Author(s): Heikki Kallasjoki, Sami Keronen, Guy J. Brown, Jort F. Gemmeke, Ulpu Remes and Kalle J. Palomaki

Title: Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments

Year: 2011

Version: Final published version

Please cite the original version:

Heikki Kallasjoki, Sami Keronen, Guy J. Brown, Jort F. Gemmeke, Ulpu Remes and Kalle J. Palomaki. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME 2011), pages 58-63, Florence, Italy, September 2011.

This publication is included in the electronic version of the article dissertation:
Kallasjoki, Heikki. Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech. Aalto University publication series DOCTORAL DISSERTATIONS, 31/2016.

Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments

Heikki Kallasjoki[1], *Sami Keronen*[1], *Guy J. Brown*[2], *Jort F. Gemmeke*[3], *Ulpu Remes*[1], *Kalle J. Palomäki*[1]

[1]Department of Information and Computer Science,
Aalto University School of Science, Finland
firstname.lastname@tkk.fi
[2]Department of Computer Science, University of Sheffield, UK
g.brown@dcs.shef.ac.uk
[3]Department ESAT, Katholieke Universiteit Leuven, Belgium
jgemmeke@amadana.nl

Abstract

This work presents an automatic speech recognition system which uses a missing data approach to compensate for environmental noise. The missing, noise-corrupted components are identified using binaural features or a support vector machine (SVM) classifier. To perform speech recognition using the partially observed data, the missing components are substituted with clean speech estimates calculated using sparse imputation. Evaluated on the CHiME reverberant multisource environment corpus, the missing data approach significantly improved the keyword recognition accuracy in moderate and poor SNR conditions. The best results were achieved when the missing components were identified using the binaural features and the clean speech estimates associated with observation uncertainty estimates.

Index Terms: noise robust, speech recognition, binaural, SVM, sparse imputation, observation uncertainties

1. Introduction

Automatic speech recognition (ASR) can reach the performance level of human listeners in controlled and noise-free conditions, but environmental noise typically degrades the system performance dramatically unless noise compensation is used. Missing data techniques (MDT) draw motivation from the human auditory processing to improve the ASR performance in noisy environments [2]. The methods are based on finding reliable information in the noise-corrupted speech signal and regarding the unreliable information as missing. Thus, the observed data is first partitioned into reliable, speech-dominated components and unreliable, noise-dominated components, and speech recognition is then attempted based on the reliable information alone. The unreliable components are either completely discarded using marginalization [2] or substituted with clean speech estimates [3].

Several approaches have been proposed for separating between reliable and unreliable information in the observed data. For example, a number of cues can be extracted from auditory models to assist with mask estimation. This includes cues related to fundamental frequency, common onset and offset, and amplitude modulation (see [4] for a review). In addition, if stereo signals (such as those recorded from a dummy head) are available, cues 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. Binaural mechanisms also suppress echoes and, therefore, counteract the negative effects of reverberation [5], and contribute to the ability to focus on a relevant sound source in the presence of other interfering sources. In mask estimation for missing-data methods, binaural cues can be used to exploit, for example, knowledge of the position of the speaker.

An alternative approach to mask estimation is to formulate the problem as a binary classification task [6]. In this case, a number of machine learning methods may be trained to associate features computed from the noisy speech observations with reliability scores obtained from suitable training material. In this work, we compare an auditory approach based on using binaural cues as proposed in [7] and a machine learning approach based on using a support vector machine (SVM) classifier as proposed in [8].

After the observed features are divided into reliable and unreliable (missing) components, the speech recognition system needs to be able to perform recognition using partial data. In the so-called reconstruction or imputation approach, the regions labeled as unreliable by the estimated mask are replaced by clean speech estimates of the missing features. Recently, the exemplar-based missing-feature reconstruction method referred to as sparse imputation [9] has shown good performance in a variety of noise robust speech recognition tasks. Sparse imputation processes the noisy data in windows that span several time frames, which leads to significant performance gains over frame-based methods especially in low-SNR conditions where reliable information is scarce. Experiments on real-world and artificially constructed noisy recordings show that sparse imputation substantially outperforms conventional imputation techniques especially when exact information about the reliable features is provided [8, 10].

In this work, a missing data ASR system using binaural cues or SVM classifier for identifying reliable data and sparse imputation for missing-feature reconstruction is evaluated on the CHiME challenge corpus [11]. To analyze the effect of mask estimation in the system performance, results using a exact information about the reliable components are also reported when possible. In order to mitigate the effect of possible reconstruction errors in the speech recognition performance, the sparse imputation results are augmented with observation uncertainty measures as proposed in [12].

2. Methods

2.1. Missing data techniques

Missing data techniques [2] use the so-called mask estimation methods to divide observed log-mel spectral features $\mathbf{Y}$ into speech and noise dominated regions. The speech-dominated time-frequency components $Y(t, f)$ are considered reliable estimates of the clean speech information, $Y_r(t, f) \approx S(t, f)$, where $S(t, f)$ 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 the noise correlation originates from an uncorrelated source, the unreliable observations provide only an upper bound to the corresponding clean speech values, $Y_u(t, f) \geq S(t, f)$. Thus, the clean speech information in the unreliable components is effectively missing. Missing-data reconstruction techniques such as sparse imputation [9] replace the unreliable values with clean speech estimates $\hat{S}_u(t, f)$. The reconstructed clean speech spectrograms $\hat{\mathbf{S}}$ may be further processed as usual or combined with observation uncertainty estimates as proposed [13].

2.2. Mask estimation methods

2.2.1. Mask estimation based on binaural features

Time-frequency masks are generated from binaural features using an approach based on [7]. First, the left-ear and right-ear signals (from the CHiME dummy-head recordings) are passed through a gammatone filterbank consisting of 21 channels between center frequencies of 171 Hz and 7097 Hz. The center frequencies and bandwidths of the gammatone filters are chosen to give mel-spaced filters that overlapped at approximately the same 3 dB points as the filters used to generate the MFCC features used for recognition (see Section 3.2). The output of each gammatone filter is half-wave rectified, and then ILD and ITD features are computed by the following parallel pathways.

In the ILD pathway, the short-term energy is computed for each channel of the gammatone filter output, for the left and right ears, over a 16 ms rectangular window with an 8 ms hop size. The ILD at each time frame is then calculated by taking the ratio of the left-ear and right-ear energies, and converting to decibels. In the ITD pathway, the cross-correlation is computed between the left-ear and right-ear gammatone filterbank outputs, for each frequency channel, with an 8 ms hop size. Time lags are computed between -1 ms and +1 ms in steps of the sampling period. The time lag at which the largest peak occurred in the cross-correlation function is taken to be the ITD. Hence, two sets of time- and frequency-dependent binaural features are obtained, corresponding to $ILD(t, f)$ and $ITD(t, f)$, where t and f index the time frame and frequency channel respectively.

Following [7], masks are estimated using the binaural features as follows (see also [14] for a related approach). First, there is a training stage in which the joint distribution of ILD and ITD features is estimated for the target source, which is known to be at zero degrees azimuth. The ILD and ITD features are computed for 120 utterances selected from the clean CHiME development set, as described above. Joint ILD-ITD distributions for the target talker are then obtained by constructing histograms of the ILD and ITD values. For ILD, the width of each histogram bin was 0.2 dB; for ITD, bins are spaced at intervals of the sampling period. This is done separately for each frequency channel, giving 21 joint ILD-ITD histograms H_f. Histograms are shown for two frequency channels in Figure 1. As expected the histograms peak at an ILD of 0 dB and an ITD of 0 ms (because the time of arrival and sound level are approximately equal at the two ears of the dummy head, for the target speaker). However, there are frequency-dependent differences due to the effects of room reverberation, and the dependency of ILD and ITD upon frequency (see [15] for a review).

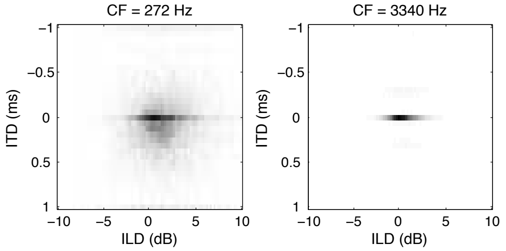

Figure 1: *Joint ILD-ITD histograms for the target speaker (at zero degrees azimuth) obtained from the clean development set. Histograms derived from two channels of the gammatone filterbank are shown, with center frequencies (CFs) of 272 Hz (left) and 3340 Hz (right).*

During recognition, ILD and ITD features are computed for each test utterance as described above, and the mask values $m(t, f)$ are set according to

$$m(t, f) = \begin{cases} 1 & \text{if } H_f[ILD(t, f), ITD(t, f)] > \theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where H_f is the joint ILD-ITD histogram for frequency channel f obtained during the training stage. Finally, any single-unit regions are removed from the masks. The threshold θ was tuned on the noisy development set to give optimum performance. An example of an estimated mask is shown in Figure 2.

2.2.2. SVM masks

Mask estimation can also be approached using machine learning methods to classify each feature as either reliable or unreliable. The classifier is trained to separate between reliable and unreliable features using training material that must necessarily consist of oracle masks, and therefore requires the use of artificially corrupted clean speech for training.

A Bayesian classification approach was proposed for mask estimation in [6], whereas in this work, we use support vector machine (SVM) classifiers. SVM is a machine learning algorithm known for its excellent performance on binary classification tasks and its generalization power when trained on relatively small data sets [16]. From the machine learning perspective, frame-based mask estimation is a multi-class classification problem with 2^F classes, where F is the number of mel-frequency bands. Since such high-dimensional multi-class classification is infeasible, we assume that the reliability estimates are independent between frequency bands and train a separate SVM classifier for each of the F mel-frequency bands. This SVM mask was shown to be effective in [8].

In this work, each classifier used the same set of single-frame-based $(7 \cdot F + 1)$-dimensional features consisting of: 1) the F-dimensional noisy speech features themselves, 2) the harmonic and 3) the aperiodic part of the harmonic decomposition

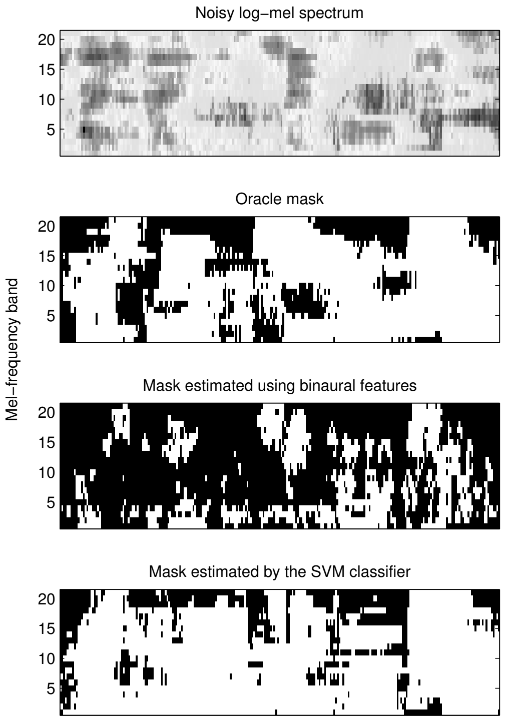

Figure 2: *From top: a noisy development signal with 0 dB SNR, the respective oracle mask with a -2 dB SNR threshold and the estimated masks derived using binaural features or the SVM classifier, respectively. Black regions in masks denote unreliable features.*

described in [17], 4) long-term energy noise estimate, and 5) a gain factor. Additionally, we used 6) the 'sub-band energy to sub-band noise floor ratio' and 7) 'flatness' features derived from the noisy mel-spectral features as described in [6], and finally 8) a single-dimensional VAD feature. The VAD was inspired by the integrated bi-spectrum method described in [18].

Figure 2 contains an example of a mask estimated by the SVM classifier.

2.3. Sparse imputation

2.3.1. Sparse representation

The sparse imputation algorithm [9, 10] is based on processing the log-mel spectral features $\mathbf{S}$ of the observed speech as a sequence of overlapping $F \times T$ dimensional spectrograms, where T is the window length in frames, and F the number of spectral channels. In the following, each window is treated as a single $D = T \cdot F$-dimensional vector by concatenating the consecutive frames. The vector $\mathbf{s}(\tau)^T = [\mathbf{S}(\tau - T/2)^T \cdots \mathbf{S}(\tau + T/2)^T]$ is then represented as a linear combination of example windows i.e. exemplars $\mathbf{a}_n$,

$$\mathbf{s}(\tau) \approx \sum_{n=1}^{N} x_n(\tau)\mathbf{a}_n = \mathbf{A}\mathbf{x}(\tau), \qquad (2)$$

where $\mathbf{x}(\tau)$ is the *activation vector* corresponding to the τ-th window, and $\mathbf{A}$ a $D \times N$ sized fixed dictionary of clean speech windows. The activation vectors $\mathbf{x}(\tau)$ are obtained by solving

$$\mathbf{x}^*(\tau) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{s}(\tau)\|_2 + \lambda \|\mathbf{x}\|_1 \right\}, \qquad (3)$$

where the $\lambda \|\mathbf{x}\|_1$ term is a sparsity-inducing penalty, in order to represent the window $\mathbf{s}(\tau)$ using as few exemplars as possible.

2.3.2. Feature reconstruction

In missing-feature reconstruction, the sparse representation $\mathbf{x}(\tau)$ is calculated based on only the reliable components in the τ-th window $\mathbf{y}(\tau)$ in the observed noisy speech spectrogram $\mathbf{Y}$. The corresponding missing data mask is denoted as $\mathbf{m}(\tau)$ and the activations are calculated as

$$\mathbf{x}^*(\tau) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \left\{ \|\mathbf{W}\mathbf{A}\mathbf{x} - \mathbf{W}\mathbf{y}(\tau)\|_2 + \lambda \|\mathbf{x}\|_1 \right\}, \qquad (4)$$

where the matrix $\mathbf{W} = \operatorname{diag}(\mathbf{m}(\tau))$ is used to select only the reliable components. The sparse representation can then be used to reconstruct an estimate of the clean speech features in the window as $\mathbf{s}^*(\tau) = \mathbf{A}\mathbf{x}^*(\tau)$. The imputed feature $\hat{s}(\tau, d)$ is then set to the observed feature $y(\tau, d)$ if the component is considered reliable, and $\min\{s^*(\tau, d),\ y(\tau, d)\}$ if the component is unreliable. The component-wise min operation reflects the additive noise assumption, under which the clean speech feature cannot exceed the observed noisy speech value. Finally, the clean speech estimates for the overlapping windows are averaged to give the enhanced features $\hat{\mathbf{S}}(t)$ as described in [10].

2.3.3. Observation uncertainties

In missing-feature reconstruction, the acoustic model likelihoods are calculated based on the reconstructed features $\hat{\mathbf{S}}(t)$. Since the estimates vary in accuracy, using observation uncertainties [19] to emphasize the reliable reconstruction can improve speech recognition performance as reported in [13, 12]. Observation uncertainties represent the expected squared error (i.e., variance in each estimated feature) and allow decoding based on the complete estimated posterior $p(\mathbf{S}(t)|\,\Lambda, \mathbf{Y}(t))$, where Λ denotes the parameters used in reconstruction. If the posterior distribution is assumed Gaussian when mapped in the acoustic model domain, decoding with observation uncertainties reduces to adding the estimated uncertainties to the model variances in the acoustic model states [19, 12].

Uncertainty estimates usually reflect the variance of the feature enhancement or reconstruction process, but since sparse imputation does not employ statistical modeling, using heuristic measures to characterize the expected uncertainty was proposed in [12]. According to the experiments reported in [12], the best results are achieved when the uncertainty $\hat{\boldsymbol{\sigma}}(t)$ associated with the reconstructed values in frame t is either set proportional to the number of clean speech exemplars (M4) or inversely proportional to the number of reliable components (M5) used in calculating $\hat{\mathbf{S}}(t)$. The uncertainties are scaled to the interval $[0, 1]$ in each utterance and the reliable components associated with zero uncertainty. For details on the uncertainty measures, see [12].

The uncertainty estimates $\hat{\boldsymbol{\sigma}}(t)$ calculated in the log-mel spectral domain are mapped to the acoustic model domain using a supervised learning approach as proposed in [13]. In this work, a Gaussian mixture model (GMM) is used for the mapping. The model is trained on stereo data (clean and noisy) to represent the statistical dependencies between the log-mel spectral domain uncertainty estimates and the observed uncertainties in the acoustic model domain (OA). These so-called oracle uncertainties are calculated as the squared error between the clean speech and the reconstructed features in the acoustic model domain and log-compressed to better fit the model. In the reconstruction phase, the acoustic model domain uncertainties used in decoding are calculated as the minimum mean squared error (MMSE) estimate of the oracle uncertainty given the model and the log-mel spectral domain uncertainty $\hat{\boldsymbol{\sigma}}(t)$.

3. Experiments

3.1. Experimental setup

The proposed system was evaluated using the CHiME challenge corpus described in [11]. The standard CHiME training set, consisting of 17000 utterances of clean speech reverberated with a binaural room impulse response (BRIR) filter, was used to train the speech recognition system. The CHiME development and evaluation sets, both consisting of 6 sets of 600 utterances at signal-to-noise ratios (SNR) of -6 dB, -3 dB, 0 dB, 3 dB, 6 dB and 9 dB, were used to evaluate the recognition performance of the systems. Only the isolated utterances from the CHiME corpus were used: in particular, the surrounding noise context was not used by the systems. In addition, a noisy training set of 2000 utterances was built following the CHiME test data construction method described in [11]. The underlying clean speech utterances were taken as a random subset of the clean speech training set. The resulting noisy training set contained utterances with SNR values uniformly distributed in the -6 dB to 9 dB range.

For the binaural feature mask described in Section 2.2.1, the masking threshold value θ of Equation (1) was set to $\theta = 0.075$, based on small-scale recognition tests on the development set.

In the case of the SVM-based masks of Section 2.2.2, an individual SVM classifier was trained for each of the 34 speakers using LIBSVM [20] on 5400 frames randomly extracted from utterances of that particular speaker in the noisy training set. Reliability labels used in training were obtained from the oracle mask, derived by using the corresponding clean and noisy utterances of the noisy training set, using a SNR threshold of -2 dB. We used an RBF-kernel and hyper-parameters of the classifier were optimized by doing 5-fold cross validation on a held-apart set of 600 additional frames.

The sparse imputation algorithm described in Section 2.3.1 was applied on 21-dimensional log-mel spectral features. The imputation was performed with the Matlab implementation used in [10]. A random selection of 34000 exemplars from the CHiME training set was used to construct the basis vector dictionary for sparse imputation. Based on development set tests, a window size of $T = 15$ frames was chosen. The 5-component GMM used for mapping the the observation uncertainty measures from the log-mel spectral domain used for the imputation to the acoustic model domain of the speech recognition system was trained using a 500 utterance subset of the constructed noisy training set. The model parameters were estimated using the expectation-maximization (EM) algorithm implemented in the GMMBAYES Matlab toolbox.

3.2. Speech recognition system

The acoustic models of the ASR system utilized 39-dimensional features composed of 12 MFCC coefficients, the logarithmic frame energy, and their first and second differentials. The features were post-processed by applying cepstral mean subtraction (CMS) and maximum likelihood linear transformation (MLLT) steps. State-clustered hidden Markov models using cross-word triphone units were used to model the features. A LVCSR system trained on the Wall Street Journal British English (WSJCAM0) corpus was used in "forced alignment" mode to generate triphone-level segmentations for the CHiME training data. Gaussian mixture models were used to model the acoustic feature domain, and Gamma distributions for explicit state duration modeling. The recognition system is described in more detail in [21]. For language modeling, a no-backoff bigram model with uniform frequencies for all valid bigrams was constructed to restrict recognized sentences to conform to the Grid utterance grammar specified in [22].

A single speaker-independent clean speech model was trained using the full 17000 utterance CHiME training set. This baseline system achieved recognition results comparable to the standard CHiME baseline system.

4. Results

All the compared systems used the speaker-independent model trained with the clean CHiME training data set. In the baseline system, this model was used as-is to recognize the noisy utterances. On the clean speech version of the development set, a keyword accuracy of 95.9% was achieved.

Table 1 presents achieved keyword accuracy rates for the noisy utterances of the CHiME development set data using the different mask estimation methods. In addition to the official CHiME baseline recognizer ("CHiME bl.") and our baseline system ("baseline"), results using sparse imputation feature enhancement are given for three different mask types: oracle masks that utilize knowledge of the clean samples ("oracle"), the binaural masks described in Section 2.2.1 ("bin.") and the SVM based masks described in Section 2.2.2 ("SVM"). Of the two mask estimation methods, only the binaural mask was able to surpass the recognition performance of the baseline recognizer at all, in the cases where the SNR was between 0 dB and -6 dB. Relative improvements in these cases range between 4% and 17%. The binaural mask was therefore selected for the observation uncertainty experiments.

Results using the three different uncertainty measures, along with the baseline and imputation-only performance, are given in Table 2. The considered measures are acoustic model domain oracle uncertainties (OA) as well as the two heuristic measures described in Section 2.3.3 (M4, M5). Here both M4 and M5 measures gave similar improvements over the SI-only system for all SNR levels. As the M4 measure achieved slightly better results, it was chosen for the final evaluation set experiments.

Table 1: Keyword accuracy rates for CHiME development set data with sparse imputation only.

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
CHiME bl.	83.1	73.8	64.0	49.1	36.8	31.1
baseline	83.3	77.9	67.3	52.9	42.2	36.8
oracle	92.7	93.1	90.3	90.6	89.3	88.4
bin.	75.8	71.9	64.8	55.1	49.5	43.1
SVM	76.5	69.4	54.6	44.4	37.6	34.6

Table 2: Keyword accuracy rates for CHiME development set data with observation uncertainty measures.

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
baseline	83.3	77.9	67.3	52.9	42.2	36.8
SI only	75.8	71.9	64.8	55.1	49.5	43.1
OA	89.3	87.4	83.1	80.2	72.6	69.7
M4	78.9	74.7	67.8	59.3	54.0	46.8
M5	78.8	74.7	67.8	58.3	53.7	46.3

Based on the results on the development set, the SI system with binaural feature masks and uncertainty measure M4 was chosen as the primary system. Results on the final CHiME evaluation set are presented in Table 3, both with and without using the observation uncertainty measure M4. As corresponding clean speech for the evaluation set was not available, the oracle mask and oracle uncertainty results are not shown. The evaluation set results closely match the corresponding development set experiments. Overall relative improvements in keyword accuracy over our baseline system for the 0 dB to -6 dB SNR test sets were between 7–13% and 14–21% for the imputation-only and observation uncertainty utilizing systems, respectively.

Table 3: Keyword accuracy rates for CHiME evaluation set data. The highest keyword accuracy for each data set is indicated in bold type.

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB
CHiME bl.	82.4	75.0	62.9	49.5	35.4	30.3
baseline	**85.6**	**77.9**	66.3	51.2	40.0	38.7
SI only	74.3	70.1	64.3	54.7	45.3	42.8
M4	77.3	73.5	**67.3**	**58.5**	**47.8**	**46.8**

5. Discussion

Significant differences can be seen when comparing the obtained keyword accuracy rates between systems using oracle masks and the binaural feature based masks. The very high recognition performance obtainable with sparse imputation when oracle masks are used suggests that mask estimation is pivotal to the performance of the system. Imputation with the SVM classifier based masks, shown to perform well with sparse imputation in realistic noise environments [8], failed to outperform the baseline system in this work. This indicates that the CHiME data set is a challenging case from the mask estimation point of view.

In the CHiME corpus, the interfering noise is often highly variable and is in many cases a voice of a single interfering talker. Models tailored specifically for separation of overlapping speech, e.g. following [23], could perform better in these cases. The SVM mask estimation method used in the present study was not specifically developed for interfering speech. In addition, in this work the SVM based masks used only monaural features based on harmonic decomposition. The use of the harmonic decomposition as a feature implicitly assumes that speech characteristics can be identified by the harmonic part of the spectrum, an assumption which may not be suited very well to the noise types encountered in the CHiME challenge.

Although using sparse imputation with the binaural mask described in Section 2.2.1 improved the recognition rates in low-SNR conditions, it should be noted that the binaural mask estimation used here is a very crude approximation of human processing. Two factors that contribute to human performance in multisource reverberant environments are not considered here, namely the precedence effect [5] and better-ear listening. Regarding the latter, Edmonds and Culling [24] have shown that better-ear listening plays a substantial role in the perceptual separation of speech from an interfering voice. Better-ear listening could be incorporated into the system described here by using the binaural model to identify the stereo channel with the most favorable SNR, and using only this channel for speech recognition.

Using sparse imputation on the development data with exact knowledge of the missing components (the “oracle” masks) resulted in impressive performance gains at all noise levels, suggesting that the reconstruction method as such is well-suited for noise conditions that are present in the CHiME data. However, when used with the estimated masks, sparse imputation degraded the recognition results compared to the uncompensated baseline system in the relatively low-noise conditions (SNR 3–9 dB). In [10], the sparse imputation method used in this work was found to perform less well than an alternative, cluster-based imputation system, when estimated masks were used in the cleanest conditions. When the estimated mask has a low number of isolated features erroneously marked as reliable, almost the entire T-frame window will be reconstructed based on the clean speech exemplars, potentially leading to insertion errors in the speech recognition stage. In the noisier conditions, however, few isolated reliable features can be all of the true, underlying clean speech signal that is left uncorrupted by noise. In this case the ability of the SI reconstruction to reconstruct longer clean speech segments can in fact improve the recognition performance.

Use of observation uncertainties can mitigate the degradation on less noisy data. For example in this work, SI when combined with the M4 uncertainty measure achieves the baseline performance also in the 3 dB case. The use of observation uncertainty estimates improves the recognition rate of the sparse imputation system for all SNR levels, with relative improvements in the keyword accuracy ranging from 4% to 9%. When “oracle” uncertainty estimates are used with the development set data, the SI system outperforms the baseline recognizer in all noise conditions, with a relative improvement of 7–90%.

6. Acknowledgments

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

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.

7. References

[1] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *ICSLP*, 1994, pp. 1555–1558.

[2] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[3] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275 – 296, 2004, special Issue on the Recognition and Organization of Real-World Sound. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4D1YV53-1/2/1b4ea43751d95a4cf54057dc4c43c21f

[4] D. Wang and G. J. Brown, Eds., *Computational Auditory scene analysis: Principles, Algorithms and Applications*. New York: Wiley/IEEE Press, 2006.

[5] P. M. Zurek, *The Precedence Effect Directional Hearing*. New York: Springer-Verlag, 1987.

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

[7] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[8] J. F. Gemmeke, Y. Wang, M. Van Segbroeck, B. Cranen, and H. Van hamme, "Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases," in *Proc. INTERSPEECH*, Brighton, UK, September 6–10 2009, pp. 1227–1230.

[9] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE J. STSP*, vol. 4, no. 2, pp. 272–287, March 2010.

[10] J. F. Gemmeke, B. Cranen, and U. Remes, "Sparse imputation for large vocabulary noise robust ASR," *Computer Speech & Language*, vol. 25, no. 2, pp. 462–479, 2011.

[11] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010.

[12] J. F. Gemmeke, U. Remes, and K. J. Palomäki, "Obsesrvation uncertainty measures for sparse imputation," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2262–2265.

[13] S. Srinivasan and D. L. Wang, "A supervised learning approach to uncertainty decoding for robust speech recognition," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 297–300.

[14] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[15] R. M. Stern, D. L. Wang, and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Application*. New York: Wiley/IEEE Press, 2006, ch. Binaural sound localization.

[16] C. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, pp. 273 – 297, 1995.

[17] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. International Conference on Audio, Speech and Signal Processing*, vol. 1, 2004, pp. 213–216.

[18] J. Ramírez, J. Górriz, J. Segura, C. Puntonet, and A. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum lrt," in *IEEE Signal Processing Letters*, 2006.

[19] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. ICSLP*, Denver, Colorado, USA, 2002, pp. 1561–1564.

[20] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9020

[21] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and S. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, 2006.

[22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[23] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. INTERSPEECH*, 2006.

[24] B. A. Edmonds and J. F. Culling, "The spatial unmasking of speech: Evidence for better-ear listening," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1539–1545, 2006.