances are much easier to compute as they can be derived directly from the AR coefficients of speech and thus avoid the power spectral density estimation problem (Gray and Markel, 1976). The cepstral distance is the most commonly used distortion measure in speech recognition (Rabiner and Juang, 1993). The COSH distance provides the most accurate estimate of spectral envelope of real speech (Wei and Gibson, 2000). Additionally the cepstral distance bounds the rms log spectral distance from below and the COSH distance from above (Gray and Markel, 1976).

The cepstral distance measures the log spectral distance between the original clean signal and the phonemically restored signal:

$$d_C = \sqrt{\left[(C_{1,0} - C_{2,0})^2 + 2\sum_{n=1}^{K}(C_{1,n} - C_{2,n})^2\right]}, \quad (8)$$

where $C_{1,n}$ are the cepstral coefficients derived from AR coefficients of the original signal and $C_{2,n}$ are the corresponding coefficients of the phonemically restored signal. We set $K = 20$. Additionally, the COSH distance (Gray and Markel, 1976) between the power spectra of the two signals is computed. Specifically, let ps_1 and ps_2 denote the power spectra of the original signal and the phonemically restored signal respectively. The COSH distance is defined as

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\cosh\left(\log\left(\frac{\mathrm{ps}_1}{\mathrm{ps}_2}\right)\right) - 1\right\}\mathrm{d}\theta.$$

The distance can be calculated conveniently in its discrete form as

$$\frac{1}{2N}\sum_{n=1}^{N}\left(\frac{\mathrm{ps}_1(\omega_n)}{\mathrm{ps}_2(\omega_n)} + \frac{\mathrm{ps}_2(\omega_n)}{\mathrm{ps}_1(\omega_n)} - 2\right). \quad (9)$$

Consistent with the feature extraction stage, N is set to 512.

Three different classes of phonemes are considered for restoration: vowels, voiced and unvoiced consonants. The vowels possess strong temporal continuity. The spectral continuity of some voiced consonants, e.g. /l/, changes smoothly but faster than vowels. Unvoiced consonants, especially stops, do not have good temporal continuity (Stevens, 1998). We use 100 tokens of isolated word utterances from the training portion of the TIDigits corpus to train each speaker-independent word template. The two isolated word utterances (for each word) of the test speaker are used to train each speaker-dependent template. The remaining 55 utterances of the test speaker form the test set. In choosing the number of isolated word utterances used to derive a speaker-dependent template, we are limited by the number of utterances available for the test speaker in the TIDigits database. Using additional utterances for training would reduce the number of utterances available for testing. The noise sources used for masking are white noise, clicks and coughs. Fig. 8 shows the magnitude spectra of the noise sources. White noise is spectrally flat as shown in Fig. 8(a). The

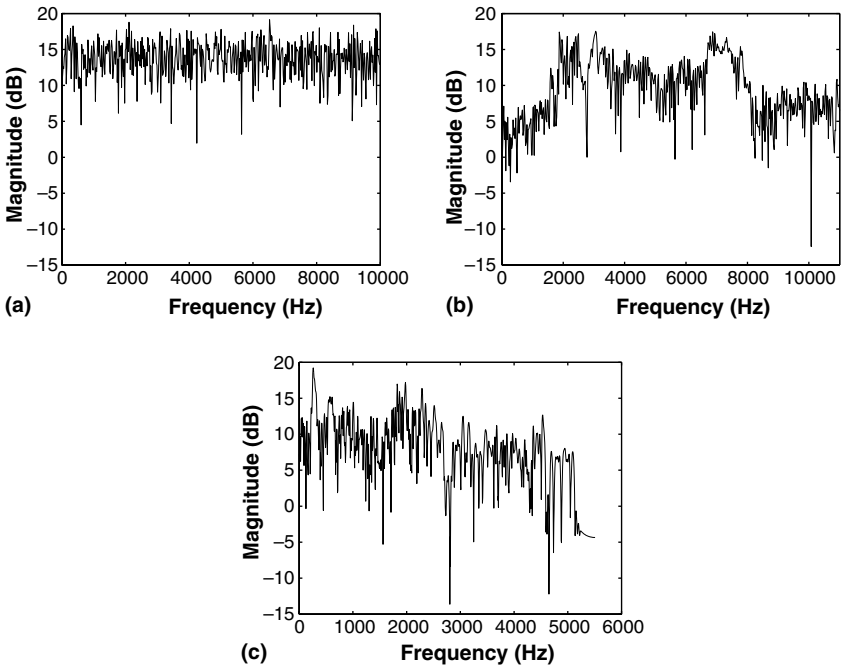

Fig. 8. Long-term spectra of the noise sources. (a) The spectrum of white noise. (b) The spectrum of a click. (c) The spectrum of cough.

spectra of a click and cough deviate in varying degrees from the spectral flatness assumption of the noise sources. The spectrum of a click in Fig. 8(b) shows some narrow peaks in the mid-and high-frequency regions. From Fig. 8(c), we can see that the spectrum of cough exhibits narrow peaks in low-and mid-frequency regions. As stated previously, phonemes are masked by overlaying them with each noise source at a local SNR of −1 dB. The length of burst in each noise source is varied to yield the desired masking of the phoneme. As the duration of a click is typically shorter than that of a phoneme, clicks are repeated to form a click train of duration equal to that of the phoneme being masked. Note that we do not consider forward masking effects in this study. Some amount of noise energy leaks into both the preceding and succeeding phonemes due to the use of the short-time Fourier transform. This effect though does not extend beyond two frames on either side of the phoneme being masked and does not cause recognition degradation.

Fig. 9 shows the performance of our model as measured by the aforementioned objective criteria with white noise as the masker, using the speaker-dependent and the speaker-independent templates. The left column shows the average cepstral distance and the right column shows the average COSH spectral distance between the original and the phonemically restored signals. For comparison, the distances between the clean and the noisy signals are also shown. In the top row we display the results of restoration for vowels. The middle row gives the results for voiced consonants, and the bottom row for unvoiced consonants. The results shown are the average of all signals in each

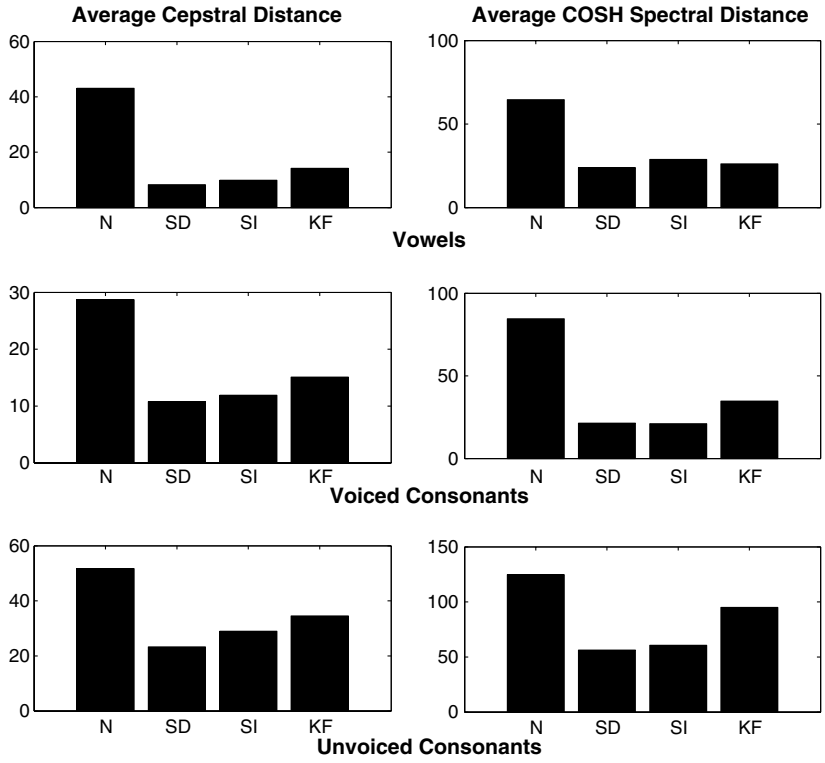

Fig. 9. Performance of the proposed method for phonemic restoration, with white noise as the masker. N refers to the distance of the noisy speech signal from the clean signal. SD refers to the performance of our model with speaker-dependent templates and SI with speaker-independent templates. The left column shows the average cepstral distance and the right column the average COSH spectral distance. The top row shows the results corresponding to vowels, the middle row voiced consonants, and the bottom row unvoiced consonants. For comparison, the results of the Kalman filter model (KF) described in Section 6, are also shown.

class in the test set. The data exclude those signals which are incorrectly recognized by the missing data ASR; recognition accuracy is 89.9%. To amplify the differences between various methods of restoration, the distance measures in Fig. 9 are plotted to different scales for the three different classes of phonemes. If a phoneme is perfectly restored, the distances of the restored signal from the original clean signal are 0 in both measures. Low values of the distance measures after the restoration of voiced phonemes indicate high quality synthesis. The restoration of the unvoiced consonants, especially with the use of speaker-dependent templates, is also good. Note that the performance is similar across both the measures. As evident from the figure, the overall performance of the model with speaker-independent template is not significantly worse than that with speaker-dependent template. The two listeners who participated in informal listening tests report improved sound quality with the use of speaker-dependent template.

Fig. 10 shows the corresponding performance with clicks as the masker. With the use of clicks as the masker, restoration of vowels is slightly better compared to that with white noise but the restoration of voiced consonants is slightly worse. The performance in restoring unvoiced consonants is similar to that with white noise. From Fig. 10, we can also see that clicks are less effective in masking phonemes than white noise, as is evident from the corresponding distances of the noisy speech signals from the original clean signals. The accuracy of the missing data recognizer is 89.2% with clicks as the masker.

Fig. 11 shows the corresponding performance of our model with cough as the masker. Vowels are restored to very high quality. The performance in restoring consonants is similar to that with clicks. Comparing Figs. 10 and 11, we can see that cough is a weaker masker than clicks, especially for voiced phonemes. The accuracy of the missing data recognizer is 92.9% with cough as the masker.

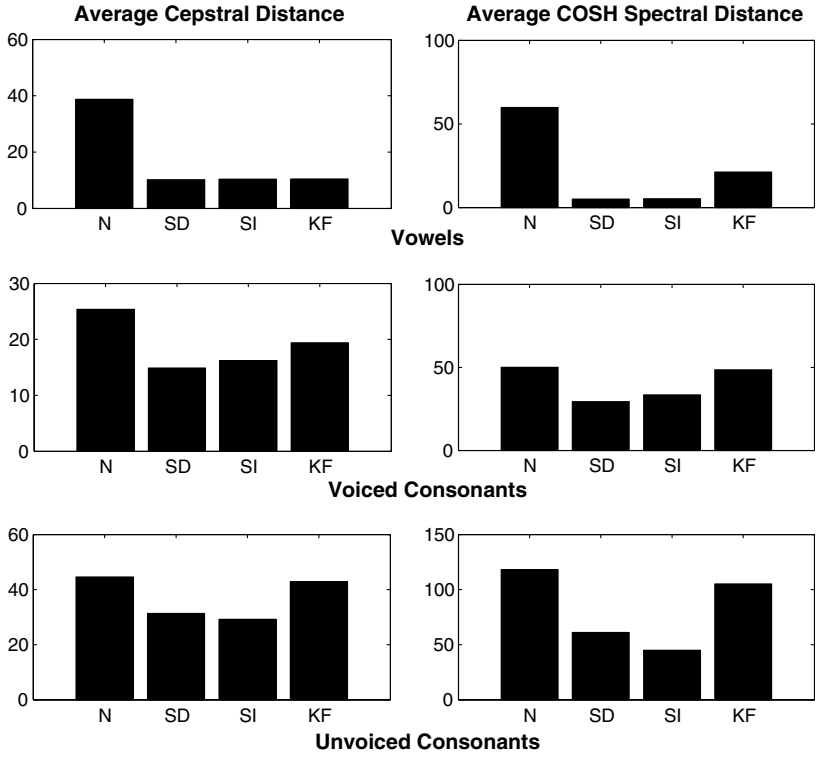

Fig. 10. Phonemic restoration results with clicks as the masker. See Fig. 9 caption for notations.

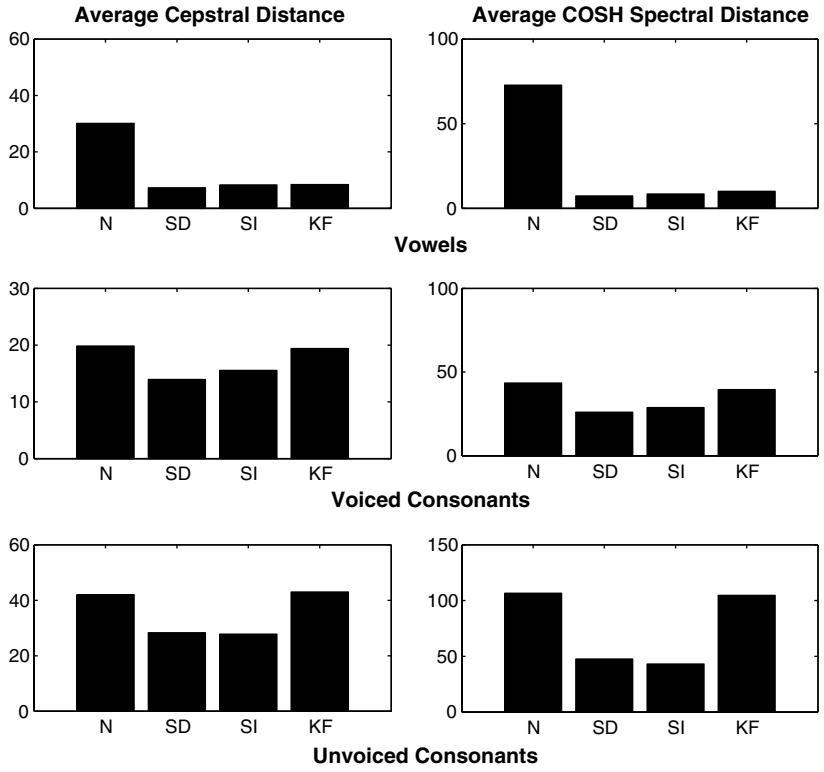

Fig. 11. Phonemic restoration results with cough as the masker. See Fig. 9 caption for notations.

The results also indicate that the performance in restoring consonants is best when white noise acts as the masker. This is not surprising; the perceptron classifier used for frame-level labeling of reliability is trained with white noise as the masker (Section 3.2) and hence performs best on the subset of the test signals which use white noise for masking too. As indicated by the COSH spectral distance, the performance in restoring vowels is better when clicks and cough are the maskers than when white noise is the masker. This indicates that the spectral tracking and smoothing operations are most effective for clicks and cough. This also illustrates that the distance in (9) is more sensitive to the smoothing action than that in (8). Finally, the performance is better when cough is used as the masker than when clicks are used. This might be due to cough being a weaker masker of speech than clicks, especially for voiced phonemes. In summary, the results indicate that the model is able to restore all classes of phonemes, with a spectral quality close to that of the original signal.

5.1. Contribution of spectro-temporal continuity and PSOLA to restoration

Our model of phonemic restoration has three contributing parts; bottom-up spectro-temporal continuity based restoration, top-down schema-based restoration, and pitch synchronization using PSOLA. In order to examine the contribution of each part in detail, we evaluate the performance of our system without one of these parts. First, the use of T–F masks of reliability based on spectro-temporal continuity results in an increase in the accuracy of recognition. Accuracy with only frame-level labels is 86.2% with white noise as the masker, 89.1% with cough as the masker and 86.3% with clicks as the masker. This is because the missing data ASR when using frame-level masks for decoding (Section 4.1), treats all fre-

quency units in a frame labeled 0 as unreliable. The additional recovery of reliable T–F units in a frame labeled 0 increases the accuracy by 3.46% on average, or decreases the error rate by 27.6%. Since PSOLA is applied on the restored frames, it does not affect the recognition results.

We next examine the effects of spectro-temporal continuity and PSOLA on the distance measures of (8) and (9). We select one of the masking noise sources, white noise, for illustration. Fig. 12 shows the influence of spectro-temporal continuity on the performance of our model. Similar to Fig. 9, the two distances in Fig. 12 are plotted to different scales for different classes of phonemes. This helps to amplify the differences in the performance of our model with and without the use of spectro-temporal continuity. Restoration of all classes is almost always better with the use of T–F masks of reliability based on spectro-temporal continuity. The biggest gain occurs in the case of restoration of vowels. This is as expected because the vowels possess the strongest spectro-temporal continuity.

We next examine the effect of PSOLA. Though our explicit motivation for using PSOLA is to provide pitch synchronization, it also affects the spectrum of the synchronized frames and hence affects the two distance measures. Fig. 13 shows the influence of PSOLA on the performance of our model. The performance is almost always better with the use of PSOLA. As observed with the use of spectro-temporal continuity, the biggest gain occurs in the case of restoration of vowels. The periodicity property of vowels is less corrupted by the addition of masking sources, compared to properties of consonants. Hence the use of PSOLA, which utilizes interpolated pitch information, works best for vowels. We also evaluate the performance without the use of either PSOLA or spectro-temporal continuity to examine the contribution of schema-based restoration alone. Note that the

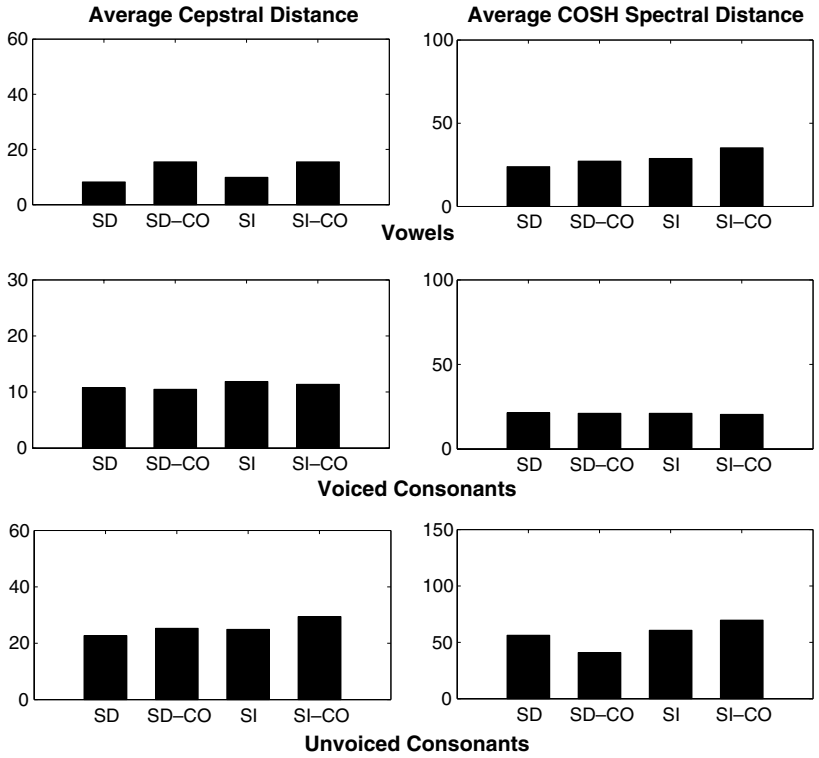

Fig. 12. Influence of spectro-temporal continuity (CO) on the performance of the proposed method in restoring phonemes masked by white noise. "–CO" refers to the performance of our model without the use of spectro-temporal continuity.

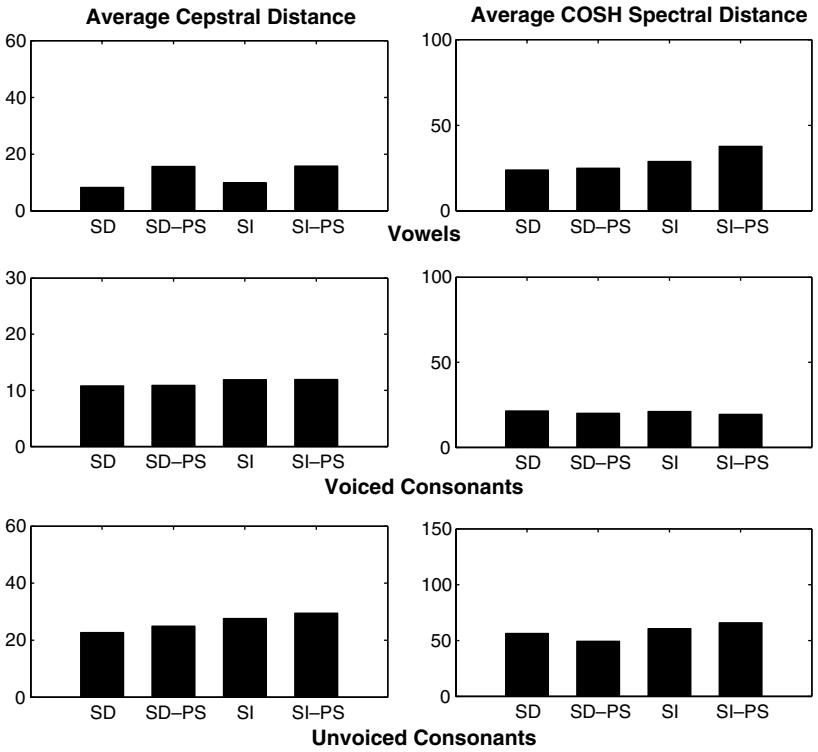

Fig. 13. Results of excluding PSOLA (PS) after restoration. "−PS" refers to the performance of our model without the use of PSOLA.

effects of PSOLA and spectro-temporal continuity are not always additive. Fig. 14 shows the combined influence of PSOLA and spectro-temporal continuity on the performance of our model. The performance is always better with the use of both spectro-temporal continuity and PSOLA. The biggest gain occurs in the case of restoration of vowels due to the aforementioned reasons. Figs. 12–14 together show that the contribution of spectro-temporal continuity and PSOLA to restoration are much smaller compared to the contribution of schema-based restoration.

5.2. Results with ideal binary masks

To reveal the full potential of the proposed model and additionally evaluate our mask generation methods, we test our model with the use of ideal frame-level and T–F binary masks. We again use white noise as the masker for illustration. The performance with ideal frame-level binary masks is shown in an earlier study (Srinivasan and Wang, 2003). An ideal frame-level mask assigns 1 to those frames that have stronger speech energy and assigns 0 otherwise. Recognition accuracy is 87.5% with ideal frame-level masks, a reduction in error rate of 9.4%. Fig. 15 shows the performance of our model using the estimated and ideal frame-level masks. Notice that the performance with the use of estimated masks is close to that with the use of ideal masks in the case of unvoiced consonants while the difference is higher for the restoration of voiced phonemes. This is probably due to SFM of noisy frames not being consistently high enough at the SNR considered in this study.

We now consider the performance with the use of ideal T–F binary masks. An ideal T–F binary mask is obtained from (3) by substituting the power spectral density coefficient of the clean speech signal for the power spectral density coefficient of the filtered signal. Recognition accuracy is 92.6% with ideal T–F masks, a reduction in error rate of 26.7%. Fig. 16 shows the performance of our model using the estimated and ideal frame

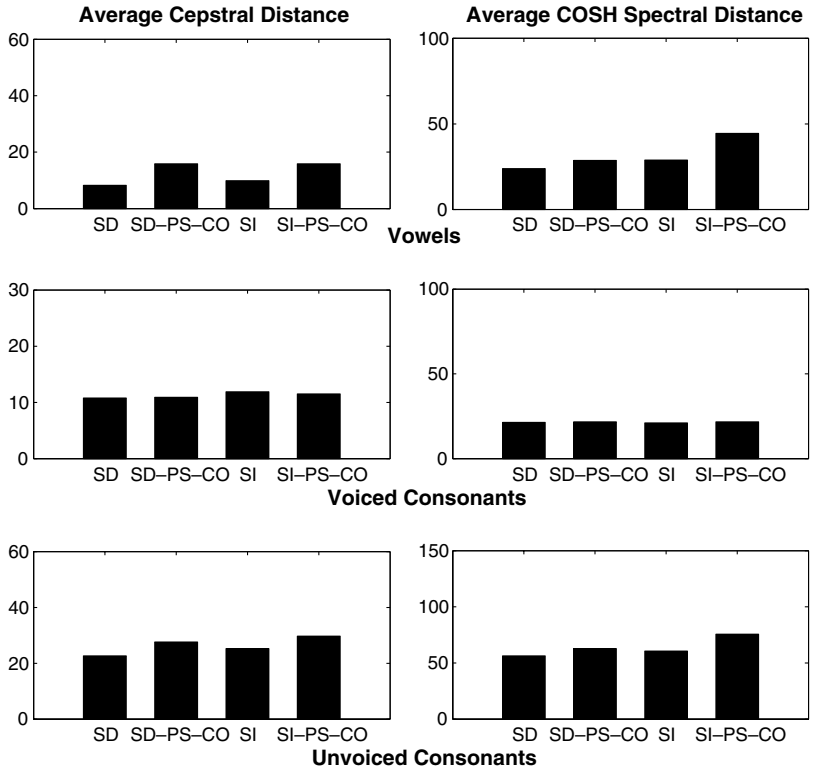

Fig. 14. Results with only schema-based restoration of phonemes masked by white noise. ''−PS−CO'' refers to the performance our model without the use of either PSOLA or spectro-temporal continuity.

T–F masks. As shown by the reduction in error rate, the performance improvement is significant with the use of ideal T–F masks when compared to the performance with the use of estimated T–F masks. This is probably due to a number of factors, including tracking by Kalman filtering not being perfect and the use of a constant value for δ. Also note that all classes of phonemes are restored to a very high quality, when using the ideal T–F masks, highlighting the potential of our approach.

6. Comparison with a Kalman filter model

We compare the performance of our model with the Kalman filter based model of Masuda-Katsuse and Kawahara (1999), which is a systematic study on phonemic restoration and produces good results. They use cepstral tracking with Kalman filtering according to the model in (4) and (5) to predict and restore the masked frames. The variance of the noise in the observation model of (5) is estimated to be proportional to the reliability of results from a previous simultaneous grouping process (based on the harmonicity cue) for the voiced speech signal. This strategy can not be employed when speech additionally contains unvoiced components. For the purpose of comparison with our model, we therefore use the same values for this variable as described in Section 3.2. Additionally, as described in our mask generation stage, we perform one step backward Kalman smoothing. Figs. 9–11 show the performance of the Kalman filter for various classes of restored phonemes.

Under both objective criteria discussed in Section 5, our method outperforms the Kalman filtering model significantly. Notice that except in restoring vowels, our model outperforms the

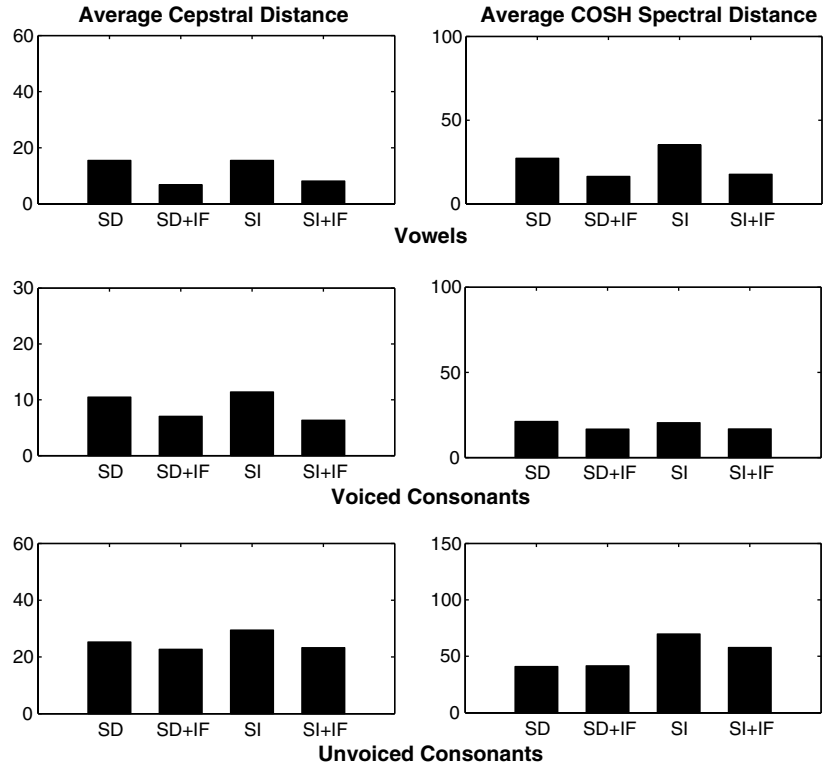

Fig. 15. Results with ideal frame-level masks. The above figure compares the restoration performance of the proposed method using estimated and ideal frame-level (IF) masks. "+IF" refers to the performance of our model using ideal frame-level masks.

Kalman filter model even without the use of PSOLA (Figs. 9 and 13). Similarly, except in restoring vowels, the performance of our model is better with the use of frame-level masks alone (Figs. 9 and 15). Note that vowels are effectively restored by the Kalman filter with sufficient spectral quality, but the restoration may not be very natural. Fig. 17(a) shows the resulting pitch track after restoration of the approximant part /j/ in the diphthong /aj/ in the utterance 'Five' using the Kalman filter model. The pitch track is discontinuous. This illustrates that the spectral magnitude restoration by Kalman filtering alone may reduce the naturalness of speech, just as the spectral magnitude restoration by our model without the use of PSOLA (see Section 4.3).

Unvoiced consonants have weak spectro-temporal continuity with neighboring phonemes and need prior knowledge for their restoration. Hence, our method performs substantially better than the Kalman filter model in restoring them. Fig. 18(a) shows the results of restoration of the unvoiced stop consonant /t/ using the Kalman filter model. As there is no spectro-temporal continuity between this phoneme and the preceding phoneme, the Kalman filter model is unable to restore the stop consonant. The rapid change in the spectrum causes inaccurate estimation of the AR parameters and hence tracking by the Kalman filter breaks down. The performance of our method in restoring voiced consonants is also superior to that of the Kalman filter. The performance of the Kalman filter model improves when clicks and cough are used as maskers (Figs. 10 and 11). This shows that errors in the identification of the noisy regions affects our model slightly more than it does the Kalman filter model. Our model restores only those frames which are labeled unreliable in the mask generation stage (Section 3.2). Kalman filter affects the information in not only the frames marked 0

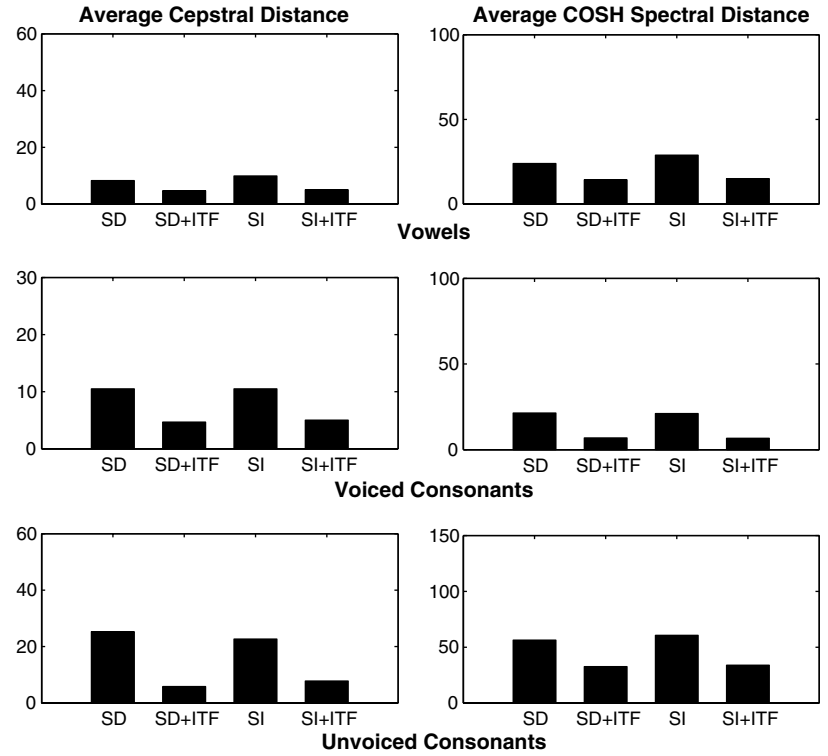

Fig. 16. Results with ideal time–frequency masks. The above figure compares the performance of the proposed method in restoring phonemes using estimated and ideal time–frequency (ITF) masks. "+ITF" refers to the performance of our model using ideal time–frequency masks.

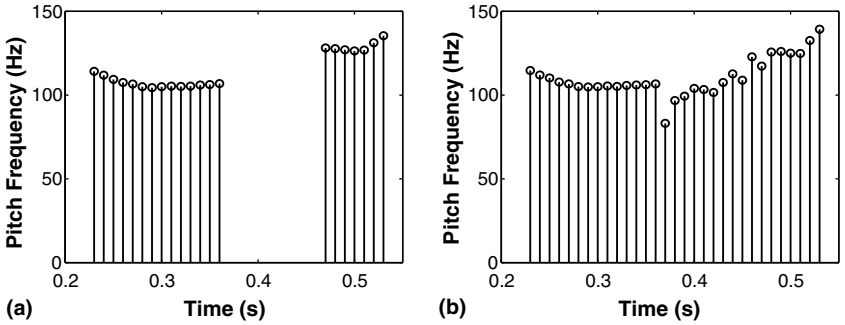

Fig. 17. Resulting pitch information after the restoration of the approximant part /j/ in the diphthong /aj/ in the word 'Five'. (a) The pitch information extracted from the resynthesized signal using the Kalman filter model. For comparison, the pitch information extracted from the resynthesized signal using speaker-independent restoration, with pitch synchronization using PSOLA is also shown in (b).

but also the neighbors of such frames. This is due to the smoothing action of the Kalman filter. Thus, if the neighbor of an unreliable frame is noisy and the mask generation stage mislabels it as 1, then the backward Kalman smoothing reduces the noise in this frame too.

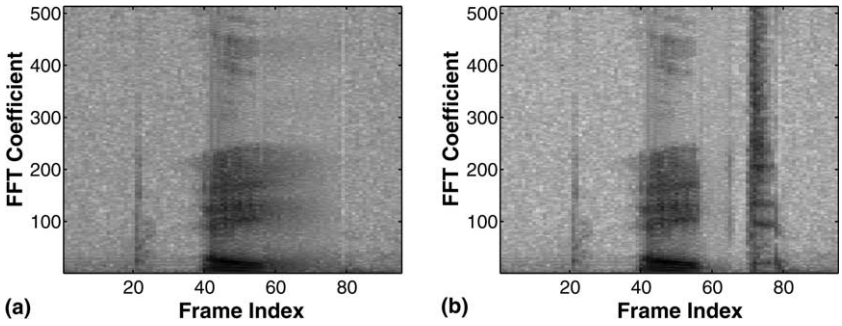

Fig. 18. (a) The restoration of the masked phoneme /t/ in the word 'Eight' by the Kalman filter model. For comparison, the restoration using the speaker-independent (SI) template, is also shown in (b).

7. Discussion

We have presented a schema-based model for phonemic restoration, which performs significantly better than a Kalman filtering model. As stated earlier, the problem for any filtering/interpolation method occurs when the speech spectrum changes rapidly. Hence, such methods perform best for voiced phonemes (especially vowels) and worst for unvoiced consonants. Models based on temporal continuity cannot restore a phoneme that lacks continuity with its neighboring phonemes. Our model is able to restore such phonemes by top-down use of word schemas. Hence, for phoneme reconstruction, we suggest that learned schemas should be employed. Such schemas represent prior information for restoration.

Our model also considers bottom-up continuity in restoration by tracking and filtering the cepstral coefficients. This is similar to the sequential grouping process in the model of Masuda-Katsuse and Kawahara (1999). The difference primarily is in the use of filtered output. Specifically, their model uses the filtered output in all frequency units of a noisy frame. Their approach works well when speech is fully voiced. When speech additionally contains unvoiced consonants, the filtered output may be significantly different from the desired output. In contrast, our model predicts which frequency units in a noisy frame, after filtering, might be close to the desired output and uses only those units for bottom-up restoration.

A system using a speech recognizer for restoration has been described previously by Ellis (1999). His study, however, does not address key issues concerning recognition of masked speech, identification of dominant speech regions in the noisy speech input, and resynthesis of speech (from ASR output labels) for restoration of noisy speech regions. Our model utilizes bottom-up properties of noise to identify the noisy regions in the input signal and applies missing data techniques for recognition based on reliable regions. The use of missing data ASR results in high accuracy of recognition, critical for any system using a speech recognizer for restoration. The use of dynamically time warped templates (based on results of recognition) for restoration followed by pitch synchronization results in high fidelity of the resynthesized phonemes.

Our model of phonemic restoration addresses sequential integration using both bottom-up spectral continuity and top-down schemas. We have shown that the use of bottom-up spectral continuity increases the recognition accuracy and given the recognition results, the top-down use of schemas enhances the original noisy signal for possible use in the following applications. The model can be used in conjunction with existing, predominantly bottom-up, CASA systems to recover masked data and, especially, to group unvoiced speech with voiced speech. Schemas, when activated, can provide top-down construction in these systems. The model may also be used for restoring lost packets in mobile and Internet telephonic applications. Though the motivation behind masking entire phonemes is to be consistent with experiments on phonemic restoration (Warren, 1999),

real-world noise may corrupt only parts of a phoneme or several phonemes at the same time. Our model can handle these conditions well as long as the masking of the speech data does not cause recognition errors. This is because the system neither makes use of the knowledge that a complete phoneme is masked nor knows the number of masked phonemes.

The model is able to simulate certain aspects of phonemic restoration by humans. First, the spectral quality of restored phonemes by our model is close to that of phonemes in clean speech. This is consistent with the observation that perceptually restored phonemes are indistinguishable from real ones (Warren and Obusek, 1971). Second, our schema-based model depends on correct recognition of the word containing the masked phoneme. Hence no improvement in recognition accuracy accrues due to restoration. This is in accordance with findings that phonemic restoration does not enhance intelligibility of words lacking sentential context (Bashford et al., 1992; Miller and Licklider, 1950).

On the other hand, studies have shown that phonemic restoration can help improve intelligibility of sentences (Bashford et al., 1992; Verschuure and Brocaar, 1983; Warren, 1999). We have not addressed this issue in this paper. A small improvement in recognition results is obtained due to restoration of some unreliable T–F units by utilizing bottom-up spectral continuity. Further increase in recognition accuracy might be obtained by increasing the role of bottom-up cues in our model. The energy in the unreliable T–F units plays an important role in phonemic restoration (Samuel, 1981). The spectral shape of the noise is related to its ability to mask a phoneme. There is also an optimal level of noise energy which results in most effective phonemic restoration (Bashford et al., 1992; Warren, 1999). However, the missing data ASR employed here treats all these as counter-evidence for recognition of certain models. A more effective use of the information in the masked regions could help increase the accuracy of the ASR. For example, the information in the masked regions may be used to score a select number of recognizer-generated hypotheses of the missing phoneme.

The model has been currently tested on digit sequences, not on meaningful sentences. To extend to sentences, one would expand the missing data recognizer to include a language model to provide constraints based on syntax and other high-level information (Young, 1996). In this view, meaningful sentences provide contextual information to improve automatic speech recognition, which is otherwise made more difficult by the use of less constrained vocabularies than that of digits.

There are different views on whether human phonemic restoration involves "top-down" synthesis of the masked phoneme. While Warren et al. (1994) suggest that phonemic restoration represents auditory induction through top-down synthesis, others argue that actual synthesis does not occur and identification of the missing phoneme is the end result of phonemic restoration (Bregman, 1990; Repp, 1992). Our model explicitly aims to synthesize the spectral information (and subsequently the waveform) of a masked phoneme, and hence our model is in accordance with the auditory induction explanation.

Our method of estimating the mask for missing data recognition is relatively simple, as it is based on only 2 frame-level and 1 intra-frame features, and masks may be more accurately estimated using a large number of features (Seltzer et al., 2000). How to generate a binary mask for missing data recognition, when maskers are band-limited and last longer than a single phoneme, also needs to be addressed. Future work will attempt to alleviate these problems by integrating the model with existing CASA systems (see e.g., Hu and Wang, 2004). The distribution of spectral tokens in words such as "Eight" may have more than one mode. Robust training of templates may not be adequate for such words. Template training by clustering should further enhance the ability of the generated templates to handle the variability in speaking style. Also, our model is based on recognition and hence not applicable when recognition fails. Combining recognition with top-down restoration and bottom-up cues should help address this problem. When recognition is successful, the top-down use of schemas can supplement bottom-up enhancement. With online detection of recognition failures (Huang et al., 2003), bottom-up processing

may be prominently applied when recognition fails. More generally, schema-based restoration and bottom-up segregation should probably interact in some iterative manner.

Acknowledgments

We thank M. Cooke, H. Kawahara and I. Masuda-Katsuse for their assistance in helping us implement their models. We also wish to thank the two anonymous reviewers for their constructive suggestions/criticisms. A preliminary version of this work was presented in 2003 EUROSPEECH. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (FA9550-04-1-0117).

References

Anderson, B.D.O., Moore, J.B., 1979. Optimal Filtering. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Bashford, J.A., Riener, K.R., Warren, R.M., 1992. Increasing the intelligibility of speech through multiple phonemic restorations. Percept. Psychophys. 51, 211–217.

Boersma, P., Weenink, D., 2002. Praat: Doing Phonetics by Computer, Version 4.0.26. Available from: <http://www.fon.hum.uva.nl/praat>.

Bregman, A.S., 1981. Asking the "what for" question in auditory perception. In: Kubovy, M., Pomerantz, J.R. (Eds.), Perceptual Organization. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 99–118.

Bregman, A.S., 1990. Auditory Scene Analysis. The MIT Press, Cambridge, MA.

Brown, G.J., Cooke, M.P., 1994. Computational auditory scene analysis. Comp. Speech Lang. 8, 297–336.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34, 267–285.

Cooke, M.P., Brown, G.J., 1993. Computational auditory scence analysis: exploiting principles of perceived continuity. Speech Commun. 13, 391–399.

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environment with combined spectral subtraction and missing data theory. In: Proc. ICASSP '98, Vol. 1, pp. 121–124.

Ellis, D.P.W., 1999. Using knowledge to orgnaize sound: the prediction-driven approach to computational auditory scene analysis, and its application to speech/non-speech mixtures. Speech Commun. 27, 281–298.

Goldinger, S.D., 1996. Words and voices: episodic traces in spoken word identification and recognition memory. J. Exp. Psychol. Learn. 22, 1166–1183.

Goldinger, S.D., Azuma, T., 2003. Puzzle-solving science: the quixotic quest for units in speech perception. J. Phonetics 31, 305–320.

Gray, A.H., Markel, J.D., 1976. Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (5), 380–391.

Hassan, M., Nayandoro, A., Atiquzzaman, M., 2000. Internet telephony: services, technical challenges, and products. IEEE Commun. 38, 96–103.

Herre, J., Allamanche, E., Hellmuth, O., 2001. Robust matching of audio signals using spectral flatness features. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics '01. pp. 127–30.

Hu, G., Wang, D.L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks 15, 1135–1150.

Huang, C.S., Lee, C.H., Wang, H.C., 2003. New model-based HMM distances with applications to run-time ASR error estimation and model tuning. In: Proc. Eurospeech '03, pp. 457–460.

Ishiguro, M., Akaike, H., 1999. DALL: Davidson's algorithm for log likelihood maximization—a subroutine for statistical model builders. In: Computer Science Monographs, No. 25. The Institute of Statistical Mathematics.

Jayant, N.S., Noll, P., 1984. Digital Coding of Waveforms. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Kato, H., Kawahara, H., 1998. An application of the Bayesian time series model and statistical system analysis for F0 control. Speech Commun. 24, 325–339.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP '84. pp. 111–114.

Masuda-Katsuse, I., Kawahara, H., 1999. Dynamic sound stream formation based on continuity of spectral change. Speech Commun. 27, 235–259.

Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22, 167–173.

Moulines, E., Charpentier, F., 1988. Diphone synthesis using a mulitpulse lpc technique. In: Proc. The Federation of Acoustical Societies of Europe International Conference '88, pp. 47–55.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467.

Nakatani, T., Okuno, H.G., 1999. Harmonic sound stream segregation using localization and its application to speech stream segregation. Speech Commun. 27, 209–222.

Nakayama, K., He, Z.J., Shimojo, S., 1995. Visual surface representation: a critical link between lower-level and higher-level vision. In: Kosslyn, S.M., Osherson, D.N. (Eds.), An Invitation to Cognitive Science. The MIT Press, Cambridge, pp. 1–70.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 335–376.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing, second ed. Prentice-Hall, Inc., Upper Saddle River, NJ.

Perkins, C., Hodson, O., Hardman, V., 1998. A survey of packet loss recovery techniques for streaming audio. IEEE Network 12, 40–48.

Principe, J.C., Euliano, N.R., Lefebvre, W.C., 2000. Neural and Adaptive Systems. John Wiley and Sons, Inc., New York, NY.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition, second ed. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Raj, B., Seltzer, M.L., Stern, R.M., 2000. Reconstruction of damaged spectrographic feautres for robust speech recognition. In: Proc. International Conference on Spoken Language Processing '00. pp. 1491–1494.

Renevey, P., Drygajlo, A., 2001. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. In: Proc. Consistent and Reliable Acoustic Cues for Sound Analysis Workshop '01. pp. 71–74.

Repp, B.H., 1992. Perceptual restoration of a "missing" speech sound: Auditory induction or illusion? Percept. Psychophys. 51, 14–32.

Samuel, A.G., 1981. The role of bottom-up confirmation in the phonemic restoration illusion. J. Exp. Psychol.: Hum. Percept. Perform. 7, 1124–1131.

Samuel, A.G., 1997. Lexical activation produces potent phonemic percepts. Cogn. Psychol. 32, 97–127.

Seltzer, M.L., Droppo, J., Acero, A., 2003. A harmonic-model-based front end for robust speech recognition. In: Proc. Eurospeech '03. pp. 1277–1280.

Seltzer, M.L., Raj, B., Stern, R.M., 2000. Classifier-based mask estimation for missing feature methods of robust speech recognition. In: Proc. International Conference on Spoken Language Processing '00. pp. 538–541.

Srinivasan, S., Wang, D.L., 2003. Schema-based modeling of phonemic restoration. In: Proc. Eurospeech '03. pp. 2053–2056.

Stevens, K.N., 1998. Acoustic Phonetics. The MIT Press, Cambridge, MA.

Stoica, P., Moses, R.L., 1997. Introduction to Spectral Analysis. Prentice-Hall, Upper Saddle River, NJ.

Verschuure, J., Brocaar, M.P., 1983. Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. Percept. Psychophys. 33, 232–240.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10 (3), 684–697.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.

Warren, R.M., 1999. Auditory Perception: A New Analysis and Synthesis. Cambridge University Press, Cambridge, UK.

Warren, R.M., Bashford, J.A., Healy, E.W., Brubaker, B.S., 1994. Auditory induction: reciprocal changes in alternating sounds. Percept. Psychophys. 55, 313–322.

Warren, R.M., Obusek, C.J., 1971. Speech perception and phonemic restorations. Percept. Psychophys. 9, 358–362.

Warren, R.M., Sherman, G.L., 1974. Phonemic restorations based on subsequent context. Percept. Psychophys. 16, 150–156.

Wei, B., Gibson, J.D., 2000. Comparison of distance measures in discrete spectral modeling. In: Proc. IEEE Digital Signal Processing Workshop '00.

Yantorno, R.E., Krishnamachari, K.R., Lovekin, J.M., Benincasa, D.S., Wenndt, S.J., 2001. The spectral autocorrelation peak valley ratio (SAPVR)–a usable speech measure employed as a co-channel detection system. In: Proc. IEEE International Workshop on Intelligent Signal Processing '01. pp. 193–197.

Young, S., 1996. A review of large-vocabulary continuous-speech recognition. IEEE Signal Process. Mag. 13, 45–57.

Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.