Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment

Sami Keronen[a,*], Heikki Kallasjoki[a], Ulpu Remes[a], Guy J. Brown[b], Jort F. Gemmeke[c], Kalle J. Palomäki[a]

[a] *Aalto University School of Science, Department of Information and Computer Science*
PO Box 15400, FI-00076 Aalto, Finland
Tel. +358-9-470-23272, Fax. +358-9-470-23277
[b] *University of Sheffield, Department of Computer Science*
Regent Court, 211 Portobello St., Sheffield S1 4DP,, UK
[c] *KU Leuven, Department ESAT-PSI*
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

Abstract

We present an automatic speech recognition system that uses a missing data approach to compensate for challenging environmental noise containing both additive and convolutive components. The unreliable and noise-corrupted ("missing") components are identified using a Gaussian mixture model (GMM) classifier based on a diverse range of acoustic features. To perform speech recognition using the partially observed data, the missing components are substituted with clean speech estimates computed using both sparse imputation and cluster-based GMM imputation. Compared to two reference mask estimation techniques based on interaural level and

*Corresponding author
Email addresses: firstname.lastname@aalto.fi (Sami Keronen), firstname.lastname@aalto.fi (Heikki Kallasjoki), firstname.lastname@aalto.fi (Ulpu Remes), g.brown@dcs.shef.ac.uk (Guy J. Brown), jgemmeke@amadana.nl (Jort F. Gemmeke), firstname.lastname@aalto.fi (Kalle J. Palomäki)

time difference-pairs, the proposed missing data approach significantly improved the keyword accuracy rates in all signal-to-noise ratio conditions when evaluated on the CHiME reverberant multisource environment corpus. Of the imputation methods, cluster-based imputation was found to outperform sparse imputation. The highest keyword accuracy was achieved when the system was trained on imputed data, which made it more robust to possible imputation errors.
Keywords: Noise robust, Speech recognition, Missing data, Binaural, Multicondition, Imputation

1. Introduction

The performance gap between human listeners and automatic speech recognition (ASR) still remains large when recognition in noisy acoustic environments is considered. One approach to reducing this performance difference is the use of missing data (MD) methods, which are motivated by studies of the human auditory system (Cooke et al., 1994). In MD methods, the observed noisy speech is partitioned into speech-dominated ("reliable") and noise-dominated ("unreliable") components. A number of ways to treat the reliable and unreliable components have been proposed. In the marginalization approach, the unreliable components are completely discarded, whereas in bounded marginalization, they are used as an upper bound to the missing clean speech values (Cooke et al., 2001). Unreliable components can also be reconstructed (imputed) by replacing them with clean speech estimates which, for example, can be obtained from statistical models of speech (Raj et al., 2004) or from a dictionary of speech exemplars (Gemmeke et al., 2011).

The quality of the classification of the observations into reliable and unreliable regions (so-called "masking") is a central factor in the performance of missing data approaches. The missing data method does not necessarily require strong assumptions about the type of background noise as information from the noise-corrupted regions does not need to be used (e.g. in marginalization or imputation). However, missing data approaches can be made more powerful by including noise models that are used in mask estimation, or provide an upper bound to observations. Missing data methods have been shown to significantly improve the noise robustness of ASR systems in experiments where knowledge of the clean speech and noise signals is available prior to mixing and exact ("a priori", or "oracle") masks of reliable regions can be used (Cooke et al., 2001).

In practice, oracle masks are not available and missing data masks must be estimated from the input signal. Several methods have been proposed to achieve this. In (Vizinho et al., 1999), the masks are created by thresholding sub-band signal-to-noise ratio (SNR) estimates, while in (Seltzer et al., 2004) a Bayesian classifier is applied to a variety of features derived from the acoustic signal, such as statistical measures of subband energies. The statistical approach has been shown to outperform SNR-based mask estimation (Seltzer et al., 2004). Support vector machine (SVM) classifiers were proposed for the mask estimation task in (Gemmeke et al., 2009) with a set of features similar to Seltzer et al. (2004) supplemented with features derived from a harmonic decomposition of the input signal (Van hamme, 2004). The mask estimation problem in both above mentioned studies was formulated as one in which each spectral channel in each time frame is classified as reliable

or unreliable, which is a more localized decision compared to e.g. classification of phonemes in ASR. This is possibly one of underlying reasons for the success of multifeature approaches (Seltzer et al., 2004) in mask estimation.

Mask estimation methods have also been proposed that are based on the output of auditory models. In particular, masks can be estimated using models of binaural hearing, which measure interaural level (ILD) and time of arrival (ITD) differences (Roman et al., 2003; Harding et al., 2006), and thereby provide information about the spatial location of the target speaker and interfering sources. Binaural mechanisms also counteract the detrimental effects of reverberation (Zurek, 1987), for example by identifying signals that are coherent at the two ears (Faller and Merimaa, 2004). However, when the target and interfering sources share the same or nearby azimuthal location, systems that rely solely on binaural cues (Roman et al., 2003; Harding et al., 2006) can not distinguish between the target and interferences. It is also noteworthy that human listeners gain only a limited advantage from binaural cues over purely monaural listening. Therefore any robust approach should also include monaural cues in the mask estimation. Another auditory-motivated technique for mask estimation of reverberation-contaminated speech is described in (Palomäki et al., 2004), in which modulation filtering is used to identify spectro-temporal regions containing strong speech energy.

While several imputation methods have been proposed for missing feature reconstruction, arguably the ones yielding the highest recognition accuracy for front-end based reconstruction are cluster-based imputation (Raj et al., 2004) and sparse imputation (Gemmeke et al., 2010). In cluster-based imputation, the statistical dependencies between clean speech features are repre-

sented using a Gaussian mixture model (GMM) and clean speech estimates for the missing components are calculated using bounded maximum *a posteriori* (MAP) estimation. In sparse imputation, the clean speech features are modeled as a linear combination of clean speech dictionary units referred to as exemplars. Estimates for the missing components are obtained from an exemplar-based representation constructed to model the reliable components using as few exemplars as possible.

In this work, mask estimation using a GMM classifier on a comprehensive set of features is proposed to address the CHiME challenge (Barker et al., 2013), which consists of stereo data recorded in a noisy, multisource reverberant environment. The feature set used in mask estimation includes features derived from binaural and monaural auditory models, together with features designed to distinguish between speech and non-speech elements. The mask estimation method is evaluated in a missing data reconstruction-based automatic speech recognition task using the cluster-based imputation and sparse imputation methods. In this work, cluster-based imputation is used with multi-frame windows as proposed in (Remes et al., 2011).

This paper makes three main contributions. Firstly, we present a comparison of cluster-based and sparse imputation methods on the CHiME corpus. While cluster-based and sparse imputation methods have performed well on various noise-robust speech recognition tasks, the CHiME data used in this work is reverberant and contains a variety of challenging noise types that neither method has been evaluated on.

Secondly, we investigate whether performance improvements can be obtained via speaker adaptation or re-training to account for the reconstruction

error. While multicondition training on enhanced speech has resulted in significant improvements when combined with a noise suppression method (Raj et al., 2010), there is no previous work on re-training acoustic models to accommodate for the reconstruction errors from missing feature reconstruction. In this work, the acoustic models re-trained on reconstructed features are referred to as *multicondition models for imputed data*.

Finally, we systematically evaluate a range of acoustic features for mask estimation, including those previously used in statistical mask estimation (Seltzer et al., 2004) and two new binaural features that have not been previously used in the context of missing data reconstruction (peak ITD and interaural coherence). To evaluate the effect of the proposed mask estimation method on the recognition accuracy, the estimated masks are compared with oracle masks, binaural reference masks adapted from (Harding et al., 2006), and masks computed by a GMM classifier trained on ILD–ITD pairs. To evaluate the contribution of individual features, we analyse their power to discriminate between reliable and unreliable values.

The remainder of the paper is organized as follows. The methods used in this work are described in Section 2, with an introduction to missing data techniques in Section 2.1, features and methods used for mask estimation in Section 2.2, and reconstruction methods in Section 2.3. The speech recognition system and noise robust speech recognition task are presented in Section 3 and the results in Section 4. We discuss the results and present our conclusions in Section 5.

2. Methods

2.1. Missing data techniques

Missing data techniques (Cooke et al., 2001) are based on the assumption that magnitude-compressed spectral features $\boldsymbol{Y}$ that represent noisy speech can be divided in speech and noise dominated components. Labels that divide the observations into speech and noise dominated regions are referred to as a *missing data mask*. The time–frequency components are denoted as $Y(\tau, d)$, where τ represents the time frame and d represents the frequency bin. The components that are speech-dominated are considered reliable estimates of the underlying clean speech features, $Y_r(\tau, d) \approx S(\tau, d)$, where $S(\tau, d)$ denotes the clean speech value that would have been observed if the signal had not been corrupted with noise. The noise-dominated components, on the other hand, are considered unreliable, and assuming that the noise originates from an uncorrelated source, the unreliable observations provide only an upper bound on the corresponding clean speech, $Y_u(\tau, d) \geq S(\tau, d)$. Thus, the clean speech information in the unreliable components is effectively missing, and must be compensated for in speech recognition.

Given the reliable and unreliable observations, the speech recognition system can either be modified to ignore the missing clean speech data, or the missing values can be reconstructed with clean speech estimates $\hat{S}(\tau, d)$. Because the reconstructed features $\hat{\boldsymbol{S}}$ do not contain any missing values, they can be subjected to any feature transformations (e.g. cepstral transformation) and can be used for speaker adaptation without a need to modify the algorithm. Reconstruction methods such as cluster-based imputation (Raj et al., 2004) and sparse imputation (Gemmeke et al., 2010) have also per-

formed well on various speech recognition tasks. The clean speech estimates (i.e. reconstructed features) are calculated using the reliable and unreliable observations and a clean speech model.

2.2. Mask estimation

In this section we describe a range of acoustic features that form the basis for classifying time-frequency regions as reliable or unreliable. The intended purpose of the multifeature approach presented here is to model various acoustic cues that could signal the reliability of time-frequency regions in CHiME data that is binaural, reverberant, and contains multiple overlapping and often highly non-stationary sources (e.g. due to a competing talker). For distinguishing between competing talkers, features taking advantage of directional cues from the target talker could be useful. For non-speech interference, cues that discriminate between speech and non-speech would be effective. Given the above considerations, the features applied here vary in their intended purpose; some are intended to provide robustness to reverberation, or to characterize the properties of the background noise and therefore provide noise robustness. Others have a focus on target detection, e.g. by using binaural cues to identify when a source is present at the known location of the target talker. They also vary in their characteristics (e.g., whether they are based on monaural or binaural signals). A summary of the feature types is given in Table 1. First, the intermediate signal representations required by the different features are described in Section 2.2.1, which is followed by the descriptions of the individual features in Section 2.2.2.

Table 1: Summary of acoustic features used for mask estimation. The table indicates whether the main focus of each feature is reverberation robustness, noise robustness or detecting the target (speech) source.

Acronym	Description	Monaural/ Binaural	Reverberation robustness	Noise robustness	Target detection
MOD	Modulation-filtered spectrogram	M	X		
MPR	Mean-to-peak-ratio of temporal envelope	M	X		
GRAD	Gradient of temporal envelope	M	X		
HA	Harmonic energy	M			X
IHA	Inharmonic energy	M		X	
LTE	Noise estimate from long-term inharmonic energy	M		X	
GAIN	Noise gain	M		X	
FLAT	Spectral flatness	M		X	
S2N	Subband energy to subband noise floor ratio	M		X	
ILD	Interaural level difference	B			X
ITD	interaural time difference	B			X
PITD	Peak interaural time difference	B			X
IC	Interaural coherence	B	X	X	
DIFF	Noise estimate from channel difference	B	X	X	

2.2.1. Intermediate signal representations

Two types of spectral acoustic features are used in our system. The features used by the ASR back-end are based on standard mel-frequency cepstral coefficients (MFCC). To compute these, the left- and right-ear signals are summed, filtered through a pre-emphasis filter, and then the magnitude spectrum is obtained via a fast Fourier transform (FFT) with a 16 ms frame length and 8 ms hop size. The magnitude-spectrum is then transformed to the mel-scale by applying 21 triangular filters with center frequencies between 171 Hz and 7097 Hz. When followed by log compression, the resulting log-

mel spectrogram is denoted by $Y(\tau, d)$, where τ denotes the frame and d the mel frequency bin. We adopt the notation that $Y^*(\tau, d)$ represents the mel spectrogram without log compression. MFCCs are obtained by decorrelating the log-mel spectrogram with a discrete cosine transformation and removing the zeroth and highest frequency components.

The standard MFCC features described above are suitable for training an ASR system, but a spectral representation that preserves the temporal fine structure within each frequency band is also required for mask estimation. In particular, temporal fine structure is required to compute the binaural features described in Section 2.2.2. Accordingly, we compute a second "mel-gammatone" spectral representation from an array of 21 gammatone filters, whose centre frequencies and bandwidths are set to match the characteristics of the MFCC front-end described above. The left-ear and right-ear signals are passed separately through a mel-gammatone filterbank and then half-wave rectified, giving a representation for the left and right ears denoted $G_l(t, d)$ and $G_r(t, d)$ respectively. Here, d represents the frequency channel and t is the discrete time in samples.

For computing ITD, peak ITD and interaural coherence features (described in Section 2.2.2), the left-ear and right-ear mel-gammatone filtered signals are cross-correlated. Here, the generalized cross-correlation method, which applies a phase transform (GCC-PHAT) (Knapp and Carter, 1976) with a parameter γ for changing the level of normalization (Tikander et al., 2003) is used. Compared to a conventional cross-correlation, GCC-PHAT suppresses secondary peaks in the cross-correlation and has been shown to produce better target localization accuracy (Perez-Lorenzo et al., 2012).

Given the two mel-gammatone signals $G_l(t,d)$ and $G_r(t,d)$ the GCC-PHAT for frame τ and channel d is defined as

$$G^F_{PHAT}(\tau,d) = \frac{G^F_l(\boldsymbol{w},d)[G^F_r(\boldsymbol{w},d)]^*}{|G^F_l(\boldsymbol{w},d)[G^F_r(\boldsymbol{w},d)]^*|^\gamma}, \quad (1)$$

where the superscript $[\ ]^F$ denotes the Fourier transform, $[\ ]^*$ denotes the complex conjugate, $\boldsymbol{w} = [\tau t, \ldots, \tau t + (W_l - 1)]$ is a indexing vector corresponding to a rectangular window of length $W_l = 256$ samples, and $\gamma = 0.8$ is the parameter for tuning the amount of magnitude normalization. GCC-PHAT is computed for each frequency channel over a 16 ms rectangular window with an 8 ms hop size.

2.2.2. Acoustic features

Modulation filtered spectrogram (MOD): Speech signals have their largest temporal modulation at the syllabic rate, which peaks at around 4 Hz. The effect of reverberation on speech is to reduce the modulation depth, as the gaps between syllable onsets are filled with reverberant energy (particularly the late reverberation component). It has been shown in e.g. (Kingsbury et al., 1998) that reverberation-robust features for ASR can be obtained by filtering spectral features along their time trajectory with filters that emphasize the syllabic modulations. Modulation filtering has also been used to discriminate "clean" speech features from those that have been contaminated by reverberant energy, in the missing feature mask generation approach of (Palomäki et al., 2004).

The modulation-filtered mel-spectrogram $MOD(\tau,d)$ is obtained by fil-

tering each channel d of $Y^*(\tau, d)^{0.3}$, described in Section 2.2.1, as follows

$$MOD(\tau, d) = \sum_{k=-\infty}^{\infty} f(k) Y^*(\tau - k, d)^{0.3}, \tag{2}$$

where filter $f(k)$ is a bandpass finite impulse response (FIR) filter with 3 dB cutoff frequencies at 1.8 and 10.2 Hz. Here, the compression factor is set to 0.3 to obtain a match to studies in which the MOD measure was originally developed (Palomäki et al., 2004) based on an auditory firing rate signal representation (for an implementation, see (Barker, 2001)), which is different from the log-compressed feature presentation used in the present study. The goal of the filtering is to find reverberation-free syllable onsets. The FIR filter is designed as a linear-phase smoothing lowpass convolved with a differentiator.

Mean to peak ratio (MPR): When speech is reverberated, the peaks in the temporal envelope are largely unaffected, but the valleys become filled with reverberant energy. Temporal smoothing of spectrograms due to reverberation can therefore be measured by computing the mean-to-peak ratio of the speech temporal envelope (Palomäki et al., 2004).

The scalar valued mean to peak ratio MPR across all channels d and all frames τ of a mel spectrogram $Y^*(\tau, d)$ is denoted as

$$MPR = \frac{1}{D} \sum_{d=1}^{D} \frac{\frac{1}{L} \sum_{\tau=1}^{L} Y^*(\tau, d)^{0.3}}{\max_\tau \left(Y^*(\tau, d)^{0.3} \right)}, \tag{3}$$

where $D = 21$ is the number of frequency channels, τ is a time frame, and L is the number of time frames in the utterance. Note that the above formulation defines MPR as a single (global) scalar measure over the whole utterance. During mask estimation for an utterance, the computed MPR is used for each

time-frequency bin within that utterance, thus $MPR(\tau, d) = MPR$. A within-channel MPR metric has also been tested in an earlier study (Brown and Palomäki, 2008), but the global metric was found to give better performance; accordingly, we only use the global metric here.

Gradient (GRAD): Reverberant regions of the speech temporal envelope are often associated with decaying tails, and it has been suggested that the detection of such tails may play a role in perceptual compensation for the effects of reverberation (Watkins and Makin, 2007). Hence, we include a gradient feature that measures the local slope of the temporal envelope within a short time window, for each frequency channel. The slope $GRAD(\tau, d)$ is determined for each frequency channel d by a linear regression line through a five-point window centered on each time frame τ. Here, the equation of $GRAD(\tau, d)$ is defined as

$$GRAD(\tau, d) = \frac{1}{10} \sum_{i=-2}^{2} Y_e(\tau + i, d)i, \tag{4}$$

where $Y_e(\tau, d)$ is a frame extended version of $Y(\tau, d)$, whose values outside the borders of the mel-spectrogram are constructed by repeating the border values.

Harmonic energy (HA): In voiced segments of speech, the speech signal will consist primarily of components harmonically related to the pitch of the speaker, while we assume no such relationship exists between the noise and the speaker pitch. Consequently, when the spectrum is decomposed into harmonic and inharmonic parts, the harmonic part will be dominated by the speech signal. The harmonic feature is formed using a harmonic decomposition based on a pitch estimate (Van hamme, 2004). In the decomposition,

the complete time-domain input utterance $y_{\text{utt}}(t)$ is windowed into overlapping frames, with a frame length of two estimated pitch periods and a frame shift of one pitch period. For each frame, the noisy signal $y(t)$ is modeled as a sum of a harmonic time-domain signal $h(t)$ and a residual $r(t)$,

$$y(t) = h(t) + r(t). \tag{5}$$

The harmonic part has the form

$$h(t) = \left(1 + \frac{ct}{N}\right) \cdot \left[\sum_{k=0}^{K} a_k \cos(2\pi f_0 k t) + \sum_{k=1}^{K} b_k \sin(2\pi f_0 k t)\right] \tag{6}$$

where f_0 is the pitch estimate given in normalized form where 1 corresponds to the sampling rate, N is the corresponding pitch period with the time index t ranging from 0 to $2N-1$, a_k, b_k and c are parameters estimated from the signal using the iterative approach of (Van hamme, 2004), and the number of harmonics K is set to the largest integer such that $f_0 K < 0.5$. Finally, the central pitch periods of each frame are concatenated to get a non-overlapping estimate $h_{\text{utt}}(t)$ for the harmonic part of the input utterance. The harmonic feature components $HA(\tau, d)$ for each frame τ and frequency channel d are then obtained as the log-mel spectral representation of the harmonic time-domain signal estimate.

Using $\text{Mel}(y, \tau, d)$ to denote the output of the process for generating the mel spectrogram $Y^*(\tau, d)$, described in Section 2.2.1, with a time-domain input signal $y(t)$, the harmonic energy features are then

$$HA(\tau, d) = 10 \cdot \log_{10} \text{Mel}(h_{\text{utt}}, \tau, d). \tag{7}$$

Inharmonic energy (IHA): The inharmonic residual of the harmonic decomposition outlined above is simply the difference between the input signal

and the harmonic signal estimate in the time domain. From Equation (5), $r_{\mathrm{utt}}(t) = y_{\mathrm{utt}}(t) - h_{\mathrm{utt}}(t)$. In voiced speech segments, the inharmonic part will mostly consist of noise. The inharmonic features $IHA(\tau, d)$ are again the log-mel spectral representation of the residual signal,

$$IHA(\tau, d) = 10 \cdot \log_{10} \mathrm{Mel}(r_{\mathrm{utt}}, \tau, d). \tag{8}$$

Noise estimate from long-term inharmonic energy (LTE): This feature is based on the inharmonic residual of the harmonic decomposition (IHA). The component $LTE(\tau, d)$ is the first quartile value of the corresponding d'th subband energies of the log-mel spectrum of the residual signal within a long time window centered at time frame τ,

$$LTE(\tau, d) = Q_1(\{IHA(\tau', d) \mid \tau' = \tau - 20 \ \ldots \ \tau + 20\}), \tag{9}$$

where $Q_1(\cdot)$ extracts the first quartile value.

Gain (GAIN): The gain feature is a rudimentary SNR estimate (Van hamme, 2004) based on the harmonic energy (HA) and the inharmonic energy noise estimate (LTE) features described earlier, as

$$GAIN'(\tau, d) = 10 \cdot \log_{10} \left(\max \left\{ \frac{HA^*(\tau, d) - 3LTE^*(\tau, d)}{HA^*(\tau, d)}, 10^{-4} \right\} \right), \tag{10}$$

where HA^* and LTE^* are the corresponding features in the uncompressed linear mel-spectral domain. The final $GAIN(\tau, d)$ features are the values of $GAIN'$ after two-dimensional mean-filter smoothing across the spectrogram elements:

$$GAIN(\tau, d) = \frac{1}{25} \sum_{\tau'=\tau-2}^{\tau+2} \sum_{d'=d-2}^{d+2} GAIN'(\tau', d'), \tag{11}$$

where values of $GAIN'(\tau', d')$ outside the borders of the mel-spectrogram are constructed by repeating the border values.

Flatness (FLAT): Qualitatively, when noise is added to a (voiced) speech signal, the effect in the spectral domain is to "flatten" the valleys between the resonances apparent in clean speech (Seltzer et al., 2004). This effect can be characterized by measuring the local variance of the subband energy within a neighborhood around each time-frequency cell in the log-mel spectrogram. The "flatness" feature used here is the sample variance of the noisy spectrogram $Y(\tau, d)$ within a 3×3 neighborhood

$$FLAT(\tau, d) = Var(Y(\tau_n, d_n)), \tag{12}$$

where τ_n and d_n run through values $[\tau - 1,\ \tau,\ \tau + 1]$ and $[d - 1,\ d,\ d + 1]$ that are not indexed outside the noisy spectrogram.

Although the *FLAT* and *MPR* features derive from the same fundamental idea and may appear redundant, there is no correlation between them; Pearson's correlation coefficient between the two is approximately -0.03 ($p < 0.01$) on both reliable and unreliable time-frequency units. The MPR is a global metric (since it sums across all frequency regions), whereas the FLAT metric is local.

Subband energy to subband noise floor ratio (S2N): A coarse estimate of the noise floor of stationary noise can be obtained by looking at the distribution of subband energy across all frames in an utterance. Such distributions tend to be roughly bimodal, with the lower mode corresponding to the noise energy during silence frames (Seltzer et al., 2004). For this feature, the subband noise floor $N(d)$ of channel d is approximated by locating the low peak from a histogram of the corresponding subband energies in the noisy log-mel spectrogram $Y(\tau, d)$, computed over all frames τ in the utterance. High values of the ratio of current subband energy to the noise floor energy,

$S2N(\tau, d) = Y(\tau, d)/N(d)$, are then indicative of regions dominated by either speech or highly non-stationary noise sources.

Interaural level difference (ILD): When stereo signals are available, features related to binaural hearing can be exploited. Human listeners are able to localize sound sources in space by measuring the interaural differences between the time of arrival (ITD) and sound level (ILD) at the two ears. The ILD at each time frame τ is calculated by taking the ratio of the energies of two windowed mel-gammatone signals, $G_r(\boldsymbol{w}, d)$ and $G_l(\boldsymbol{w}, d)$ described in Equation (1), and converting to decibels as follows

$$ILD(\tau, d) = 10 \log_{10} \frac{G_r(\boldsymbol{w}, d)^2}{G_l(\boldsymbol{w}, d)^2}. \tag{13}$$

Interaural time difference (ITD): Given the definition of G_{PHAT} in Equation (1), the cross-correlation for an interaural time difference k is given by:

$$g_{PHAT}(\tau, d, k) = \arg_k \mathrm{IFFT}\big(G^F_{PHAT}(\tau, d)\big), \tag{14}$$

where k is an index to vector $G^F_{PHAT}(\tau, d)$. The dominant ITD within channel d is then given by

$$ITD(\tau, d) = \operatorname{argmax}_k g_{PHAT}(\tau, d, k), \tag{15}$$

where the time lags k are computed between -1 ms and +1 ms in steps of the sampling period and $k = 0$ corresponds to the ITD for a source at zero degrees azimuth.

Peak ITD (PITD): The peak ITD metric is the ratio between the height of the highest peak and the height at zero delay in the cross-correlation. PITD is computed using Equation (14) as

$$PITD(\tau, d) = \frac{g_{PHAT}(\tau, d, 0)}{\max_k g_{PHAT}(\tau, d, k)}, \tag{16}$$

where time lags k are computed between -1 ms and +1 ms in steps of the sampling period. For sources at zero degrees azimuth, PITD should be close to unity.

Interaural coherence (IC): In anechoic environments with a single active sound source, the signals arriving at the two ears of a listener are highly coherent. However, in complex listening situations, such as in the presence of several sound sources and room reflections, sound from several different directions concurrently reaches the position of the listener and the interaural coherence is much reduced. Furthermore, the superposition of sound emanating from several directions results in instantaneous ITD and ILD cues that most of the time do not correspond to any of the source directions.

Faller and Merimaa (Faller and Merimaa, 2004) suggest that source localization in complex listening situations should be improved by retaining ITD and ILD cues only when they coincide with a high IC; when the IC is low, these cues are unlikely to give an accurate source direction and should be discarded. Here, a simplified version of the Faller-Merimaa model is used since the location of the target (zero degrees azimuth) is known and coherence is just the generalized cross-correlation value at zero lag. IC is computed using Equation (14) as $IC(\tau, d) = g_{PHAT}(\tau, d, 0)$.

Channel difference (DIFF): A rough estimate of the additive noise sources and reverberation can be formed by a simple channel difference measure obtained by subtracting the left-ear audio channel from the right-ear audio channel. The difference signal is converted to a log-mel spectral representation as described in Section 2.2.1. In this work, this is effective since we assume the spatial location of the speaker is known a priori and there is no

time delay between the left and right channels in the direct sound component of the target speech signals.

2.2.3. Classifier

In this work, the missing data masks are estimated using a two-class GMM classifier as proposed in (Seltzer et al., 2004). Each time–frequency component is represented as an N-dimensional feature vector $\boldsymbol{o}(\tau, d)$, where N is the number of features used in mask estimation. Since the features $\boldsymbol{o}(\tau, d)$ may vary depending on the frequency channel, a separate two-class classifier is trained for each frequency channel d. The training data for channel d is divided into two sets depending on whether the time–frequency component associated with the feature vector $\boldsymbol{o}(\tau, d)$ is labelled as reliable or unreliable, and an M-component GMM is estimated for both sets. An M-component GMM is a weighted sum of M Gaussian densities as given by the equation,

$$P(\boldsymbol{o}(\tau, d)|\Gamma) = \sum_{i=1}^{M} w_i g(\boldsymbol{o}(\tau, d)|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_I), \tag{17}$$

where $\boldsymbol{o}(\tau, d)$ is a N-dimensional continuous-value feature vector, $w_i, i = 1, \ldots, M$ are the mixture weights, and $g(\boldsymbol{o}(\tau, d)|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, \ldots, M$, are the component Gaussian densities. Each component density is a N-variate Gaussian function of the form,

$$g(\boldsymbol{o}(\tau, d)|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{o}(\tau, d) - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{o}(\tau, d) - \boldsymbol{\mu}_i)\right\}, \tag{18}$$

with the mean vector $\boldsymbol{\mu}_i$ and full covariance matrix $\boldsymbol{\Sigma}_i$. The mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights

from all component densities. These parameters are denoted by,

$$\Gamma = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \qquad i = 1, \ldots, M. \tag{19}$$

Thus, a GMM classifier consists of separate M-component full covariance GMM models for reliable and unreliable data independently for each frequency channel D. This makes the total number of GMMs $2 * D$. In this work, the training data is labelled into reliable and unreliable components based on oracle mask information as described in Section 3.3.

Mask estimation is based on using the GMMs for the reliable and unreliable classes as a scaled maximum likelihood classifier. First, the feature vector $\boldsymbol{o}(\tau, d)$ associated with the d-th channel in the τ-th frame is used to calculate the likelihoods $P_{d,\,r}(\boldsymbol{o}(\tau, d))$ and $P_{d,\,r}(\boldsymbol{o}(\tau, d))$ using Equations (17) and (18), where $P_{d,\,r}(\boldsymbol{o}(\tau, d))$ denotes the probability of $\boldsymbol{o}(\tau, d)$ evaluated on the GMM that represents the class of reliable feature components in channel d and $P_{d,\,u}(\boldsymbol{o}(\tau, d))$ the probability of $\boldsymbol{o}(\tau, d)$ evaluated on the GMM that represents the class of unreliable components in channel d. In maximum likelihood classification, the component would be classified as reliable or unreliable based on the likelihood scores, but in practice, the results improve if likelihood scores are scaled with a factor C as discussed in (Seltzer et al., 2004). In this work, a time–frequency component $Y(\tau, d)$ is classified as reliable if

$$C \cdot P_{d,\,r}(\boldsymbol{o}(\tau, d)) > P_{d,\,u}(\boldsymbol{o}(\tau, d)), \tag{20}$$

and unreliable otherwise. The scale factor C is optimized for a development data set and the two missing feature reconstruction methods as described in Section 3.3.

Since the GMM classifier processes each time-frequency component independently, the estimated mask can contain isolated reliable components that are unlikely to contain usable information (Cooke, 2006). In missing-feature reconstruction, even a single isolated reliable component can result in reconstruction errors and notably degrade the system performance. The estimated masks are therefore post-processed by removing groups of reliable features containing less than 20 connected reliable elements. The group size is optimized on the development data set for both imputation methods separately and also separately for the binaural reference mask. A similar post-processing step is also used in mask estimation in (Gemmeke et al., 2011).

2.2.4. Binaural reference mask

For the purposes of evaluation, we compare our GMM-based mask estimation scheme with a reference mask estimation method, which was the best-performing approach in our CHiME workshop paper (Kallasjoki et al., 2011). The reference method is based on the approach described by Harding et al. (2006), in which the detection of the target source in each time-frequency bin is based on joint statistics of interaural level (ILD) and time (ITD) difference cues.

In the training phase, ITD and ILD data are collected for target sources known to be at zero degrees azimuth in clean conditions. ILD is estimated as described in Section 2.2.1, whereas ITD is estimated via a standard cross-correlation rather than the GCC-PHAT approach described previously. For the training material 120 utterances were selected from the clean CHiME development set (increasing the amount of data decreased the performance). ILD–ITD distributions for the target speech were then represented by joint

ILD–ITD histograms collected independently for each frequency channel. For ITD, histogram bins were spaced at intervals of 62.5μs (the sampling period) and for ILD, the width of each histogram bin was 0.2 dB. Normalization was then performed by dividing every histogram bin value with the maximum value observed within the same histogram. The resulting histogram is denoted by $H_d[o_h(\tau, d)]$ with $o_h(\tau, d) = [ILD(\tau, d), ITD(\tau, d)]$, indicating the observation in time frequency component τ, d.

During recognition the mask elements $M(\tau, d)$ are set reliable for $H_d[o_h(\tau, d)] > \theta_b$, otherwise they are set as unreliable. Time–frequency regions that contain less than 10 connected reliable elements are removed from the masks. The threshold θ_b value was selected for cluster-based imputation such that it yielded the best keyword accuracy rate on the development sets with 6 dB and -6 dB SNRs.

This baseline mask estimation approach differs from the original Harding et al. (2006) paper in a number of respects; in particular, we use a binary mask and measure statistics from clean training signals, whereas Harding et al. used a real-valued mask and collected statistics from noisy speech data.

2.3. Reconstruction

Two imputation approaches are utilized in this work: the cluster-based imputation described in Section 2.3.1 and the sparse imputation algorithm described in Section 2.3.2. Both methods work on vectors formed by the concatenation of T consecutive mel-spectral feature vectors, in order to include some amount of time context in the reconstruction of a single observation.

Cluster-based imputation (Raj et al., 2004) is based on building a Gaussian mixture model, essentially a soft clustering solution, to represent the

distribution of the concatenated feature vectors of clean speech. The unreliable components of observed feature vectors are then filled by the most probable observation according to the model, under the constraints that the reliable components equal the actual observation, while the unreliable components are bounded by it. These bounds are based on the additive noise assumption, under which the energy of the clean speech signal cannot exceed that of the observed noisy signal.

In contrast, the sparse imputation approach (Gemmeke et al., 2010, 2011) represents the T consecutive features as a linear combination of sample clean speech segments (*exemplars*) of similar length, with the weights chosen to use as few exemplars as possible. In the case of feature imputation, the linear combination weights are estimated based only on the corresponding reliable components of the features, after which the unreliable components are set to the values given by the linear combination of the full clean speech exemplars, if the resulting values are lower than in the noisy observation. This selection of the lower value reflects the bounds used for cluster-based imputation.

2.3.1. Cluster-based imputation

When missing-feature reconstruction is applied on a continuous speech recognition task, the log-compressed spectral features are processed in T-frame windows where $T \geq 1$. Each window is represented as a TD-dimensional vector $\boldsymbol{s}(\tau)$, where D is the number of spectral channels in the observed data. In cluster-based imputation (Raj et al., 2004), the clean speech feature vectors $\boldsymbol{s}(\tau)$ are modeled as independent and identically distributed (*i.i.d.*)

samples from a Gaussian mixture model (GMM)

$$P(\boldsymbol{s}(\tau)) = \sum_m c(m)\mathcal{N}(\boldsymbol{s}(\tau); \boldsymbol{\mu}(m), \boldsymbol{\Sigma}(m)), \tag{21}$$

where $c(m)$ is the prior weight and $\boldsymbol{\mu}(m)$ and $\boldsymbol{\Sigma}(m)$ the mean vector and the full covariance matrix of the m-th GMM component. Maximum likelihood estimates for the model parameters $\Lambda = \{\boldsymbol{\mu}(m), \boldsymbol{\Sigma}(m)\}_m$ are obtained from clean speech training data using the EM algorithm.

Given the observed feature vector $\boldsymbol{y}(\tau)$ divided into reliable and unreliable components, the reconstructed features are calculated as a weighted sum of cluster-conditional bounded maximum a posteriori (MAP) estimates as proposed in (Raj et al., 2004). The reconstructed features are given as

$$\hat{\boldsymbol{s}}(\tau) = \sum_m \omega(m) \arg\max_{\boldsymbol{s}} \{P(\boldsymbol{s}|\boldsymbol{s}_r = \boldsymbol{y}_r(\tau), \boldsymbol{s}_u \leq \boldsymbol{y}_u(\tau), \Lambda, m)\}, \tag{22}$$

where $\boldsymbol{y}_r(\tau)$ and $\boldsymbol{y}_u(\tau)$ denote the vectors constructed from the observed reliable and unreliable components of $\boldsymbol{y}(\tau)$ in frame τ, and $\boldsymbol{s}_r$ denotes the clean speech vector components that correspond to $\boldsymbol{y}_r(\tau)$ and $\boldsymbol{s}_u$ the clean speech vector components that correspond to $\boldsymbol{y}_u(\tau)$. The weight $\omega(m)$ is the posterior probability of cluster m calculated as described in (Raj and Stern, 2005). The cluster-conditional bounded MAP estimation in Equation (22) can be formulated as a constrained optimization task,

$$\min_{\boldsymbol{s}} \{\frac{1}{2}(\boldsymbol{s} - \boldsymbol{\mu}(m))^{\mathsf{T}} \boldsymbol{\Sigma}(m)^{-1} (\boldsymbol{s} - \boldsymbol{\mu}(m))\} \tag{23}$$

subject to

$$\boldsymbol{s}_r = \boldsymbol{y}_r(\tau) \text{ and } \boldsymbol{s}_u \leq \boldsymbol{y}_u(\tau),$$

where $\boldsymbol{\mu}(m)$ and $\boldsymbol{\Sigma}(m)$ are the mean vector and full covariance matrix of the m-th component in Equation (21). The feature vector that optimizes Equation (23) corresponds to the m-th bounded MAP estimate in Equation (22). Note that the clean speech estimate for the reliable features is constrained to match the observed value.

Finally, when the frame shift between two consecutive windows is less than the window length T, several reconstructed vectors $\hat{\boldsymbol{s}}(\tau)$, which represent a multi-frame window centered at frame τ, will contain an estimate for the same time–frequency component $S(\tau', d)$. The clean speech estimates for the individual time–frequency components, $\hat{S}(\tau', d)$, are calculated as the average of all the components of reconstructed vectors $\hat{\boldsymbol{s}}(\tau)$ that represent the channel d of frame τ'.

2.3.2. Sparse imputation

The sparse imputation algorithm (Gemmeke et al., 2010, 2011) provides an alternative approach to the task of reconstructing the clean speech signal. Similarly to the cluster-based imputation method, the reconstruction is processed in T-frame windows, and the compressed spectral features are concatenated to give a single TD-dimensional vector $\boldsymbol{y}(\tau)$. The underlying clean speech vector $\boldsymbol{s}(\tau)$ is represented as a linear combination of *exemplars* (i.e. clean speech frames) of the same size,

$$\boldsymbol{s}(\tau) \approx \sum_{n=1}^{E} x_n(\tau)\boldsymbol{a}_n = \boldsymbol{A}\boldsymbol{z}(\tau), \tag{24}$$

where $\boldsymbol{z}(\tau)$ is the *activation vector* corresponding to the τ-th frame, and $\boldsymbol{A}$ is the $TD \times E$ sized fixed dictionary of clean speech.

For reconstruction, the activation vector components are estimated using the reliable components in each frame, while the reconstructed clean speech estimate is computed using the full clean speech exemplars. In particular, the activation vector $\boldsymbol{z}(\tau)$ is obtained by solving the minimization problem

$$\boldsymbol{z}(\tau) = \underset{\hat{\boldsymbol{z}} \in \mathbb{R}^E}{\operatorname{argmin}} \left\{ \left\| \boldsymbol{W}(\tau)\boldsymbol{A}\hat{\boldsymbol{z}} - \boldsymbol{W}(\tau)\boldsymbol{y}(\tau) \right\|_2 + \lambda \left\| \hat{\boldsymbol{z}} \right\|_1 \right\}, \tag{25}$$

where given the binary missing data mask $\boldsymbol{m}(\tau)$, the matrix $\boldsymbol{W}(\tau) = \operatorname{diag}(\boldsymbol{m}(\tau))$ is used to select only the reliable components of frame τ. The $\lambda \left\| \boldsymbol{z} \right\|_1$ term is a sparsity-inducing penalty in order to represent the observation using as few exemplars as possible.

After obtaining the activation vectors, a clean speech estimate is then reconstructed as $\boldsymbol{s}^*(\tau)$. The reliable components of the imputed features $\hat{\boldsymbol{s}}(\tau)$ are set to $\boldsymbol{y}(\tau)$, while the unreliable components are set to $\min\{\boldsymbol{s}^*(\tau), \boldsymbol{y}(\tau)\}$, reflecting the additive noise assumption, under which the observation gives an upper bound for the value of the clean speech features. Finally, as in the cluster-based imputation method, the clean speech estimates for the individual time–frequency components, $\hat{S}(\tau', d)$, are calculated as the average of all the components of reconstructed vectors $\hat{\boldsymbol{s}}(\tau)$ that represent the channel d of frame τ'.

3. Experimental setup

3.1. Data

The proposed system is evaluated using the CHiME challenge corpus described in (Barker et al., 2013). Here only a short overview is given. The CHiME challenge defines standard training, evaluation and test sets. The

speech material in CHiME is taken from the Grid corpus (Cooke et al., 2006) which is reverberated with binaural room impulse responses (BRIRs) measured from a dummy head at a source-to-receiver distance of 2 meters. A typical Grid corpus utterance could be, for example, "*bin green in r eight please*". Environmental noises were recorded in a family home over a long period of time, using the same dummy head setup described above. The standard training set consists of 17000 utterances of reverberated (but noise-free) speech. The CHiME development and evaluation sets consist of 600 different utterances with shared speakers. Speech in the development and evaluation sets are reverberated and mixed with environmental noise samples at SNRs ranging from -6 to 9 dB at 3 dB intervals. Neither the speech nor the noise is scaled to obtain a desired SNR; rather, particular SNRs were obtained by selecting noise samples of an appropriate intensity.

In addition, we constructed a multicondition training set by mixing the standard clean training set with utterance-length samples extracted from random locations in the background noise recordings. Following the process used for the CHiME test sets, no scaling was applied to either the speech or the noise samples. Rather, the selected noise samples were chosen to provide an approximately uniform SNR distribution in the -6 dB to 9 dB range for the utterances of the final multicondition training set.

3.2. Speech recognition system and setup

The baseline system used in this work is a large vocabulary continuous speech recognizer (LVCSR) based on hidden Markov models (HMM) with state likelihoods modeled by Gaussian mixture models and trained on CHiME training data. The speaker independent acoustic models of the sys-

tem are state-tied triphones constructed with a decision-tree method. The triphone-level segmentations of the CHiME training data were generated by a LVCSR trained on the Wall Street Journal British English (WSJCAM0) corpus (Robinson et al., 1995) in "forced alignment" mode. Each state is modeled with at most 100 Gaussian components and the state durations are modeled with gamma distributions. Each frame of the speech signal is represented by 12 MFCC features and a frame power feature, together with their first- and second-order temporal derivatives. Post-processing of the features is done by applying cepstral mean subtraction (CMS) before scaling and mapping with a maximum likelihood linear transformation (MLLT) (Gales, 1999) optimized in training. Finally, the covariance matrices of the Gaussians are diagonalized. A more detailed description of the baseline system is found in (Hirsimäki et al., 2006).

For language modeling, a no-backoff bigram (i.e. word-pair) model with uniform frequencies for all valid bigrams is constructed to restrict recognized sentences to conform to the restricted Grid utterance grammar specified in (Cooke et al., 2006). The training data described in Section 3.1 was used to train a clean baseline, multicondition baseline and multicondition model for imputed data ASR-systems. The multicondition model for imputed data was trained on the multicondition training set processed by cluster-based imputation using GMM estimated masks (i.e. the best performing MD system; see Tables 3 and 4). The development set was used for parameter adjustment and training the mask estimation GMMs (see the detailed description in Section 3.3). Finally, an independent evaluation was performed on the standard evaluation set. For additional speaker adaptation tests, unsupervised con-

strained maximum likelihood linear regression (CMLLR) was applied to both multicondition trained systems to further improve the accuracy rate. The adaptation data for an individual speaker was obtained from the first-pass recognition hypotheses of the corresponding multicondition trained system using all the SNRs for the speaker in question of either the development or evaluation set.

Our baseline system is evaluated against a CHiME baseline system in Table 5. The differences between the CHiME challenge baseline and our baseline system are that the CHiME baseline system is trained speaker dependently and the words are modeled as whole-word HMMs. Our baseline also utilizes MLLT post-processing in addition to CMS.

3.3. Missing feature reconstruction and mask estimation setup

The GMMs used for mask estimation were trained using the EM algorithm implemented in the GMMBAYES Matlab toolbox (Kämäräinen and Paalanen, 2005). The model parameters were initialized with fuzzy c-means and the EM algorithm was used for at most 100 iterations. The training data was constructed from 1800 sentences randomly selected from the development data (all SNRs) as follows. First, oracle masks were constructed based on stereo data (noisy and clean) that allows to exactly calculate the local SNR in each time–frequency component. In the oracle mask estimation, a time–frequency component $Y(\tau, d)$ is considered reliable if $\big(10\log_{10}(\exp S(\tau,d)) - 10\log_{10}(\exp Y(\tau,d) - \exp S(\tau,d))\big) > \theta$, where $Y(\tau, d)$ denotes the noisy log-mel-spectral feature component and $S(\tau, d)$ the clean log-mel-spectral feature component, and θ is a threshold parameter in decibels determined based on speech recognition experiments on the development

data. The threshold θ giving the best keyword accuracy rate is optimized separately for each SNR in the development set and separately for cluster-based (θ ranges from 0 dB to 2 dB depending on SNR) and sparse imputations (θ is between -2 dB and 2 dB). Separate threshold optimizations for both imputation methods are required since the optimal values depend on the imputation method. For example, the optimal θs for cluster based imputation are 0 dB for SNRs of 6 dB and 0 dB, whereas the respective θs for sparse imputation are both 2 dB. Each time–frequency component in the noisy training data was represented as an N-dimensional feature vector $\boldsymbol{o}(\tau, d)$ and the reliable–unreliable classifications were used to divide the training samples into two sets that were further divided into separate sets for each frequency channel d.

In this work, the feature vectors $\boldsymbol{o}(\tau, d)$ are 14-dimensional, which are referred to as a *full* feature set (with features described in Section 2.2.2). We also use other systems in comparisons presented in Sections 2.2.4 and 4.1.3, that use only a subset of features (feature vectors are either two or 13-dimensional). Based on experiments on the CHiME development data set, a 14-component GMM was chosen to model each channel-dependent training dataset. Note that separate training datasets and therefore separate GMM classifiers were constructed for cluster-based imputation and sparse imputation because the above described optimal oracle mask thresholds were different.

The scale factor C introduced in Section 2.2.3 was determined based on experiments on the development data, which indicated that the optimal value is $C = 1$ for cluster-based imputation and $C = 0.6$ for sparse imputation.

The missing feature components are reconstructed in the $D = 21$-dimensional log-compressed mel-spectral domain. In cluster-based imputation, the features are processed in $T = 5$ frame windows with a window shift of $\Delta = 1$ frame. 1500 randomly selected utterances from the CHiME training set were used to train a 13-component clean speech GMM with 105-variate component densities and full covariance matrices. The model was trained using the EM implementation in the GMMBAYES toolbox. The simply constrained quadratic programming problem in Equation (23) was solved with an active-set method implemented in the QPC toolbox (Willis, 2010).

The sparse imputation algorithm was applied using windows of $T = 15$ frames and a window shift $\Delta = 1$ frame. The window size was selected based on small-scale tests on the CHiME development data and the clean speech dictionary consisted of 34000 randomly selected 15-frame exemplars, which resulted in 315 dimensional exemplar vectors, taken from the CHiME training set. For the sparse imputation, a Matlab implementation by Gemmeke et al. (2011) was used.

Increasing the model size i.e. increasing the number of components in the GMM or using a larger dictionary for CI and SI, respectively, may improve the performance at the cost of an increase in computation time. For both methods, pilot investigations (not shown) have shown that any improvement is only minor and not worth the increased computational effort.

4. Experiments

4.1. Feature analysis

The mask estimation approach discussed in Section 2.2 classifies each time–frequency component as reliable or unreliable based on a vector containing $N = 14$ features. The features are constructed to utilize known speech characteristics or binaural cues in order to discriminate between the reliable, speech-dominated components and the unreliable, noise-dominated components. To analyze how well each feature discriminates between the reliable and unreliable data, the distributions of reliable and unreliable features are compared, and mask estimation experiments using feature sets with $N-1$ features are conducted. The experiments were conducted on the SNR 6 dB and 0 dB development data described in Section 3.1, and the reliable and unreliable components were determined based on oracle masks with a 0 dB threshold. Using a 0 dB threshold means that a component is interpreted as speech-dominated when the speech energy exceeds the noise energy, and noise-dominated when the noise energy exceeds the speech energy. 0 dB is also the optimal threshold for cluster-based imputation in SNR 6 dB and 0 dB conditions, whereas the optimal for sparse imputation is 2 dB.

4.1.1. Feature histograms

In this experiment, the reliable and unreliable feature distributions are represented as normalized histogram vectors. The histogram vectors were constructed as follows. The N-dimensional data vectors $\boldsymbol{o}(\tau, d)$ are each associated with an oracle mask label that indicates whether the vector describes a reliable or an unreliable time–frequency component $Y(\tau, d)$. Based

on the mel-spectral channel d and oracle mask label, the data vectors can be divided into $D \cdot 2$ sets, where D is the number of mel-spectral channels. The components of the data vector $\boldsymbol{o}(\tau, d)$ correspond to the different features used for mask estimation, so we represent the distribution of each component within each of the $D \cdot 2$ datasets as a normalized histogram vector $\boldsymbol{h}$. The histogram vector dimension depends on the feature n, for continuous features are represented with 1000-bin histograms but discrete features are not represented with more histogram bins than they have discrete values. A histogram vector calculated for the n-th feature based on a set where the data vectors are associated with channel d and labeled reliable is denoted as $\boldsymbol{h}_r(n, d)$ and a histogram calculated based on a set where the vectors are labeled unreliable as $\boldsymbol{h}_u(n, d)$.

In order to rank features based on the dissimilarity between their reliable and unreliable values, the normalized histogram vectors were compared using two dissimilarity measures: *cosine distance*, which measures similarity based on the normalized inner product between the vectors, and *information radius* or Jensen–Shannon divergence, which measures the difference in relative entropy between the distributions P and Q and the average distribution $(P + Q)/2$. $L1$ and Bhattacharyya distances were also computed but the results were similar to the results of cosine distance or information radius, respectively, and are hence not presented. For a comprehensive survey of dissimilarity measures, see (Cha, 2007).

To measure the feature discrimination power, the dissimilarity between the normalized histogram vectors $\boldsymbol{h}_r(n, d)$ and $\boldsymbol{h}_u(n, d)$ for each feature n and each frequency channel d was calculated. The results are displayed in

Figure 1: Dissimilarity between the reliable and unreliable feature distributions measured using cosine distance and information radius. The cosine distances and information radiuses between the reliable and unreliable histograms of each feature in each channel are displayed on a scale from 0 (white) to 0.18 (black). The average across-channel cosine distance or information radius of each feature is displayed as a separate plot. Results from the SNR 6 dB and 0 dB conditions are reported in separate plots. Note that the differences between the results from the two conditions are due to differences in both the noise level and noise type.

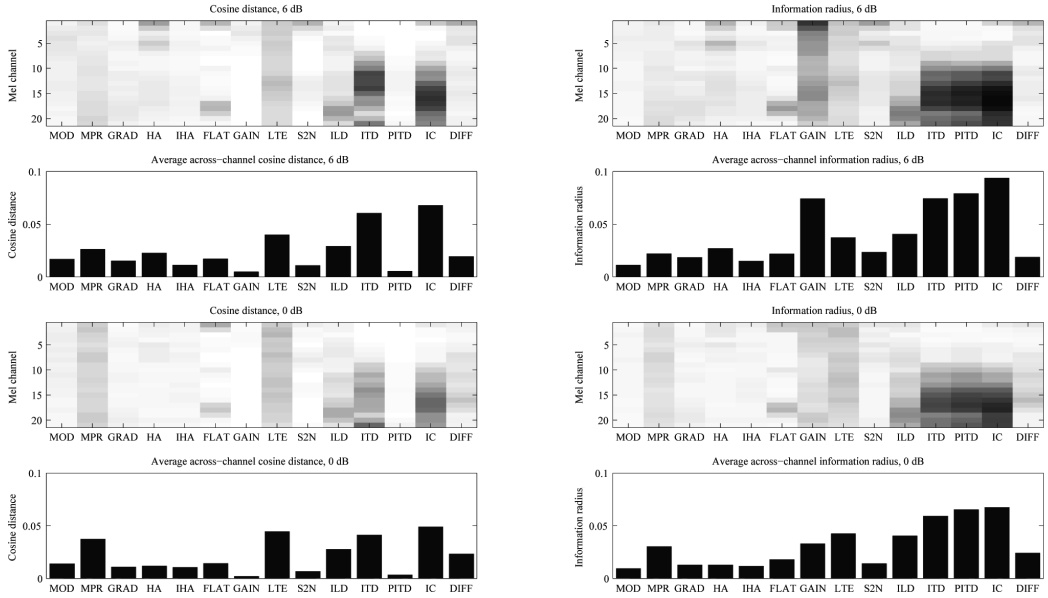


Figure 1. Both dissimilarity measures give the reliable and unreliable feature distributions low dissimilarity scores in general. The average cosine distance is 0.02 and the average information radius 0.03 on a scale from 0 to 1 where 0 indicates exactly overlapping distributions and 1 indicates that the distributions do not overlap. Highest dissimilarity scores are observed between

the reliable and unreliable ITD and interaural coherence distributions in the mel channels above $d = 10$. Additionally, the dissimilarity scores between the reliable and unreliable LTE distributions are consistently above average in every mel channel.

While the dissimilarities measured with cosine distance and information radius mostly follow the same pattern, a few notable differences may be observed. For example, information radius indicates an above-average dissimilarity between the reliable and unreliable peak ITD distributions in the mel channels above $d = 10$, whereas their cosine distance is almost zero. This is because the cosine distance emphasizes differences between histogram bins with large normalized counts and because the peak ITD distributions have two components: a continuous, unimodal component and a peak at PITD = 1. The continuous parts in the reliable and unreliable peak ITD distributions have an average across-channel cosine distance of 0.04, but since 1 is the single most observed value in both reliable and unreliable components, the concurrent peaks dominate the inner product, and hence, the cosine distance between the distribution vectors.

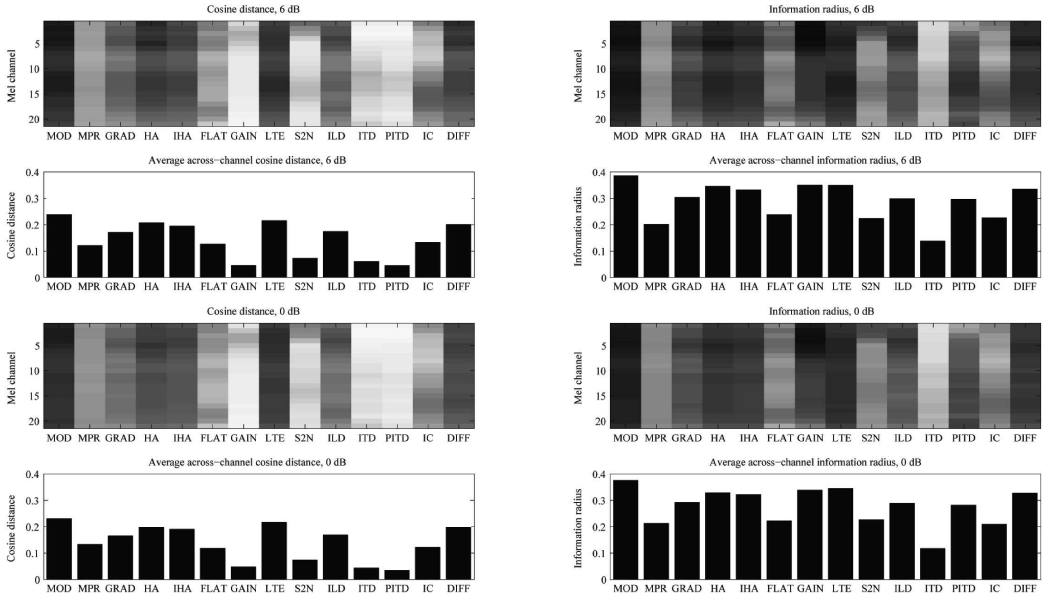
Finally, we have also calculated the average of the measured channel–dependent dissimilarities (Figure 1). We note that while the average across-channel dissimilarities conveniently characterize the discrimination power of each feature with one number, the unweighted average does not completely correlate with the missing data task performance. This is because the mask estimation errors in different frequency channels do not have an equal effect in the missing-feature reconstruction performance; errors in certain channels have a larger effect in the overall result (Gemmeke et al., 2008).

4.1.2. Feature-pair histograms

Since the features are used together with other features rather than independently, experiments with feature combinations were also conducted. In these experiments, two-feature combinations were considered. Normalized feature-pair histograms were calculated from the same $D \cdot 2$ datasets that were used to calculate the feature histograms in Section 4.1.1. The histogram vectors constructed for feature pairs are different from the histogram vectors constructed for single features in that each histogram bin defines an applicable range for both features rather than one feature. Other than using 2-dimensional bins and data points, the feature-pair histograms are constructed as described in Section 4.1.1. A histogram vector calculated for the n-th and n'-th feature based on a set where the data vectors are associated with channel d and labeled reliable is denoted as $\boldsymbol{h}_r(n, n', d)$ and a histogram calculated based on a set where the vectors are labeled unreliable as $\boldsymbol{h}_u(n, n', d)$.

The normalized histogram vectors were compared using the cosine distance and information radius. Again, L1 and Bhattacharyya distances were found to be very similar to the results of cosine distance and information radius and hence are not presented. To facilitate the comparisons between individual features, the average dissimilarities across all configurations with the given feature n were calculated. The results are displayed in Figure 2. The average dissimilarity between feature pair distributions is 0.12 with the cosine distance and 0.24 with the information radius. Both dissimilarity measures suggest that while the modulation filtered spectrogram, gradient, harmonic and inharmonic energy, and channel difference were not particularly

Figure 2: The dissimilarity scores reported here are the average cosine distance or information radius between the reliable and unreliable feature-pair distributions that include the indicated feature. The average cosine distances between the reliable and unreliable feature-pair histograms that include the indicated feature are displayed for each channel in scale from 0 (white) to 0.3 (black) and the information radiuses in scale from 0 (white) to 0.45 (black). The average across-channel cosine distance and information radius are displayed separately and the results from the SNR 6 dB and 0 dB conditions reported in separate plots.



effective in separating between reliable and unreliable components when analyzed independently, they notably increase the dissimilarity between reliable and unreliable component distributions when combined with other features. In addition, the dissimilarity scores between the reliable and unreliable LTE feature distributions are above average in every mel channel, and finally, the information radius between the reliable and unreliable gain feature distribu-

tions is notable especially in the mel channels $d = 2 \ldots 5$.

4.1.3. Feature discrimination power in mask estimation

Mask estimation experiments with a 7-component GMM classifier trained either using the full feature set or a set with $N - 1$ features was conducted. The classifier was trained on a 600 utterance training dataset constructed by randomly selecting 100 utterances from each development set SNR. The estimated masks were (a) compared to the oracle masks and (b) evaluated in the missing feature reconstruction and keyword recognition task described in Section 3.1. The features were reconstructed using cluster-based imputation. Comparison with the oracle masks allows us to determine how often a component labelled as reliable corresponds to a reliable component in the oracle mask (precision) and how often a reliable component is labelled as reliable in the estimated mask (recall). The F_1 scores calculated based on comparison with the oracle masks and the keyword accuracies obtained in the reconstruction and recognition task are reported in Table 2.

Using all $N = 14$ features in mask estimation results in the best F_1 score and also in the best keyword accuracy when the average results from 6 dB and 0 dB conditions are compared. The features whose removal most degrades the system performance are inharmonic energy, flatness, and channel difference. Additionally, when mean to peak ratio is not included in the feature set, the keyword accuracy degrades in the 6 dB condition where removing a single feature often improves the performance. We find it noteworthy that inharmonic energy, flatness, and channel differences, whose removal most degrades the results, have continuous, unimodal distributions, i.e. distributions that a GMM can efficiently model. The histogram experiments, which re-

Table 2: (a) F_1 scores and (b) keyword accuracies (%) measured on the development data when each feature is in turn excluded from the feature set used for mask estimation. The labels indicate which feature is not used in the experiment, and "none" indicates that all features are used. An underlined average result suggests that the feature is one of the three most important features for mask estimation.

(a)

	MOD	MPR	GRAD	HA	IHA	FLAT	GAIN	LTE	S2N	ILD	ITD	PITD	IC	DIFF	None
6 dB	0.79	0.79	0.79	0.79	0.77	0.77	0.79	0.79	0.80	0.80	0.80	0.80	0.78	0.76	0.80
0 dB	0.71	0.71	0.71	0.70	0.69	0.67	0.70	0.71	0.70	0.72	0.71	0.72	0.70	0.68	0.72
Avg.	0.75	0.75	0.75	0.74	<u>0.73</u>	<u>0.72</u>	0.75	0.76	0.75	0.76	0.76	0.76	0.74	<u>0.72</u>	0.76

(b)

	MOD	MPR	GRAD	HA	IHA	FLAT	GAIN	LTE	S2N	ILD	ITD	PITD	IC	DIFF	None
6 dB	83.3	81.9	82.9	82.1	81.7	82.2	82.6	83.1	82.8	83.0	82.8	83.6	82.8	82.2	82.3
0 dB	63.0	62.8	62.7	63.8	60.9	60.4	63.7	63.9	64.1	61.8	62.6	62.3	63.1	60.9	65.0
Avg.	73.1	72.3	72.8	72.9	<u>71.3</u>	<u>71.3</u>	73.1	73.5	73.4	72.5	72.7	73.0	73.0	<u>71.6</u>	73.7

sulted in a different ranking for the features, did not take such factors into account. The histogram experiments also did not consider $N - 1$ feature combinations but individual features or feature pairs.

4.2. Speech recognition performance of imputation and mask estimation methods

The speech recognition results are collected in Tables 3, 4 and 5, and the system ranking is based on their average results. The highest scores on each evaluation set SNR is shown in bold type. First, cluster-based imputation (CI) was evaluated with three mask estimation methods: 14-component GMM classifier trained on the full 14-feature set (GALL), 14-component GMM classifier trained on ILD–ITD pairs (GBIN), and an ILD–ITD his-

Table 3: Keyword accuracy rates of cluster-based imputation (CI) using binaural histogram (HBIN), binaural GMM (GBIN) and GMM (GALL) mask estimation techniques for the CHiME development and evaluation sets.

	Devel set							Eval set						
	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CI+HBIN	87.8	83.9	75.3	62.4	53.1	47.8	68.4	88.6	79.8	70.8	58.9	47.4	46.3	65.3
CI+GBIN	88.7	84.3	77.1	67.1	56.8	52.6	71.1	88.5	83.2	73.5	63.6	54.9	48.6	68.7
CI+GALL	88.6	85.3	78.1	68.6	60.6	55.1	72.7	**90.3**	**84.3**	**76.9**	**68.2**	**58.2**	**56.3**	**72.3**

togram (HBIN). The keyword accuracy rates obtained with the mask estimation methods are compared in Table 3. HBIN mask estimation receives the lowest scores in almost in every SNR condition on both development and evaluation data sets, with respective averages of 68.4% and 65.3%. GBIN, with an average of 71.1% for the development set and 68.7% for the evaluation set, outperforms HBIN at every SNR, excluding the 9dB evaluation set. GALL achieves the highest average scores, of 72.7% and 72.3% on the development and evaluation sets respectively. Based on these results, GMM mask estimation trained on the full feature set (GALL) was selected for subsequent evaluation.

Statistical significance of the keyword accuracy difference between each system pair on the evaluation set was computed by the Wilcoxon signed-rank test with a 95 % confidence level. The details of the statistical analysis are presented in Appendix A, with the results of SNR-wise system comparisons shown in Tables A.6 and A.7. In this section, we present only the statistics of the pairwise comparisons based on the average results, except for the

Table 4: Keyword accuracy rates of cluster-based imputation (CI) and sparse imputation (SI) using oracle (ORA) and GMM (GALL) estimated masks for the CHiME development and evaluation sets.

	Devel set							Eval set						
	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CI+ORA	93.5	93.0	92.3	90.1	86.2	86.6	90.3				N.A.[a]			
CI+GALL	88.6	85.3	78.1	68.6	60.6	55.1	72.7	**90.3**	**84.3**	**76.9**	**68.2**	**58.2**	**56.3**	**72.3**
SI+ORA	93.7	93.8	92.3	90.5	90.3	88.2	91.5				N.A.[a]			
SI+GALL	81.7	80.3	68.1	59.5	49.5	42.7	63.6	84.3	78.3	67.3	57.1	44.2	42.7	62.3

[a] Oracle masks can not be generated for the evaluation set.

statistics computed for Table 4 which are fully described. In Table 3, all the differences between the system averages are statistically significant.

Keyword accuracy rates for cluster-based imputation (CI) and sparse imputation (SI) using oracle masks (ORA) and masks estimated by the GMM mask estimation (GALL) for the CHiME development and evaluation sets are collected in Table 4. When oracle masks are used, the accuracy of SI is higher or equal to the accuracy of CI in every development set SNR condition. Respective averages for SI and CI are 91.5% and 90.3%. Both imputation systems using oracle masks outperform those using GMM estimated masks (GALL) by a large margin, especially at low SNRs. However, when GMM estimated masks are used, CI achieves higher accuracies in every development and evaluation set SNR condition. Development and evaluation set averages for CI with GMM estimated masks are 72.7% and 72.3%, and the respective averages of SI are 63.6% and 62.3%. For further system analysis, CI was selected as the imputation method. Statistically significant differences are

Table 5: Keyword accuracy rate improvements over CHiME baseline (CBL) and our baseline (BL) for CHiME development and evaluation sets. Abbreviations: Cluster-based imputation using GMM estimated masks (CI+GALL), multicondition trained system (MC), multicondition models for CI+GALL imputed data (MCI), unsupervised speaker adaptation (SA).

	Devel set							Eval set						
	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CBL	83.1	73.8	64.0	49.1	36.8	31.1	56.3	82.5	75.0	62.9	49.5	35.4	30.3	55.9
BL	83.3	80.0	69.8	55.2	46.0	40.6	62.5	86.3	78.3	68.5	53.9	44.3	41.9	62.2
CI+GALL	88.6	85.3	78.1	68.6	60.6	55.1	72.7	**90.3**	84.3	76.9	68.2	58.2	56.3	72.3
MC	87.7	85.7	80.9	69.2	62.2	54.4	73.3	88.4	84.3	78.8	71.3	61.3	53.9	73.0
MC+SA	88.1	85.9	81.7	70.5	63.0	56.8	74.3	89.2	84.9	79.7	72.0	62.4	54.7	73.8
MCI	88.3	87.5	83.0	73.1	65.5	60.9	76.4	89.0	86.3	82.4	74.3	65.2	61.8	76.5
MCI+SA	88.3	88.5	83.3	73.8	65.9	61.4	76.9	89.6	**86.7**	**83.2**	**75.4**	**65.7**	**62.3**	**77.1**

found for all pairwise system comparisons in Table 4 on keyword spotting accuracy over each evaluation set SNR and average.

In Table 5, keyword accuracy rates of the CHiME baseline system (CBL) and our own baseline system (BL) are evaluated against the multicondition trained systems and the best performing missing data system. The BL system achieves higher rates than CBL in every SNR condition on both data sets. The development set averages for CBL and BL are 56.3% and 62.5%, and for evaluation set 55.9% and 62.2%, respectively. Cluster-based imputation using GMM estimated masks outperforms BL on every SNR condition for both data sets, with averages of 72.7% on the development set and 72.3% on the evaluation set (CI+GALL). The multicondition trained system (MC) gives

average rates of 73.3% and 73.0% for the development and evaluation sets, respectively, which are comparable to CI+GALL; higher rates are obtained by CI+GALL on 9 dB and −6 dB cases on both sets, otherwise MC performs better or equal. Applying unsupervised speaker adaptation (SA) to MC, all the rates in every SNR condition on both data sets are improved. Averages of 74.3% and 73.8% are achieved for the development and evaluation sets, respectively. The multicondition model for imputed data (MCI) outperforms MC+SA almost in all SNR conditions and the average rates are increased to 76.4% in the development set and to 76.5% in the evaluation set. Comparing MCI to CI+GALL system, relative improvements of 2.4%, 7.2%, 8.9%,12.0%, and 9.8% are obtained on SNRs ranging from 6 dB to −6 dB, respectively. Applying SA to the MCI system, average rates of 76.9% and 77.1% are reached for the development and evaluation sets, which are either equal or slightly increased over MCI in every SNR condition on both sets.

We also observe that systems based on MCI gain higher average rates on the evaluation set than on the development set, which is not observed for other cases. Overall, MCI+SA receives the highest accuracy rates on almost every SNR condition of the evaluation set. Comparing the best performing method MCI+SA to BL, relative improvements ranging from 3.8% to 48.7% are observed.

All the differences in the average accuracies of the evaluation set in Table 5 are statistically significant except between system pairs CI+GALL and MC, and CI+GALL and MC+SA.

5. Discussion

In this study, a mask estimation method based on a GMM classifier utilizing a set of fourteen features was proposed including both monaural and binaural representations. Notably the set included two new binaural features (peak ITD and interaural coherence) not previously used in the context of missing data reconstruction. The proposed method was evaluated against two kinds of binaural reference masks; one adapted from (Harding et al., 2006) and the other based on ILD–ITD features applying the GMM classifier. The evaluations were performed using CHiME corpus in applying missing data reconstruction with cluster-based (CI) and sparse imputation (SI) methods. The proposed method using the full feature-set outperformed the reference methods using only ILD and ITD. Regarding comparisons of the two imputation methods, CI clearly outperformed SI when the estimated masks were used. Additional performance improvements were obtained with speaker adaptation and re-training HMMs in the ASR back-end for the reconstruction error.

While the present work on the CHiME challenge is based on a rather small vocabulary recognition, our ultimate goal is to apply the proposed methods in large vocabulary tasks. This justifies our choice of focusing on imputation approaches that make it possible to use cepstral or other decorrelated features, rather than on classifier modifications that require spectral features. It also justifies the use of our own complex LVCSR-system rather than the CHiME baseline system. We also note that our LVCSR baseline system produced better results than the CHiME baseline.

Section 2.2.4 described our adaptation of Harding et al.'s binaural histogram-

based mask estimation technique (Harding et al., 2006), which provided reference masks (HBIN) for the present study with feature representation based on ILD and ITD cues alone. Our implementation of Harding et al.'s scheme does not gather statistics from noisy conditions, as was done in their paper – instead, we compare observed ILD–ITD feature pairs to a histogram that represents the ILD–ITD distribution for clean speech. A more faithful implementation of Harding et al.'s approach was also tried on CHiME data, but poor results were obtained; most likely this was due to the noise background in the CHiME corpus, which is much more spatially diffuse than the strongly localized noise sources used in Harding et al.'s study.

The reasons why the proposed mask estimation technique outperforms the reference HBIN system are now considered. Most importantly, our approach uses a diverse range of fourteen features (including ILD and ITD), whereas the HBIN system uses only ILD and ITD. Secondly, the data representations in the classifier model are different; our approach uses GMMs for both the reliable and unreliable data, whereas HBIN uses histograms only for reliable data. To investigate the role of the ILD–ITD data representation, we also tested a binaural reference mask that was directly comparable to the full feature set approach (GALL). We therefore trained a GMM classifier for ILD–ITD data, with settings of the model (GBIN) that exactly matched the main approach (GALL), including the same number of Gaussians and separate models for reliable and unreliable components. The GBIN approach provided a performance improvement over HBIN, but was poorer than GALL. In addition, it is worth noting that our implementation of ITD estimation (in GBIN and GALL) uses normalized GCC-PHAT cross-correlation, which

provides more accurate time delay estimation than the conventional cross-correlation (Perez-Lorenzo et al., 2012) used in HBIN and in the original work of Harding et al. (2006). Taken together, GBIN most likely gave a performance improvement over HBIN because the GMMs were better able to construct a representation of the data, and some additional improvement could have been achieved by using a more accurate ITD estimation algorithm. However, we can not completely rule out the possibility that there may still exist a substantially better configuration for the histograms used in HBIN. Thus with more effort to refine the parameters of a histogram based representation, one could possibly improve also the performance of the HBIN approach.

There are a several probable reasons why the full fourteen feature set mask estimation was more effective than the ILD–ITD pair based estimation. First, the classifier had access to interaural coherence, and may therefore have placed less weight on ILD and ITD cues with low coherence. Second, ILD and ITD are degraded by reverberation to some extent, but the 14-feature set contains features (e.g. modulation filtered spectrogram) that are more reverberation-robust. Finally, the noise backgrounds contain interference from a range of spatial locations; if the noise and target locations overlap, then monaural cues are needed to determine the mask.

In the experiments where a single feature was removed from the mask estimation feature set, we observed that in the SNR 6 dB condition, using one less feature often improved the keyword accuracy. This may be due to correlations between the different features, and suggests that the feature set used for mask estimation could benefit from dimensionality reduction. We

also note that the results in Table 2 represent a GMM that was essentially trained from a data dependent initialization. Depending on the initialization (i.e. the feature set), training may have focused on minimizing the error between the GMM and sharp, discrete peaks such as the peak in the PITD distribution (discussed in Section 2.2.2). It is therefore possible that some features may have potential that we were not able to exploit with a GMM classifier. A support vector machine (SVM), which does not attempt to model the class distributions but the decision border between classes, could be used with the proposed mask estimation method for improved classification accuracy. Further classification approaches will be investigated in the future.

Feature analysis indicated that the reliable and unreliable feature distributions are most dissimilar when binaural features are used (Figure 1), and in the mask estimation experiments, using a GMM classifier trained on ILD–ITD pairs resulted in acceptable missing feature reconstruction and speech recognition performance (Table 3). However, the results from mask estimation experiments in Tables 2 and 3 indicate that on average, the best performance is obtained with the full feature set that includes both binaural and monaural features, and the feature pair analysis (Figure 2) suggested modulation filtered spectrogram is more important than the binaural features when features are used as a set. Like binaural features, the modulation filtered spectrogram can indicate speech components corrupted with reverberation.

This combination-of-features-theory is also promoted by the observation that the proposed mask estimation method combined with cluster-based im-

putation imparts high generalization power, since the decrease in average accuracy rate is considerably smaller than on the reference methods when shifting from the development set to the evaluation set (Table 3). The multicondition trained system and multicondition model for imputed data offer similar generalization power with higher performance.

Considering the comparison between the two imputation methods, it is perhaps surprising that CI achieves higher accuracies than SI when estimated masks are used. In previous work, SI achieved substantially higher accuracies than CI in a large vocabulary task using the SPEECON corpus (Iskra et al., 2002), although also in that work, the gap between oracle mask and estimated mask recognition accuracies was large. In that work however, no time-context was used in CI, whereas in (Remes et al., 2011) CI using multi-frame windows showed superior performance over SI on speech corrupted by impulsive noises, while SI did outperform CI on speech corrupted with babble noise. From this, we conclude that SI is more sensitive to the type of mask estimation errors made on the noise types with which the CHiME data is corrupted.

It is known that SI is more sensitive to features incorrectly labeled as reliable, as the presence of these *false reliables* results in finding incorrect exemplar representations (Iskra et al., 2002). The fact that $C < 1$ for SI (resulting in fewer reliable features) while the oracle mask thresholds do not differ much from those used in CI shows that this issue is relevant to the CHiME dataset. Moreover, we observe that even at the highest SNRs, there is a substantial drop in recognition accuracy when compared to the use of oracle masks. Although a detailed analysis of the mask estimation errors that lead to a lower performance of SI with respect to CI is out of the scope of this

paper, these findings lead us to hypothesize that the mask estimation method presented in this work consistently labels some features incorrectly as reliable. Future work will have to test the validity of this hypothesis, for example by experimenting with missing data masks containing only mask estimation errors on reliable features or on unreliable features. Another, perhaps more pragmatic line of research would be to investigate to what extent the retraining of the acoustic model on SI-processed speech can compensate for the errors introduced by SI.

In conclusion, in this paper we presented a mask estimation method that employs a comprehensive set of features, which focus on reverberation, noise robustness and on monaural and binaural aspects of the noisy speech signal. We presented an analysis of the discriminative power of each of the features as well as evaluations on a noise robust speech recognition task. Although the performance gap with error-free oracle masks remains large, the evaluations show that the presented mask estimation technique works substantially better than previously used mask estimation methods that only employ a subset of these features.

Finally, it should be noted that the standard approach in missing data techniques has been to use models in the ASR back-end that have been trained using clean (noise free) speech (Cooke et al., 1994), which is in contrast to many practical speech recognition applications that commonly use multicondition training. In the present study training the system on the imputed speech (MCI) was rather successful. Improvements using the MCI systems were at their largest at low SNRs. This could be explained by the fact that at low SNRs both mask estimation and imputation are more prone

to errors than at high SNRs, and the effect of multicondition training is expected to counteract these errors. We would also expect that in the CHiME data the improvements due to multicondition training might be relatively large as the data is very challenging for both the mask estimation and imputation due to non-stationary noises and interference from competing speech. In conclusion, a key finding from the present study is that the imputation missing data approach can benefit from multicondition training when trained on imputed speech, which in turn might make missing data techniques a more viable option for practical noise robust ASR applications.

6. Acknowledgments

The work was supported by Langnet (Sami Keronen) and Hecse (Heikki Kallasjoki, Ulpu Remes) graduate schools, and by the Academy of Finland projects 136209 (Kalle J. Palomäki) and AIRC (Sami Keronen, Heikki Kallasjoki, Ulpu Remes, Kalle J. Palomäki). The research of Guy J. Brown and Jort F. Gemmeke was supported by EPSRC grant EPG0098051 and by IWT-SBO project ALADIN contract 100049, respectively.

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

Appendix A. Statistical analysis

For the Wilcoxon signed-rank test in separate SNR cases, the recognition data of two compared systems are paired speakerwise (with CHiME data, the number of ranked pairs i.e. speakers is 34) and the ranking is based on the correct keyword counts. For computing the statistics of the average results, the correct keyword counts are collected speakerwise from all SNR cases so that the number of ranked pairs remains the same as in the separate SNR cases.

Statistical significances of pairwise system comparisons of the evaluation set for Table 3 are gathered in Table A.6 and the statistics of Table 5 are gathered in Table A.7. In Table 4, all differences in pairwise system comparisons are statistically significant thus no separate statistics table is presented.

Table A.6: Statistical significances of pairwise system comparisons of the evaluation set presented in Table 3. "−" and "+" denote negative and positive statistical significance, respectively.

Pair	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CI+HBIN – CI+GBIN	-	+	+	+	+	+	+
CI+HBIN – CI+GALL	-	+	+	+	+	+	+
CI+GBIN – CI+GALL	+	-	+	+	+	+	+

Table A.7: Statistical significances of pairwise comparisons of the evaluation set presented in Table 5. "–" and "+" denote negative and positive statistical significance, respectively.

Pair			9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
CBL	–	BL	+	+	+	+	+	+	+
CBL	–	CI+GALL	+	+	+	+	+	+	+
CBL	–	MC	+	+	+	+	+	+	+
CBL	–	MC+SA	+	+	+	+	+	+	+
CBL	–	MCI	+	+	+	+	+	+	+
CBL	–	MCI+SA	+	+	+	+	+	+	+
BL	–	CI+GALL	+	+	+	+	+	+	+
BL	–	MC	+	+	+	+	+	+	+
BL	–	MC+SA	+	+	+	+	+	+	+
BL	–	MCI	+	+	+	+	+	+	+
BL	–	MCI+SA	+	+	+	+	+	+	+
CI+GALL	–	MC	+	-	-	+	-	+	-
CI+GALL	–	MC+SA	-	-	-	+	+	-	-
CI+GALL	–	MCI	-	-	+	+	+	+	+
CI+GALL	–	MCI+SA	-	+	+	+	+	+	+
MC	–	MC+SA	-	-	-	-	+	-	+
MC	–	MCI	-	-	+	+	+	+	+
MC	–	MCI+SA	-	+	+	+	+	+	+
MC+SA	–	MCI	-	-	+	+	+	+	+
MC+SA	–	MCI+SA	-	-	+	+	+	+	+
MCI	–	MCI+SA	-	-	-	-	-	-	+

References

Barker, J., 2001. RESPITE CASA Toolkit, User's quide.

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech and Language this issue.

Brown, G.J., Palomäki, K.J., 2008. A reverberation-robust automatic speech

recognition system based on temporal masking. J. Acoust. Soc. Am. 123, 2978.

Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Science 1, 300–307.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562–1573.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120, 2421–2424.

Cooke, M., Green, P., Crawford, M., 1994. Handling missing data in speech recognition, in: Proc. ICSLP, Yokohama, Japan. 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, 267–285.

Faller, C., Merimaa, J., 2004. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. The Journal of the Acoustical Society of America 116, 3075–3089.

Gales, M.J., 1999. Semi-tied covariance matrices for hidden markov models. IEEE Transactions on Speech and Audio Processing 7, 272–281.

Gemmeke, J.F., Cranen, B., ten Bosch, L., 2008. On the relation between statistical properties of spectrographic masks and recognition accuracy,

in: Proceedings of the Fifth IASTED International Conference on Signal Processing, Pattern Recognition and Applications, ACTA Press, Anaheim, CA, USA. pp. 200–206.

Gemmeke, J.F., Cranen, B., Remes, U., 2011. Sparse imputation for large vocabulary noise robust ASR. Computer Speech & Language 25, 462–479.

Gemmeke, J.F., Van hamme, H., Cranen, B., Boves, L., 2010. Compressive sensing for missing data imputation in noise robust speech recognition. IEEE J. STSP 4, 272–287.

Gemmeke, J.F., Wang, Y., Van Segbroeck, M., Cranen, B., Van hamme, H., 2009. Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases, in: Proc. INTERSPEECH, Brighton, UK. pp. 1227–1230.

Harding, S., Barker, J., Brown, G.J., 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. IEEE Transactions on Audio, Speech and Language Processing 14, 58–67.

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 & Language 20, 515–541.

Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Kiessling, A., 2002. SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation, in: Proc. LREC, pp. 329–333.

Kallasjoki, H., Keronen, S., Brown, G.J., Gemmeke, J.F., Remes, U.,

Palomäki, K.J., 2011. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments, in: CHiME 2011 Workshop on Machine Listening in Multisource Environment, Florence, Italy. pp. 58–63.

Kämäräinen, J., Paalanen, P., 2005. GMMBAYES - Bayesian classifier and Gaussian mixture model toolbox V1.0.

Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25, 117–132.

Knapp, C.H., Carter, G.C., 1976. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing 24, 320–327.

Palomäki, K.J., Brown, G.J., Barker, J., 2004. Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. Speech Communication 42, 123–142.

Perez-Lorenzo, J.M., Viciana-Abad, R., Reche-Lopez, P., Rivas, F., Escolano, J., 2012. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. Applied Acoustics 73, 698–712.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Communication 43, 275–296.

Raj, B., Stern, R.M., 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine 22, 101–116.

Raj, B., Virtanen, T., Chaudhuri, S., Singh, R., 2010. Non-negative matrix factorization based compensation of music for automatic speech recognition, in: Proc. INTERSPEECH, Makuhari, Chiba, Japan. pp. 717–720.

Remes, U., Nankaku, Y., Tokuda, K., 2011. GMM-based missing feature reconstruction on multi-frame windows, in: Proc. INTERSPEECH, Florence, Italy. pp. 2407–2410.

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., 1995. WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition, in: Proc. ICASSP, Detroit, MI, USA. pp. 81–84.

Roman, N., Wang, D.L., Brown, G.J., 2003. Speech segregation based on sound localization. The Journal of the Acoustical Society of America 114, 2236–2252.

Seltzer, M., Raj, B., Stern, R., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Communication 43, 379–393.

Tikander, M., Härmä, A., Karjalainen, M., 2003. Binaural positioning system for wearable augmented reality audio, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA. pp. 153–156.

Van hamme, H., 2004. Robust speech recognition using cepstral domain missing data techniques and noisy masks, in: Proc. ICASSP, Montreal, Quebec, Canada. pp. 213–216.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study, in: Proc. Eurospeech, pp. 2407–2410.

Watkins, A., Makin, S., 2007. Perceptual compensation for reverberation in speech identification: effect of single-band, multiple-band, and wideband noise contexts. Acta Acustica united with Acustica 93, 403–410.

Willis, A., 2010. QPC - Quadratic Programming in C.

Zurek, P.M., 1987. The Precedence Effect Directional Hearing. Springer-Verlag, New York. pp. 85–105.