
Article:
Remes, U., López, A.R., Juvela, L. et al. (2015) Comparing human and automatic speech recognition in a perceptual restoration experiment. Computer Speech and Language, 35. 14 - 31. ISSN: 0885-2308

https://doi.org/10.1016/j.csl.2015.06.005

Comparing human and automatic speech recognition in a perceptual restoration experiment

Ulpu Remes[a], Ana Ramírez López[a], Lauri Juvela[a], Kalle Palomäki[a], Guy J. Brown[b], Paavo Alku[a], Mikko Kurimo[a]

[a]*Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, PO Box 13000, Espoo, Finland*
[b]*Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom*

Abstract

Speech that has been distorted by introducing spectral or temporal gaps is still perceived as continuous and complete by human listeners, so long as the gaps are filled with additive noise of sufficient intensity. When such *perceptual restoration* occurs, the speech is also more intelligible compared to the case in which noise has not been added in the gaps. This observation has motivated so-called 'missing data' systems for automatic speech recognition (ASR), but there have been few attempts to determine whether such systems are a good model of perceptual restoration in human listeners. Accordingly, the current paper evaluates missing data ASR in a perceptual restoration task. We evaluated two systems that use a new approach to bounded marginalisation in the cepstral domain, and a bounded conditional mean imputation method. Both methods model available speech information as a clean-speech posterior distribution that is subsequently passed to an ASR system. The proposed missing data ASR systems were evaluated using distorted speech, in which spectro-temporal gaps were optionally filled with additive noise. Speech recognition performance of the proposed systems was compared against a baseline ASR system, and with human speech recognition performance on the same task. We conclude that missing data methods improve speech recognition performance in a manner that is consistent with perceptual restoration in human listeners.

Keywords: automatic speech recognition, missing data, observation uncertainties, perceptual restoration, uncertainty propagation

1. Introduction

Human listeners have a remarkable ability to perceptually restore sounds that have been masked by noise. This ability is particularly evident in the perception of speech sounds. If spectral or temporal sections of a speech signal are removed, listeners perceive the resulting stimulus as distorted. However, when the removed speech regions are filled with loud additive noise, listeners commonly perceive the removed acoustic content as being present (Warren, 1970; Miller and Licklider, 1950). It therefore appears that perceptual restoration only occurs if there is sufficient evidence that the speech has been masked by additive noise (Bregman, 1990). Furthermore, perceptual restoration is accompanied by a benefit in terms of speech intelligibility; listeners are better able to repeat speech content when additive noise is introduced in the removed portions (Warren, 1984; Verschuure and Brocaar, 1983; Powers and Wilcox, 1977; Warren et al., 1997).

Repp (1992) proposed that perceptual restoration theories could be categorised into those based on segregation or top-down completion. These two theories differ in the role that they ascribe to the additive noise. Top-down completion theories propose that restoration is an auditory illusion; the additive noise provides evidence that a speech sound has been masked, therefore activating an (illusory) phonetic percept that depends on the speech context (Bregman, 1990). In contrast, segregation theories explain restoration in terms of sound separation; some of the noise energy is used to reconstruct the missing speech at the acoustic level, with the remainder being attributed to an extraneous sound (Warren, 1984).

Motivated by these findings, several authors have proposed computational models of perceptual restoration. Such models are useful both as a means of clarifying the underlying perceptual mechanisms, and also as components of automatic speech recognition (ASR) systems that are robust in noise. Perceptual restoration has been modelled within source separation systems using continuity of spectral change and fundamental frequency (Cooke and Brown, 1993;

Masuda-Katsuse and Kawahara, 1999) and using a prediction-driven approach (Ellis, 1999; Srinivasan and Wang, 2005). These systems were evaluated by assessing the extent to which they could segregate acoustic mixtures and restore acoustic content that was masked by noise. Perceptual restoration has also motivated *missing data* approaches to noise-robust ASR (Cooke et al., 2001). These include techniques that recognise noise-corrupted speech based on observed, incomplete information, and approaches that restore unobserved clean speech information prior to recognition (Barker, 2012; Gemmeke and Remes, 2012).

Despite the close link between missing data ASR and perceptual restoration, we are not aware of any previous work that directly compares human performance with a missing data system. Cooke (2006) compared human and missing data ASR performance in a consonant recognition task, but noise was added to the speech without prior deletion of speech information; hence his task was different to that used to demonstrate perceptual restoration, in which spectral or temporal gaps are filled by noise. A comparison of human and machine performance using stimuli that induce perceptual restoration is therefore the aim of the current paper. Such a comparison can suggest new developments of missing data ASR systems that might bring them closer to human performance; indeed, in the current study it motivated a particular approach to modelling the uncertainty of observed features.

The system proposed here models unobserved clean speech information as a random variable whose full posterior distribution is used in the ASR task. This can be regarded as a top-down completion model, in that speech is recognised based on partial information and unobserved clean speech information is not restored prior to recognition. Two approaches are used to derive information from the occluding noise about the uncertainty in the underlying observations, namely *bounded marginalisation* (Cooke et al., 2001) and *bounded conditional mean imputation* (Faubel et al., 2009) with uncertainty propagation.

We evaluate our missing data ASR system on a perceptual restoration task, in which the goal is to recognise speech utterances in which acoustic content has either been removed, or substituted with additive noise. First, we confirm that perceptual restoration occurs for human listeners using the speech material selected for our study, and that human speech recognition performance is higher in conditions where perceptual restoration occurs. Subsequently, we show that perceptual restoration also occurs in our missing data ASR system, and that this gives it a performance advantage compared to a baseline system when the speech signal is distorted. Finally, we report a direct comparison between the performance of human listeners and our missing data ASR system, on a subset of the speech material that was used in our listening test.

The remainder of the paper is structured as follows. First, we present the missing data system in Section 2 and the perceptual restoration task in Section 3. The experiments that we conducted are then presented in three sections. Transcription tests conducted to evaluate listener performance, and ASR tests conducted to determine whether the proposed system exhibits perceptual restoration, are reported in Sections 4 and 5, respectively. A direct comparison between human and ASR performance is reported in Section 6. Our results are discussed in Section 7, and conclusions are presented in Section 8.

2. Missing data system

Missing data processing includes components for mask estimation and missing data compensation. The former divides noise-corrupted speech data into reliable and unreliable features, whereas the latter compensates for unreliable information. The approach for mask estimation used in the current study is discussed in Section 2.1. The proposed system does not restore the unobserved clean speech features; instead, the unobserved clean speech features are modelled as a random variable whose posterior distribution is used in the acoustic model likelihood calculation. We calculate the distributions with bounded marginalisation and bounded conditional mean imputation which are introduced in Sections 2.2 and 2.3, respectively. Calculation of acoustic model likelihoods based on full distributions is discussed in Section 2.4.

2.1. Mask estimation

Missing data methods compensate for environmental noise that is additive in the time domain. The methods operate on observed speech and additive noise mixtures in a magnitude-compressed spectral domain, where additive noise can be modelled as an occluder (Nádas et al., 1989). The missing data front-end used in the current work operates on log-magnitude-compressed mel-spectral features. We denote the observed speech and additive noise

mixture in channel d of time frame τ in the log-mel-spectral domain by $Y(\tau, d)$. According to the so-called log-max approximation (Nádas et al., 1989; Varga and Moore, 1990) the log-magnitude-compressed observations can be approximated as

$$Y(\tau, d) \approx \max\{X(\tau, d), N(\tau, d)\}, \tag{1}$$

where $X(\tau, d)$ denotes the speech and $N(\tau, d)$ is the additive noise component. Since $Y(\tau, d) \approx N(\tau, d)$ when $N(\tau, d) > X(\tau, d)$, additive noise behaves like an occluder. The additive-noise-dominated components are therefore referred to as *unreliable* observations and speech-dominated components as *reliable* observations.

The current work focuses on top-down completion modelled with clean speech posterior distributions and observation uncertainties, and does not propose a solution to mask estimation. Instead, we assume access to the speech stimuli and additive noise stimuli used to construct the observed data, so that the reliable and unreliable components can be determined by comparing the speech and noise energy in each time-frequency region. Of course, separate speech and noise components would not be available in practice, and therefore a technique would be required to estimate the mask from the observed noisy speech only. A review of such mask estimation techniques is given by Cerisara et al. (2007).

2.2. Bounded marginalisation

The baseline approaches in missing-data based noise-robust ASR are marginalisation and bounded marginalisation (Cooke et al., 1994, 2001). The present work uses bounded marginalisation. We assume that the observed features are represented in the log-mel-spectral domain and have been divided into reliable and unreliable components. The unobserved clean speech feature that corresponds to an observed feature component $Y(\tau, d)$ is modelled as a random variable ξ. Since the log-mel-spectral features used are non-negative, we know *a priori* that the unobserved clean speech features $\xi \geq 0$. Then, the observations provide information as follows. Since reliable observations are assumed to represent clean speech, for a reliable observation we set $\xi = Y(\tau, d)$. In contrast, an unreliable observation represents additive noise. Since the noise is assumed to be more intense than speech, the unobserved clean speech feature cannot exceed the observed value, $\xi < Y(\tau, d)$.

A posterior distribution encompasses the observed and prior information: when $Y(\tau, d)$ is assumed unreliable, the clean speech posterior distribution in channel d in time frame τ is a continuous uniform distribution between 0 and $Y(\tau, d)$. Since a continuous uniform distribution between a lower bound a and an upper bound b has mean $\frac{1}{2}(b - a)$ and variance $\frac{1}{12}(b-a)^2$, the clean speech feature ξ corresponding to an unreliable feature component $Y(\tau, d)$ has mean and variance

$$\upsilon(\tau, d) = \frac{1}{2} Y(\tau, d), \tag{2}$$

$$\Delta(\tau, d) = \frac{1}{12} Y(\tau, d)^2, \tag{3}$$

where $\upsilon(\tau, d)$ denotes the posterior mean and $\Delta(\tau, d)$ the posterior variance in mel-spectral channel d in time frame τ. Since the acoustic models used in the current work are trained on normalised cepstral features, the clean speech posterior cannot be compared to acoustic models in the magnitude-compressed spectral domain as described by Cooke et al. (2001). Instead, the clean speech posterior in the log-mel-spectral domain is used to calculate the clean speech posterior in the acoustic model domain, and acoustic model likelihoods are calculated as discussed in Section 2.4. Previous work on cepstral-domain marginalisation has proposed a similar approach, in which cepstral-channel weights were used to suppress information from unreliable components (Häkkinen and Haverinen, 2001).

2.3. Bounded conditional mean imputation

Bounded marginalisation and bounded conditional mean imputation (BCMI) assume that the unobserved clean speech features ξ are constrained between 0 and $Y(\tau, d)$. The difference between bounded marginalisation and BCMI is that the latter utilises prior information about statistical dependencies between the clean speech features. The statistical dependencies are represented as a Gaussian mixture model (GMM) whose parameters are trained on clean speech data. A model trained on D-dimensional log-mel-spectral feature vectors captures dependencies between spectral channels, but does not model temporal dependencies between clean speech components. However, a GMM

can capture short-term temporal dependencies if the features are processed in windows that span multiple time frames, as demonstrated in Remes et al. (2011) and González et al. (2013).

The current work evaluates frame and window-based approaches to BCMI. When speech data is processed in multi-frame windows, we concatenate the D-dimensional feature vectors in T consecutive time frames into TD-dimensional vectors $\boldsymbol{y}(t)$. The reliable feature components in window t are represented as subvector $\boldsymbol{y}_r(t)$ and the unreliable components as subvector $\boldsymbol{y}_u(t)$. The unobserved clean speech information that corresponds to $\boldsymbol{y}(t)$ is modelled as a TD-dimensional random variable ζ. Prior information on ζ is represented in the prior distribution $p(\zeta)$, which is a GMM trained on clean speech data, and an approximate clean speech posterior is calculated as follows. First, the reliable observations $\boldsymbol{y}_r(t)$ are used to calculate a conditional distribution $p(\zeta|\boldsymbol{y}_r(t))$. This is a clean speech posterior distribution which does not take into account unreliable observations. The distribution is approximated with a normal distribution $N(\zeta|\boldsymbol{y}_r(t))$ that has a diagonal covariance matrix. The calculations needed to determine $p(\zeta|\boldsymbol{y}_r(t))$ and $N(\zeta|\boldsymbol{y}_r(t))$ are described in (Remes et al., to appear).

Since our acoustic model processes speech data in frames τ, rather than windows t, window-based distributions must be converted to frame-based distributions prior to recognition. In the current work, approximate frame-based clean speech posterior distributions are calculated based on window-based distributions $N(\zeta|\boldsymbol{y}_r(t))$ as follows. The approximate posterior distribution $N(\zeta|\boldsymbol{y}_r(t))$ models each feature component $\zeta(k)$ in window t as an independent random variable with mean $m(t,k)$ and variance $S(t,k)$, where k indexes the component. When features are processed in windows that overlap in time, a clean speech feature ξ in mel-spectral channel d and time frame τ is associated with several window-based posterior distributions that arise from the consecutive windows. Here, the posterior distribution is calculated as an average of the window-based posterior distributions. The distribution is normal, with a mean $m'(\tau,d)$ and variance $S'(\tau,d)$ that are calculated from the means $m(t,k)$ and variances $S(t,k)$ corresponding to the clean speech component in mel-spectral channel d and time frame τ.

The posterior distribution associated with an unobserved clean speech feature ξ is now a normal distribution with mean $m'(\tau,d)$ and variance $S'(\tau,d)$. To encode the information that ξ is constrained between 0 and $Y(\tau,d)$, we box-truncate the approximate posterior distribution. Thus, the approximate posterior distribution associated with ξ becomes a truncated normal distribution. The distribution mean $\upsilon(\tau,d)$ and variance $\Delta(\tau,d)$ are calculated as

$$\upsilon(\tau,d) = m'(\tau,d) + \frac{f(L) - f(U)}{F(U) - F(L)} \sqrt{S'(\tau,d)}, \tag{4}$$

$$\Delta(\tau,d) = S'(\tau,d)\left[1 + \frac{L \cdot f(L) - U \cdot f(U)}{F(U) - F(L)} - \left(\frac{f(L) - f(U)}{F(U) - F(L)}\right)^2\right], \tag{5}$$

where $L = -m'(\tau,d) \,/\sqrt{S'(\tau,d)}$ and $U = (Y(\tau,d) - m'(\tau,d)) \,/\sqrt{S'(\tau,d)}$. Here, $m'(\tau,d)$ denotes the posterior mean and $S'(\tau,d)$ the posterior variance prior to truncation. The probability density function of the standard normal distribution is denoted by $f(Z)$, whereas $F(Z)$ represents its cumulative distribution function. The means and variances are used to calculate the clean speech posterior in the acoustic model domain, and acoustic model likelihoods are calculated as described in the next section.

2.4. Observation uncertainties

Our approach models perceptual restoration as a top-down completion process. This means that the clean speech features are not restored at the acoustic level, but the available information is modelled as a clean speech posterior distribution. The information encoded in posterior distributions is communicated to the ASR system as proposed by Arrowood and Clements (2002) and Deng et al. (2005). We assume that the acoustic model states are modelled as GMMs whose components are indexed by m. A clean speech feature vector in time frame τ in the acoustic model domain is modelled as a random variable $\boldsymbol{x}$ that follows a multivariate normal distribution with mean $\hat{\boldsymbol{\mu}}(\tau)$ and diagonal covariance $\hat{\boldsymbol{\Sigma}}(\tau)$. The likelihood that random variable $\boldsymbol{x}$ pertains to the acoustic model component m is calculated as

$$E\{L(\tau,m)\} = \int N(\boldsymbol{x}\,|\,\hat{\boldsymbol{\mu}}(\tau), \hat{\boldsymbol{\Sigma}}(\tau))\, N(\boldsymbol{x}\,|\,\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})\, d\boldsymbol{x}, \tag{6}$$

where $N(\boldsymbol{x}\,|\,\hat{\boldsymbol{\mu}}(\tau), \hat{\boldsymbol{\Sigma}}(\tau))$ is the clean speech posterior distribution associated with time frame τ and $N(\boldsymbol{x}\,|\,\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$ is the clean speech distribution associated with the acoustic model component m. We further assume that the covariance

matrices $\hat{\boldsymbol{\Sigma}}(\tau)$ and $\boldsymbol{\Sigma}^{(m)}$ are diagonal, in which case the convolution has a closed-form solution,

$$E\{L(\tau, m)\} = N(\hat{\boldsymbol{\mu}}(\tau) \,|\, \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)} + \hat{\boldsymbol{\Sigma}}(\tau)). \tag{7}$$

The modified likelihood calculation compares the mean of the clean speech posterior distribution to the acoustic model distribution with mean $\boldsymbol{\mu}^{(m)}$ and variance $\boldsymbol{\Sigma}^{(m)} + \hat{\boldsymbol{\Sigma}}(\tau)$. Thus, the clean speech posterior mean is interpreted as a feature estimate, and the clean speech posterior variance is introduced into the acoustic model parameters as a dynamic variance component.

The ASR system used in the current work operates on normalised cepstral features (see Section 5.1), whereas the clean speech posterior distributions discussed in the previous sections represent information in a log-mel-spectral domain. Therefore, in order to employ a posterior distribution given by Equations (2)–(3) or (4)–(5) in the ASR system, the distribution is used to calculate a posterior distribution for the clean speech features in the acoustic model domain. The clean speech posterior distribution in the acoustic model domain is modelled as a multivariate normal distribution with mean $\hat{\boldsymbol{\mu}}(\tau)$ and variance $\hat{\boldsymbol{\Sigma}}(\tau)$. The distribution parameters are calculated with piecewise uncertainty propagation (Astudillo et al., 2010), with system-related details as presented in our previous work (Remes et al., to appear).

While the clean speech feature components are assumed to be independent in the log-mel-spectral domain, the feature transformations introduce correlation between components. Distribution parameters needed to model the correlations increase the computational cost in each successive posterior transformation, and thus make observation uncertainties an unattractive alternative compared to other noise-robust speech recognition frameworks. Here, we model the correlation introduced in the cepstral transformation but do not model the inter-frame correlation introduced by cepstral mean subtraction or the inter-frame and intra-frame correlation introduced by differential transformations. The correlations introduced by cepstral mean subtraction and differential transformations that operate on multi-frame windows are ignored to limit the increase in computational cost.

3. Perceptual restoration task

The system described in the previous section was evaluated using a perceptual restoration task in which stimuli were constructed from clean speech utterances and additive noise. Perceptual restoration studies indicate that the restoration effect depends on acoustic cues that relate to the additive noise (Powers and Wilcox, 1977; Samuel, 1981; Warren et al., 1997) and context cues that relate to the speech material (Warren and Sherman, 1974; Verschuure and Brocaar, 1983; Bashford et al., 1992). The current section describes the stimuli and evaluation metric used in the present work. The stimuli include (i) read sentences in which certain spectral or temporal portions have been removed and (ii) read sentences in which additive noise has been introduced in the removed portions. Sections 3.1 and 3.2 discuss the characteristics of the additive noise, specifically the noise type and level. The stimuli used in the perceptual restoration experiments are then described in Sections 3.3 and 3.4, and the evaluation approach is discussed in Section 3.5.

3.1. Additive noise

To construct noisy speech stimuli, we paired each clean speech utterance with a speech-shaped noise sample constructed as described below. Since similarity of spectral shape is known to enhance perceptual restoration (Samuel, 1981), a noise sample was produced for each utterance that matched its spectral shape. Our approach was motivated by the stimuli used in Warren et al. (1997), where the speech stimuli were presented in two narrow spectral channels that were equalised so that the peaks in each spectral band were within ±2 dB(A) of the overall presentation level. This procedure effectively whitens the speech spectrum, so that the removed part between the narrow spectral channels can be replaced with bandpass-filtered white noise. In the present study, we chose to shape the frequency content of the noise to match the speech, rather than to equalise the spectral gains of the clean speech utterances. This ensured both that the speech retained its naturalness, and also that the spectral shape of the speech matched the acoustic models in our ASR system.

In practice, white noise was shaped to each utterance as follows. First, the clean speech utterances were processed with a simple voice activity detection (VAD) algorithm to remove silence frames. The approach used an estimate of the speech energy envelope that was calculated as the absolute value of the speech waveform smoothed with a

moving average filter. The moving average was calculated using a 380 ms triangular window. Envelope values which exceeded a threshold of 0.2 times the standard deviation of the envelope were considered active, and values below the threshold were discarded as silence. Active speech frames were used to estimate the spectral shape of the utterance, which was modelled as a second-order linear prediction (LP) filter. The model order was considered sufficient for capturing the broad spectral shape, whilst also guaranteeing that the utterance-dependent filters did not capture local spectral components such as formants that could bias the speech recognition results. The utterance-dependent filters were applied to white noise to construct utterance-dependent noise samples, which were then combined with the corresponding speech utterance as described below.

3.2. Noise level

Experiments conducted on human listeners indicate that the perceptual restoration effect depends on the level difference between the speech and additive noise stimuli (Powers and Wilcox, 1977; Warren et al., 1997). For this reason, we determined the average level in each clean speech utterance and scaled the additive-noise levels according to the estimated speech levels. The levels were calculated in MATLAB with a virtual sound level meter (Lanman, 2006). The levels were determined on the dB(A) scale, and the average speech level in each utterance was calculated based on the active speech frames, detected as proposed in the previous section. Then, the speech and noise pairs were processed as follows. To prepare data for human listeners, we determined the sound pressure level of a reference noise sample and scaled each clean speech utterance to the same level. Noise samples were then scaled with respect to the speech level. The utterances used in ASR experiments, on the other hand, were not normalised to a specific sound pressure level, but the noise samples were scaled with respect to the speech level in each utterance. This means that the level difference between each speech and noise pair was fixed, and the exact same level differences were used in human and ASR experiments when experiments were conducted on the same speech and noise pairs. The only difference was that the speech level in each utterance was not fixed in the ASR experiments. The exact noise levels used in the human and ASR experiments are provided in Sections 4.3 and 5.3.

3.3. Filtered speech

The perceptual restoration experiments conducted in the current study include both spectral and temporal restoration. Experimental data was prepared based on the clean speech and noise pairs constructed and adjusted as described in the previous sections. For the spectral restoration experiments, the clean speech utterances were filtered with a 1458-point finite impulse response (FIR) bandstop filter with linear phase and $>$ 1000 dB/octave attenuation in the transition bands. The 3 dB cutoff frequencies were 472 Hz and 5680 Hz, and minimum attenuation in the stopband between 505 Hz and 5650 Hz was 80 dB. The additive noise used in the spectral restoration task was constructed in a similar manner. The noise samples paired with the utterances were filtered with a 1457-point FIR bandpass filter with a linear phase and $>$ 1000 dB/octave attenuation in the transition bands. To avoid masking between the speech and noise stimuli, the bandstop and bandpass filter cutoffs were separated at 6 dB cutoff frequencies by one equivalent rectangular bandwidth (ERB) (Edmonds and Culling, 2005). The 3 dB cutoff frequencies in the bandpass filter were 560 Hz and 5030 Hz, and the minimum attenuation in the stopbands below 526 Hz and above 5060 Hz was 80 dB. To construct the speech and noise stimuli used in perceptual restoration experiments, the filtered speech and noise stimuli were added as illustrated in Figure 1.

3.4. Interrupted speech

Interrupted speech stimuli used in temporal restoration experiments were constructed from alternating 200 ms speech and silence segments or 200 ms speech and noise segments. Raised cosine ramps of 10 ms duration were applied to the onset and offset of each segment, and the interrupted speech and noise stimuli were added to construct interrupted speech with additive noise (Figure 2). We also informally tested 50 ms and 100 ms interruptions, but listening to the stimuli indicated that – while the interrupted speech sounded discontinuous without additive noise – there was no reduction in human speech recognition performance. The same observation has been reported in previous studies (Miller and Licklider, 1950; Powers and Speaks, 1973).

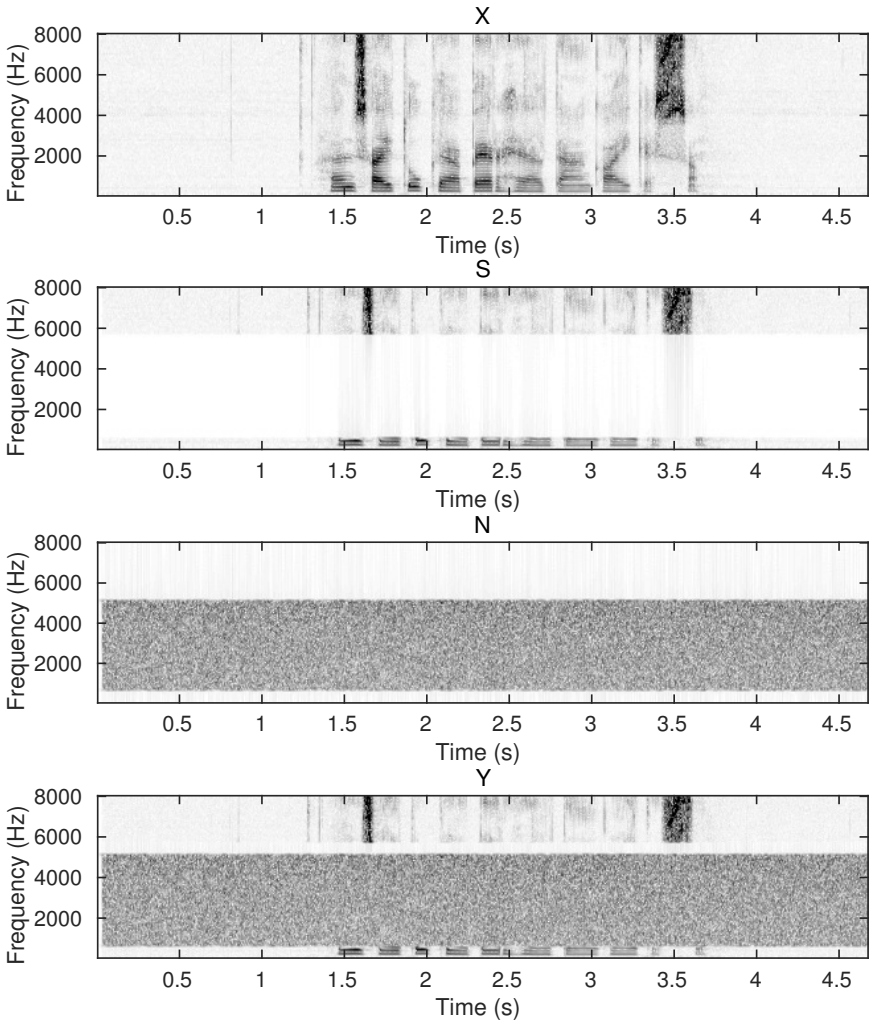

Figure 1: Filtered speech example. A clean speech utterance has been paired with noise, and in this example, the speech and noise have been scaled to the same level. The clean speech utterance in the spectral domain is denoted by X. The utterance is bandstop-filtered to create the speech stimulus denoted by S and the noise is bandpass-filtered to create the noise stimulus denoted by N. The speech and noise stimuli are added to create filtered speech with additive noise $Y = S + N$.

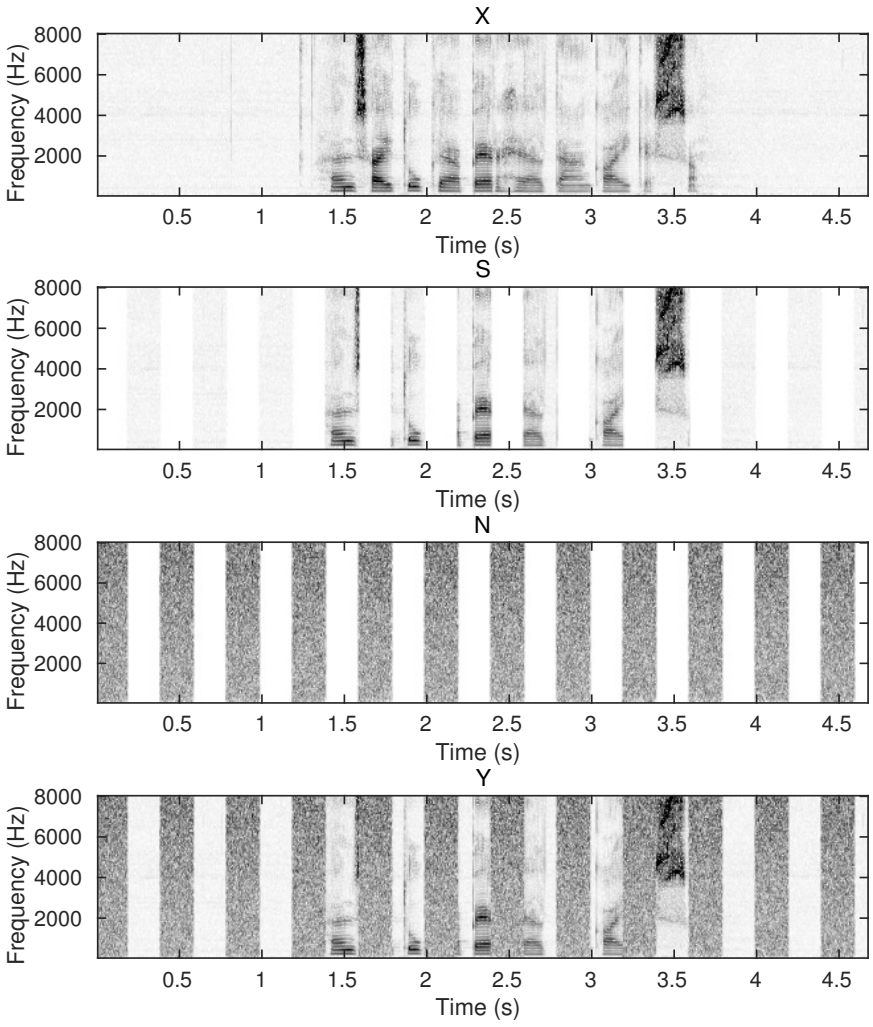

Figure 2: Interrupted speech example. A clean speech utterance has been paired with noise, and in this example, the speech and noise have been scaled to the same level. In the utterance, 200 ms interruptions are introduced to create the interrupted speech stimuli denoted by S. Interruptions are introduced in the noise with an inverse pattern, to create the interrupted noise stimuli denoted by N. The speech and noise stimuli are added to create interrupted speech with additive noise $Y = S + N$.

3.5. Evaluation

To evaluate whether perceptual restoration improves human speech recognition performance, we asked human listeners to transcribe bandstop-filtered and interrupted utterances presented with and without additive noise. The utterances used in this work were read sentences spoken in Finnish. The transcriptions were compared to a reference text and accuracy was measured in letter error rate (LER). Letter error rate is preferred over word error rate because it provides a more sensitive measure of transcription accuracy. Letter errors also approximate phoneme errors, due to the regular sound–letter correspondence in standard modern Finnish. The same measure was also used to evaluate ASR performance. To calculate the LER between a transcription and reference text, both texts are converted into letter sequences that include markers for word boundaries and aligned. Letter error rate is then calculated as

$$\mathrm{LER} = 100 \times \frac{S + D + I}{N}, \tag{8}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions between the transcription and reference sequences, and N is the number of tokens in the reference sequence. Since a transcription sequence can contain more tokens than the reference, letter error rate can exceed 100.

4. Experiment 1

Transcription tests were conducted with human listeners to ensure that perceptual restoration improved listeners' speech recognition performance on the speech material used in the current work. Test participants are described in Section 4.1 and the method is explained in Section 4.2. Then, test data and results are presented in Section 4.3 and Section 4.4, respectively. Two tests were conducted because the first test did not provide conclusive evidence as to whether perceptual restoration improved listener performance in the transcription task. Both tests are described in the following sections.

4.1. Participants

12 listeners participated in the first test and 12 new listeners in the second test. The listeners were native Finnish speakers who were not familiar with the experimental setup. They were offered financial compensation for participation.

4.2. Method

Each test included a familiarisation section and four test sections, visualised as a sequence of modules in Figure 3. A module has an utterance list parameter that determines which utterances are loaded, and a test condition parameter that determines how the utterances are processed. The test conditions evaluated were (i) filtered speech, (ii) filtered speech with additive noise, (iii) interrupted speech and (iv) interrupted speech with additive noise. The utterance lists L1–L4 and test conditions C1–C4 were presented so that each listener heard each list once and each test condition once. Since the utterances in each list were unique, a listener did not hear the same utterance more than once. Instead, the utterance lists and test conditions were allocated across listeners so that each list was transcribed in each condition (Figure 4). Test conditions were presented to the listeners in a non-repeated non-repeating random order.

The test sections were structured as follows. First, example utterances processed according to the current test condition were presented to the listener, and then the listener proceeded to transcribe the test utterances. The example utterances included one male and one female utterance, and the same sentences were used as examples in each section. The test utterances, which the listener transcribed, were loaded based on the input utterance list and the current test condition. The listeners were allowed to hear each utterance once and were encouraged to transcribe the utterances even if they were uncertain about the content (or even if the utterances appeared nonsensical). The utterances were presented to the listeners via Sennheiser HD650 headphones in a sound-attenuated listening booth. The listeners completed the test in 10–15 minutes.

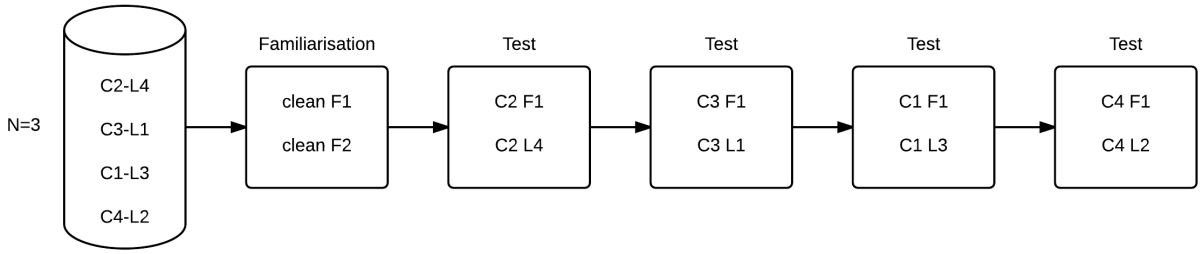

Figure 3: Example test with a familiarisation section and four test sections. The sections are visualised as modules that have a condition parameter and a list parameter. The utterances presented in the familiarisation section are undistorted clean speech. The utterances in list F1 (two utterances) are used as examples and the utterances in list F2 (three utterances) are used to familiarise listeners with the transcription task. The utterances presented in a test module are processed according to certain conditions C1–C4. The order in which the conditions are presented is determined based on the listener number N, and thus varies between listeners. Listener number three is used in the above example, and the condition order in this example is C2–C3–C1–C4. To familiarise listeners with each new condition, sections start with the example utterances F1 processed according to the current condition. Listeners then hear and transcribe the test utterances in a specific list L1–L4.

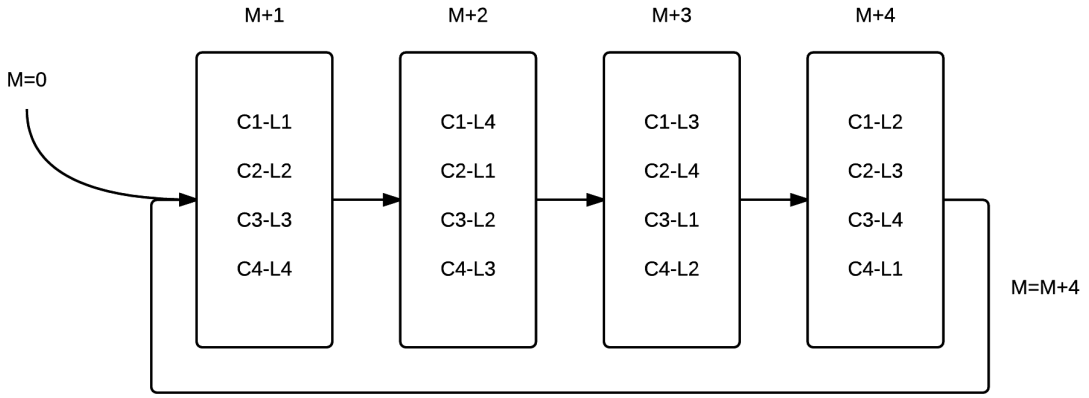

Figure 4: List and condition circulation procedure. Each listener is presented with conditions C1–C4 and lists L1–L4 in a certain combination. *M* is a counter incremented by four each time four new subjects participated the experiment. Four combination sets (rectangles in above diagram) are used. Each combination set contains four unique list and condition pairs, that are fixed across all the experiments. In the test each listener (in groups of four) is assigned to one of the four sets. Note that, while the illustration shows list and condition pairs ordered based on condition index, in the actual experiments the order of list and condition pairs within each combination set is randomized so that a non-repeated random order is obtained for each listener. The circulation procedure ensures that each list is transcribed in each condition three times when 12 listeners participate in the test.

Table 1: Human speech recognition performance in the first test evaluated using the letter error rate. Performance was evaluated on filtered and interrupted speech with and without additive noise. When noise was included, the difference between speech and noise levels was scaled to 0 dB(A).

	Filtered	Interrupted
Speech	50.0	38.5
Speech and noise	37.6	35.7

4.3. Data

The utterances used in the tests were read sentences selected from the Finnish SPEECON database (Iskra et al., 2002). Five utterances were used in the familiarisation and example sections to introduce the test setup and test conditions (Figure 3). These utterances were not used in the final evaluation. The test utterances that were transcribed and evaluated included 32 utterances (2 minutes) that represented 16 male and 16 female speakers. The utterances were divided into lists L1–L4 that were circulated across test conditions as discussed in the previous section. The utterances were hand-selected to ensure that they had a clear meaning and that there were no rare or unusual words that listeners might be unable to recognise. There were on average 5.3 and maximum 6 words in each utterance.

The clean speech utterances were processed to create bandstop-filtered (Section 3.3) and interrupted (Section 3.4) versions with and without additive noise, and a reference noise was used to ensure that the presentation level remained constant across the listeners in one test. The reference noise level was measured in the test setup with an RS 33-2055 sound-level meter operating in slow response mode. In the first test, the reference noise presentation level was adjusted to 60 dB(A) and the speech and additive-noise levels were scaled to the reference noise level (Section 3.2). In the second test, the speech level was scaled to the reference noise level and the additive-noise level was scaled to exceed the reference noise level by 10 dB(A). To compensate for the increase in additive-noise level, the reference noise presentation level was lowered to 55 dB(A).

4.4. Results

The transcription tests were conducted in order to ensure that perceptual restoration led to improved speech recognition performance, for the particular speech material used here. In the first test, listeners were presented with (i) bandstop-filtered or interrupted speech without additive noise and (ii) bandstop-filtered or interrupted speech and noise at the same level. Test results are reported in Table 1. To evaluate whether perceptual restoration enhances listener performance in the transcription task, we compare the LER observed in test conditions with and without additive noise. We note that the LERs calculated based on listener transcriptions depend on the test condition, listener, and utterance. Since a listener never transcribed the same utterance in conditions with and without additive noise, the LERs that we wish to compare are not paired. Hence, statistical significance is determined with the Wilcoxon rank-sum test. Comparison between listener performance in conditions with and without additive noise indicates that human speech recognition performance improves ($p < 0.001$) in the filtered speech condition when additive noise is introduced in the data. While additive noise also appears to improve listener performance in the interrupted speech condition, the difference between listener performance in conditions with and without additive noise is not statistically significant at a significance level of $\alpha = 0.05$.

The first test therefore indicated that additive noise induced perceptual restoration and improved listener performance in the bandstop-filtered speech condition, but listener performance did not improve when additive noise was introduced in the interrupted speech condition. Possible explanations for this finding are that (i) perceptual restoration did not improve listener performance in the interrupted speech condition, or (ii) that perceptual restoration did not occur in the interrupted speech condition. To eliminate the second of these possibilities, we conducted a second listening test in which the additive-noise level was increased to ensure perceptual restoration. The noise level in the second test exceeded the speech level by 10 dB(A). Comparison between listener performance in the filtered and interrupted speech conditions, with and without additive noise (Table 2), indicates that the higher noise level improves human speech recognition performance in both conditions ($p < 0.01$ in both comparisons). We therefore conclude that the aforementioned second explanation cannot be eliminated, and it is indeed possible that perceptual restoration did not

Table 2: Human speech recognition performance in the second test evaluated using the letter error rate. Performance was evaluated on filtered and interrupted speech with and without additive noise. When noise was included, the noise level exceeded the speech level by 10 dB(A).

	Filtered	Interrupted
Speech	52.7	43.0
Speech and noise	38.0	31.7

occur in the interrupted speech condition in the first test. Moreover, we conclude that perceptual restoration can be observed as an increase in transcription accuracy in both test conditions when the noise level is sufficiently high.

5. Experiment 2

To evaluate whether perceptual restoration can be realised as top-down completion in a machine system, we compare the performance of baseline and missing data ASR systems in a perceptual restoration task. The baseline system and missing data system are introduced in Sections 5.1 and 5.2 respectively. The data used in the experiments is explained in Section 5.3, and ASR results are presented in Section 5.4.

5.1. Baseline system

The large-vocabulary continuous speech recognition (LVCSR) system used in the current study processes speech data in 16 ms frames, with a 8 ms overlap between consecutive frames. Each frame is represented by a feature vector of 12 mel-frequency cepstral coefficients (MFCC) and log-compressed frame energy. Feature vectors are normalised by cepstral mean subtraction (CMS), augmented with their first and second order differentials, and decorrelated with the maximum likelihood linear transformation (MLLT). Decorrelated cepstral features are modelled as continuous-density hidden Markov models (HMM) whose states are modelled with at most 100 Gaussian components. State durations are modelled with gamma distributions (Pylkkönen and Kurimo, 2004). Acoustic model parameters are trained on 30-hours of data drawn from the Finnish SPEECON database, consisting of clean speech utterances recorded with a headset in quiet environments (SNR 16–44 dB). The decoder is a time-synchronous beam-pruned Viterbi token-pass system, and the language model is a morph-based growing n-gram model (Hirsimäki et al., 2009) trained on Finnish book and newspaper data with 145 million words. The vocabulary is in practice unlimited, since all words and word forms can be represented with statistical morphs (Hirsimäki et al., 2006).

5.2. Missing data system

The missing data system proposed here operates on uncertain clean speech information represented as a multivariate normal distribution in the acoustic model domain. The system is based on the baseline system introduced in the previous section, which is modified to use observation uncertainties (Section 2.4). Clean speech posterior distributions presented to the system are calculated based on posterior distributions constructed in a log-mel-spectral domain. In the current work, the clean speech posterior distribution is calculated with bounded marginalisation (Section 2.2) or with bounded conditional mean imputation (Section 2.3). BCMI calculates the posterior distribution based on a GMM prior trained on 500 read sentences (52 minutes) included in the acoustic model training data (Section 5.1). Speech data is processed in 1-frame or 5-frame windows, and the statistical dependencies between feature components are modelled as a 5-component full-covariance GMM. The models were initialised with fuzzy c-means and trained with the expectation–maximisation (EM) algorithm implemented in the GMMBAYES toolbox (Kämäräinen and Paalanen, 2005).

The missing data system was evaluated on filtered or interrupted clean speech and additive noise (Section 3). The features that represent the filtered or interrupted clean speech and noise are denoted by $\boldsymbol{S}$ and $\boldsymbol{N}$, respectively, and the observed features are denoted by $\boldsymbol{Y}$. Since we have access to the speech and noise independently, the reliable and unreliable components can be determined based on so-called oracle (*a priori*) information: component $Y(\tau, d)$ is labelled unreliable if $S(\tau, d) < N(\tau, d)$ and reliable otherwise. This means that all feature components are assumed reliable when speech is presented without additive noise, $\boldsymbol{Y} = \boldsymbol{S}$.

Table 3: Level difference between speech and noise (dB(A)) and SNR (dB) for filtered and interrupted conditions with and without additive noise. Additive noise conditions are indicated using the labels LL–HH.

Noise condition	None	LL	L	0	H	HH
Level difference (dB(A))		-20	-10	0	10	20
Filtered SNR (dB)	-2.1	-2.1	-2.2	-3.0	-7.1	-15.4
Interrupted SNR (dB)	3.1	3.1	2.4	-1.3	-9.6	-19.3

5.3. Data

ASR performance was evaluated on 1093 utterances (115 minutes) from 40 speakers (22 female and 18 male) extracted from the Finnish SPEECON database. The utterances were processed to create filtered (Section 3.3) and interrupted (Section 3.4) speech datasets with and without additive noise presented at several noise levels. The level differences between speech and additive noise are labelled in different test conditions as follows. In one condition, speech and noise are scaled to the same level (i.e., the level difference between speech and additive noise is 0 dB(A)). This is labelled as test condition 0. In test conditions L and LL, the noise is attenuated so that the level difference between speech and additive noise is -10 dB(A) in test condition L and -20 dB(A) in test condition LL. In contrast, in test condition H and HH, additive noise is scaled to exceed the speech level so that the level difference between speech and noise is 10 dB(A) in test condition H and 20 dB(A) in test condition HH.

Normally, when noise is added to speech at a level that is low compared to the speech level, observed features correspond to undistorted clean speech features, $\boldsymbol{Y} \approx \boldsymbol{X}$, and when the additive-noise level increases, more and more observations become unreliable until $\boldsymbol{Y} \approx \boldsymbol{N}$. In contrast, in our perceptual restoration task the noise level does not determine the ratio between reliable and unreliable features, because speech information in certain spectrotemporal areas is removed and then *substituted* with additive noise. To evaluate how much acoustic-level information remains in the filtered and interrupted signals with and without additive noise, we can compare the stimuli to undistorted clean speech utterances. The SNR in each test condition is calculated based on the undistorted clean speech utterances and the perceptual restoration stimuli as

$$\mathrm{SNR} = 10 \log_{10} \left[\frac{\sum_n x(n)^2}{\sum_n (y(n) - x(n))^2} \right] \tag{9}$$

where $x(n)$ is the undistorted clean speech amplitude and $y(n)$ is the amplitude of the observed stimulus. The SNR calculated in each test condition is reported in Table 3. While an increase in additive-noise level decreases SNR in test conditions L–HH, SNR cannot be determined based on the additive-noise level, and indeed, additive noise in test condition LL does not decrease SNR.

5.4. Results

The two systems evaluated here are an uncompensated baseline system and a missing data system using bounded marginalisation (BM) and bounded conditional mean imputation (BCMI). BCMI is further evaluated with priors trained on 1-frame windows (BCMI-T1) and 5-frame windows (BCMI-T5). The systems' performance on speech stimuli with and without additive noise at noise levels LL–HH is reported in Table 4. Performance is reported as letter error rate, and statistical significance is tested in pairwise comparisons with the Wilcoxon signed-rank test at a significance level of $\alpha = 0.05$. The uncompensated baseline system and missing data system performance are identical when evaluated on speech without additive noise. However, when the systems are evaluated on speech with additive noise, BM and BCMI systems have better ($p < 0.05$) performance than the uncompensated baseline system. We also observe that BCMI performance is better than BM performance in most test conditions. While there is no conclusive difference between BCMI-T1 and BCMI-T5 when systems are evaluated on filtered speech, BCMI-T5 performance is consistently better than, or comparable to, BCMI-T1 performance in the interrupted speech conditions.

To evaluate whether a particular system models perceptual restoration, we compare system performance in test conditions with and without additive noise. We observe that additive noise improves ($p < 0.001$) baseline system performance in filtered speech condition L and interrupted speech conditions L–LL, while in interruped speech condition 0, the observed difference is not statistically significant. The additive noise level in conditions L–LL is lower than the

Table 4: Automatic speech recognition performance on (a) filtered (b) interrupted speech data with additive noise. Performance is reported in terms of letter error rate, and the additive-noise levels are indicated by LL–HH.

		None	LL	L	0	H	HH
(a)	Baseline	88.9	89.4	86.9	91.7	93.8	95.3
	BM	88.9	77.2	77.8	80.1	83.0	89.9
	BCMI-T1	88.9	78.3	73.0	70.4	72.1	77.2
	BCMI-T5	88.9	79.3	73.0	70.7	70.2	73.6

		None	LL	L	0	H	HH
(b)	Baseline	72.8	59.2	65.3	71.8	74.5	80.9
	BM	72.8	57.6	56.8	55.0	60.1	72.0
	BCMI-T1	72.8	56.7	55.1	53.8	56.6	58.7
	BCMI-T5	72.8	56.5	54.5	52.6	52.3	53.6

speech level. In contrast, when the baseline system is evaluated in test conditions H–HH, where the additive noise level exceeds the speech level, the system performance is worse ($p < 0.01$) than the performance without additive noise. Thus, additive noise does not consistently improve the uncompensated baseline system performance on the proposed stimuli.

In contrast to the baseline system performance, missing data system performance improves with the addition of noise in most test conditions. Additive noise improves ($p < 0.001$) BM performance in test conditions LL–H, and differences between BM performance on stimuli without noise and with intense additive noise (HH) are not statistically significant. BCMI-T1 and BCMI-T5 performance on stimuli with additive noise is consistently better ($p < 0.001$) than BCMI-T1 and BCMI-T5 performance on stimuli without noise. However, while additive noise improves BCMI-T1 and BCMI-T5 performance at all noise levels, performance also depends on the noise level. Pairwise comparison across noise levels indicates that BCMI-T1 performance is best ($p < 0.001$ in all pairwise comparisons) in test condition 0, and BCMI-T5 is best in test conditions 0 and H. Differences in missing-data system performance with and without additive noise are illustrated in Figure 5.

6. Experiment 3

To realise a direct comparison between human and ASR performance in a perceptual restoration task, we evaluated ASR performance on the dataset used in the second transcription test (Section 4.4, Table 2). The evaluations were conducted with the uncompensated baseline system and the missing data system with window-based bounded conditional mean imputation (BCMI-T5). In this evaluation, letter error rates were calculated at the utterance level and compared across human and ASR evaluations.

Human and ASR performance with and without additive noise is reported in Table 5. The human performance rates were determined in the second transcription test (Table 2) and are repeated here for convenience. We note that the relative error reduction is similar for listeners and the BCMI-T5 system when noise is added in both filtered (listeners: 28%, ASR: 24%) and interrupted (listeners: 26%, ASR: 29%) conditions. In contrast, adding noise resulted in increased errors for the baseline system. To evaluate whether ASR system performance correlates with the listener results, Spearman rank-order correlation was calculated between the listener and ASR performance on the same utterance in the same condition. Furthermore, random permutations were applied to the utterance-level data to create a null distribution for the rank correlation in each condition, and to determine statistical significance. Permutation analysis indicates that correlation between the performance of listeners and the uncompensated baseline system is not statistically significant at a significance level of $\alpha = 0.05$. In contrast, listener and BCMI-T5 performance are correlated with $\rho = 0.22$ ($p < 0.005$) in the filtered speech condition and with $\rho = 0.41$ ($p < 0.001$) in the interrupted speech condition.

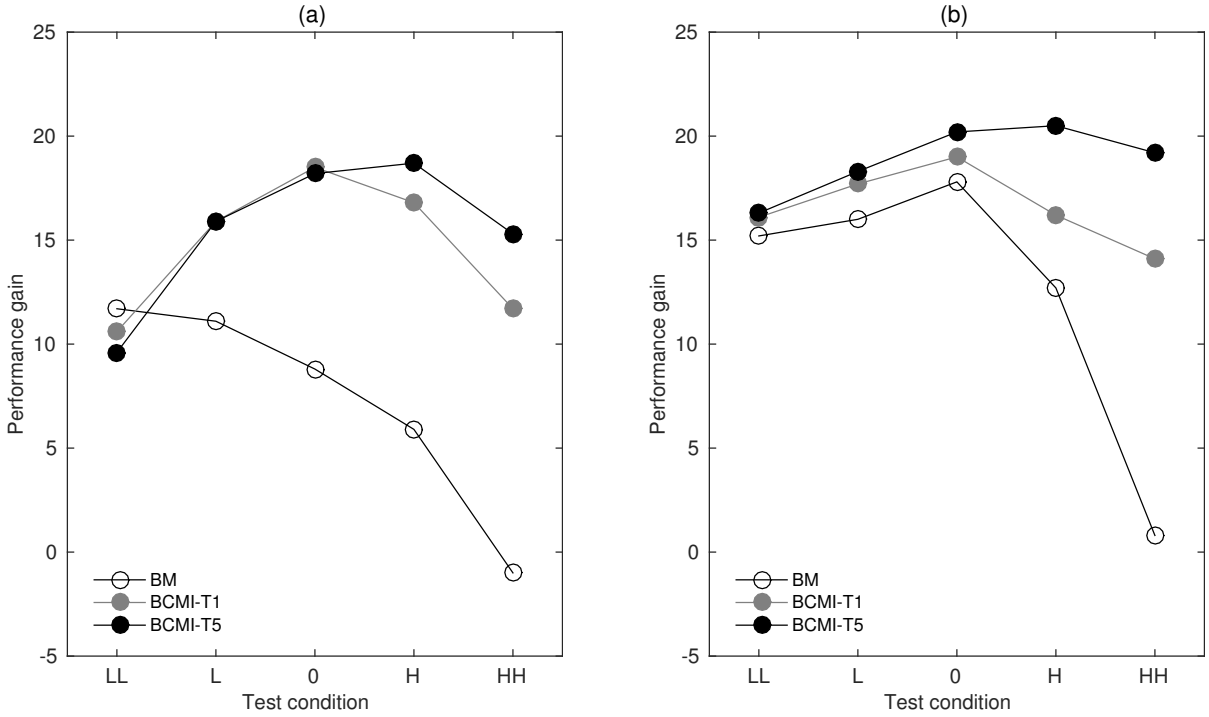

Figure 5: Performance gains when the missing data system is evaluated on (a) filtered and (b) interrupted speech with and without additive noise. The performance is measured in terms of letter error rate (LER) and the gain is calculated as the difference between LER in the conditions with and without additive noise (Table 4).

Table 5: Human and automatic speech recognition results on (a) filtered and (b) interrupted speech with and without additive noise. Performance is reported in terms of letter error rate. When noise was included, the noise level exceeded the speech level by 10 dB(A).

(a)

	Human	Baseline	BCMI-T5
Speech	52.7	87.7	87.7
Speech and noise	38.0	92.9	66.9

(b)

	Human	Baseline	BCMI-T5
Speech	43.0	70.4	70.4
Speech and noise	31.7	77.4	49.9

The correlations between listener and BCMI-T5 performance do not indicate that the proposed system simulates human speech perception or the mechanisms underlying perceptual restoration. However, since perceptual restoration occurred in listeners, it is reasonable to interpret the correlation between listener and ASR performance as an indication that the missing data system replicates listener performance, and hence achieves perceptual restoration, to some extent. For reference, we also calculated the correlation between listener performance rates and average listener performance. The average performance was calculated without the listener whose performance was being compared to the average. The listener and average listener performance are correlated with $\rho = 0.48$ ($p < 0.001$) in the filtered speech condition and with $\rho = 0.57$ ($p < 0.001$) in the interrupted speech condition.

7. Discussion

Experiments were conducted that compared human listeners, a baseline ASR system and a missing data ASR system in a perceptual restoration task. The results are discussed in Section 7.1. In addition, the missing data system performance in the filtered and interrupted conditions is discussed in more detail, and compared to relevant previous studies conducted with human listeners, in Sections 7.2 and 7.3.

7.1. Perceptual restoration in machine and human listeners

Perceptual restoration studies indicate that restoration occurs when removed acoustic content is substituted with additive noise, and that restoration is most potent when the additive noise could have plausibly masked the removed speech content. Hence, in order to determine whether an ASR system models perceptual restoration, we compared the system performance in conditions where sections of the speech signal were removed, and the resulting spectro-temporal gaps were either left as silence or filled with additive noise. Our data shows that the performance of the uncompensated baseline ASR system is not consistent with human perceptual restoration. Filling gaps in the speech with low-level noise improved the baseline system performance, but performance worsened as the noise level was increased, and was worse with intense additive noise than in conditions without additive noise. In contrast, additive noise consistently improved the performance of frame-based and window-based BCMI systems in the perceptual restoration task. For completeness, system performance was also evaluated on the dataset presented to human listeners in the second transcription test, and correlations between ASR and listener performance were calculated to illustrate the difference between baseline and missing-data system performance as perceptual restoration models.

Since the baseline system performance is not consistent with perceptual restoration, we conclude that the missing data front-end and observation uncertainties improved system performance in the perceptual restoration task. The difference in system performance can be explained in terms of the acoustic model likelihood calculation used by the different systems. When the baseline system compares observed feature vectors to acoustic model mean vectors, it does not differentiate between an additive distortion such as noise and a subtractive distortion such as the removal of speech content. Since the missing data system assumes that additive-noise-dominated components are unreliable, it differentiates between additive and subtractive distortions.

Previous studies have indicated that perceptual restoration can be modelled as source separation. Here, the missing data front-end and observation uncertainties improved system performance in the perceptual restoration task, meaning that perceptual restoration can be modelled as top-down completion. However, while the missing data system outperformed the uncompensated baseline system, the performance rates were not comparable to human performance. For example, our results are not in line with a study by Cooke (2006) in which human listeners and an ASR system using bounded marginalisation showed comparable performance in a consonant identification task. This discrepancy is most likely due to differences in speech material and recognition task. Cooke (2006) studied vowel-consonant-vowel data, whereas the speech stimuli in the current work consisted of read sentences with syntactic and semantic context. Semantic cues are expected to improve listener performance in a perceptual restoration task (Verschuure and Brocaar, 1983) and listeners can take advantage of semantic contexts spanning over sentences. In contrast, our ASR system uses a variable length n-gram language model, and can only benefit from rather short contexts (commonly no more than a few words).

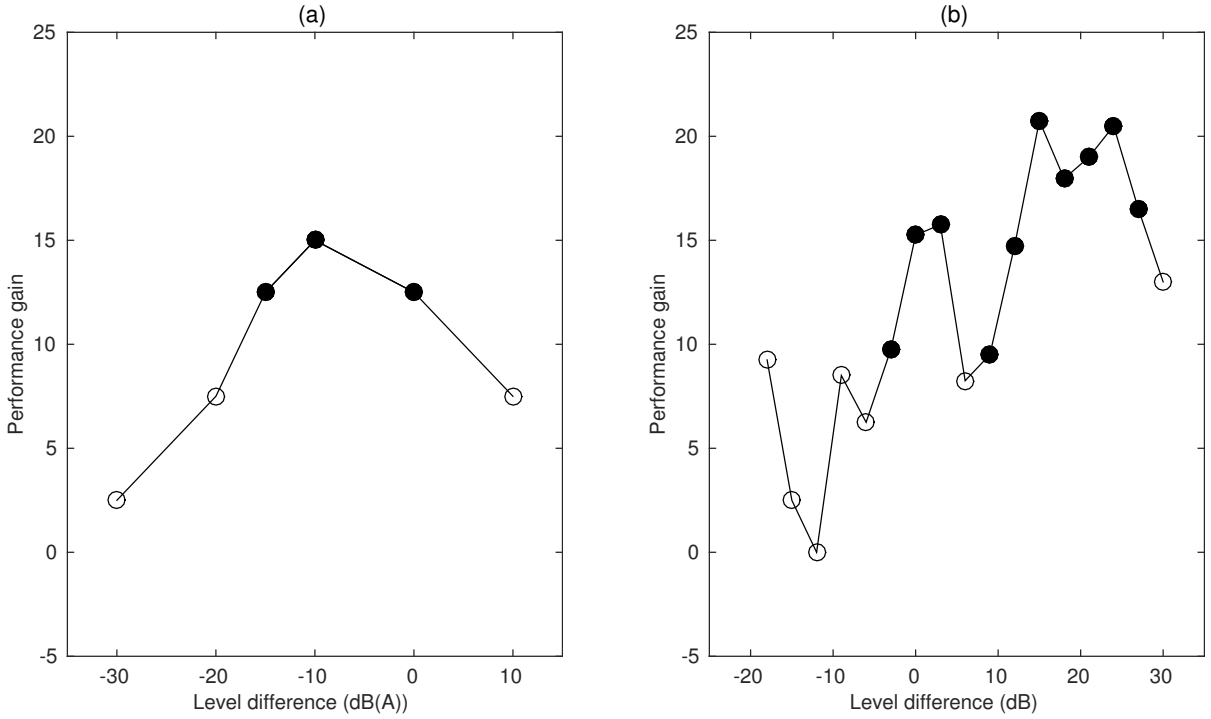

Figure 6: Listener data from other perceptual restoration studies. (a) Performance on filtered speech reported by Warren et al. (1997, Figure 2). (b) Performance on interrupted speech reported by Powers and Wilcox (1977, Table III–IV). In both cases, performance gains are calculated as the difference between keyword identification performance in conditions with and without additive noise, and the level differences indicate the difference in speech and additive noise levels. Test conditions in which additive noise improves listener performance ($p < 0.05$) are marked with filled circles.

7.2. Filtered speech

Filtered speech has been studied in perceptual restoration experiments by Warren et al. (1997). They used read sentences presented in two narrow spectral bands centred at 370 Hz and 6000 Hz, and measured listener performance in a keyword identification task. The stimuli they used retained less information than the bandstop-filtered speech stimuli used in the current work, and their listeners achieved a keyword identification percentage of 20% when the stimuli were presented without additive noise. While the addition of noise improved the performance of their listeners, it remained low compared to that reported here. The listener performance reported in Warren et al. (1997) is presented in Figure 6a.

Due to the differences in stimuli and in the evaluation measure, we do not consider performance comparisons with respect to the exact noise level relevant. However, we observe certain common trends in the listener performance rates (Figure 6a) and ASR performance (Figure 5a). First, the listener and representative missing data system (BCMI-T5) performance at low noise levels improve as the additive-noise level increases. Moreover, in both cases there is an optimal additive noise level, beyond which the listener and ASR system performance decline. This is because additive noise that is much more intense than the removed speech content does not provide acoustic-level cues that would aid restoration, and the observed features cease to constrain the clean speech posterior distribution in Equations (4)–(5).

7.3. Interrupted speech

Powers and Wilcox (1977) conducted perceptual restoration experiments using interrupted speech and additive noise at several interruption rates and noise levels. Listener performance at different additive-noise levels was evaluated with $1/3$ second interruptions and white noise inserts. The sentence lists and evaluation measure were the same as those used in the spectral restoration study of Warren et al. (1997). Listener performance across noise levels in the Powers and Wilcox (1977) study is shown in Figure 6b. The method used in our listening test differs in a number

of respects from these previous studies (e.g., in interruption rate, speech material, evaluation measure, and additive-noise type). Nonetheless, our data is in agreement with previous findings that additive noise either does not improve listener performance, or improves performance by a small amount when the additive-noise level is close to speech level. A similar observation has been reported by Verschuure and Brocaar (1983), where pilot experiments indicated that perceptual restoration is most effective when the additive-noise level exceeds the speech level by 10 dB.

The experiments conducted in Powers and Wilcox (1977) and Warren et al. (1997) indicate a difference between spectral and temporal restoration tasks. While the noise level does not need to exceed the speech level for additive noise to improve listener performance in the spectral restoration task (Figure 6a), listener performance evaluated on interrupted stimuli with additive noise does not improve at low noise levels (Figure 6b). In contrast, the representative missing data system (BCMI-T5) performance is near identical in the filtered and interrupted speech conditions (Figure 5). However, we note that the ASR system and listener performance on interrupted stimuli are consistent, in that performance improves when additive-noise level increases and declines when an optimal noise level is exceeded. Also, the difference observed at low noise levels is related to the difference between performance with and without additive noise. Thus, the difference between system and listener performance at low noise levels may not even relate to the perceptual restoration model. A notable difference between human listeners and ASR systems, or between the spectral and temporal restoration tasks, is that human listeners, when presented with interrupted speech without additive noise, can perceive and locate what was removed, and can compensate for the information loss at a conscious level. Verschuure and Brocaar (1983) hypothesised that conscious restoration and perceptual restoration are separate processes, and that, unlike perceptual restoration, conscious restoration improves with practice.

8. Conclusions

Previous studies indicate that listener performance in perceptual restoration tasks improves as additive-noise level increases, and reaches an optimal level when the noise is not too intense. In the current work, the same trends were observed in the performance of a missing data ASR system. In addition, direct comparison between the missing data system and human listeners on the same utterances indicated that the performance of the two was correlated. A baseline ASR system performance did not correlate with listener performance evaluated in the current or previous studies; we therefore conclude that the missing data methods and observation uncertainties improved the ASR system performance in the perceptual restoration task.

While the proposed missing-data approach improved system performance in perceptual restoration task, our results also suggested that the missing data system does not utilise available semantic and syntactic context to their full extent. Future experiments are needed to evaluate system performance on isolated words, and the performance on isolated words and complete sentences must be compared in order to determine whether the additional context information available in complete sentences improves system performance. Since previous studies indicate that semantic context is important to human listeners in the perceptual restoration task, future experimenters may wish to evaluate top-down completion approaches with a whole-sentence language model (Rosenfeld et al., 2001).

Acknowledgements

This work received financial support from the Academy of Finland under grants no 136209, 272710, and 251170 (Finnish Centre of Excellence in Computational Inference Research), from Tekes under the FUNESOMO project, and from EC FP7 under grant agreement 287678. GJB was supported by the EC FP7 grant Two!Ears (ICT-617075).

References

Arrowood, J. A., Clements, M. A., 2002. Using observation uncertainty in HMM decoding. In: Proc. ICSLP. pp. 1561–1564.

Astudillo, R. F., Kolossa, D., Mandelartz, P., Orglmeister, R., 2010. An uncertainty propagation approach to robust ASR using the ETSI advanced front-end. IEEE J. Selected Topics in Signal Processing 4 (5), 824–833.

Barker, J., 2012. Missing-data techniques: Recognition with incomplete spectrograms. In: Virtanen, T., Singh, R., Raj, B. (Eds.), Techniques for Noise Robustness in Automatic Speech Recognition. John Wiley & Sonds, Ltd, pp. 371–398.

Bashford, Jr., J. A., Riener, K. R., Warren, R. M., 1992. Increasing the intelligibility of speech through multiple phonemic restorations. Perception & Psychophysics 51 (3), 211–217.

Bregman, A. S., 1990. Auditory Scene Analysis. MIT Press.

Cerisara, C., Demange, S., Haton, J. P., 2007. On noise masking for automatic missing data speech recognition: A survey and discussion. Computer Speech & Language 21 (3), 443–457.
Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.
Cooke, M., Green, P., Crawford, M., 1994. Handling missing data in speech recognition. In: Proc. ICSLP. pp. 1555–1558.
Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.
Cooke, M. P., Brown, G. J., 1993. Computational auditory scene analysis: exploiting principles of perceived continuity. Speech Communication 13 (3–4), 391–399.
Deng, L., Droppo, J., Acero, A., 2005. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Speech and Audio Processing 13 (3), 412–421.
Edmonds, B. A., Culling, J. F., 2005. The spatial unmasking of speech: evidence for within-channel processing of interaural time delay. J. Acoust. Soc. Am. 117 (5), 3096–3078.
Ellis, D. P. W., 1999. Using knowledge to organize sound: The prediction-driven computational auditory scene analysis and its application to speech/nonspeech mixtures. Speech Communication 27 (3–4), 281–298.
Faubel, F., McDonough, J., Klakow, D., 2009. Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features. In: Proc. ICASSP. pp. 3869–3872.
Gemmeke, J. F., Remes, U., 2012. Missing-data techniques: Feature reconstruction. In: Virtanen, T., Singh, R., Raj, B. (Eds.), Techniques for Noise Robustness in Automatic Speech Recognition. John Wiley & Sonds, Ltd, pp. 399–432.
González, J. A., Peinado, A. M., Ma, N., Gómez, A. M., Barker, J., 2013. MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. IEEE Trans. Audio, Speech, and Language Processing 21 (3), 624–635.
Häkkinen, J., Haverinen, H., 2001. On the use of missing feature theory with cepstral features. In: Proc. CRAC Workshop.
Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language 20 (4), 515–541.
Hirsimäki, T., Pylkkönen, J., Kurimo, M., 2009. Importance of high-order n-gram models in morph-based speech recognition. IEEE Trans. Audio, Speech, and Language Processing 17 (4), 724–732.
Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., Kiessling, A., 2002. SPEECON - speech databases for consumer devices: Database specification and validation. In: Proc. LREC. pp. 329–333.
Kämäräinen, J., Paalanen, P., 2005. GMMBAYES Toolbox V1.0.
URL `www.it.lut.fi/project/gmmbayes/`
Lanman, D. R., 2006. Sound level meter.
URL `http://www.mathworks.com/matlabcentral/fileexchange/9603-sound-level-meter`
Masuda-Katsuse, I., Kawahara, H., 1999. Dynamic sound stream formation based on continuity of spectral change. Speech Communication 27, 235–259.
Miller, G. A., Licklider, J. C. R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22 (2), 167–173.
Nádas, A., Nahamoo, D., Picheny, M. A., 1989. Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoustics, Speech, and Signal Processing 37 (10), 1495–1503.
Powers, G. L., Speaks, C., 1973. Intelligibility of temporally interrupted speech. J. Acoust. Soc. Am. 54 (3), 661–667.
Powers, G. L., Wilcox, J. C., 1977. Intelligibility of temporally interrupted speech with and without intervening noise. J. Acoust. Soc. Am. 61 (1), 195–199.
Pylkkönen, J., Kurimo, M., 2004. Duration modeling techniques for continuous speech recognition. In: Proc. INTERSPEECH. pp. 385–388.
Remes, U., Nankaku, Y., Tokuda, K., 2011. GMM-based missing-feature reconstruction on multi-frame windows. In: Proc. INTERSPEECH. pp. 1665–1668.
Remes, U., Ramírez López, A., Palomäki, K., Kurimo, M., to appear. Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation. IEEE/ACM Trans. Audio, Speech and Language Processing.
Repp, B. H., 1992. Perceptual restoration of a "missing" speech sound: Auditory induction or illusion? Perception & Psychophysics 51 (1), 14–32.
Rosenfeld, R., Chen, S. F., Zhu, X., 2001. Whole-sentence exponential language models: a vehicle for linguistic-statistical integration. Computer Speech & Language 15 (1), 55–73.
Samuel, A. G., 1981. Phonemic restoration: Insights from a new methodology. J. Experimental Psychology: General 110 (4), 474–494.
Srinivasan, S., Wang, D. L., 2005. A schema-based model for phonemic restoration. Speech Communication 45, 63–87.
Varga, A., Moore, R. K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP. pp. 845–848.
Verschuure, J., Brocaar, M. P., 1983. Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. Perception & Psychophysics 33 (3), 232–240.
Warren, R. M., 1970. Perceptual restoration of missing speech sounds. Science 167, 393–393.
Warren, R. M., 1984. Perceptual restoration of obliterated sounds. Psychological Bulletin 96 (2), 371–383.
Warren, R. M., Riener Hainsworth, K., Brubaker, B. S., Bashford, Jr., J. A., Healy, E. W., 1997. Spectral resotration of speech: Intelligibility is increased by inserting noise in spectral gaps. Perception & Psychophysics 52 (2), 275–283.
Warren, R. M., Sherman, G. L., 1974. Phonemic restorations based on subsequent context. Perception & Psychophysics 16 (1), 150–156.