Author(s): Heikki Kallasjoki, Ulpu Remes, Jort F. Gemmeke, Tuomas Virtanen and Kalle Palomaki

Title: Uncertainty measures for improving exemplarbased source separation

Year: 2011

Version: Final published version

Please cite the original version:

Heikki Kallasjoki, Ulpu Remes, Jort F. Gemmeke, Tuomas Virtanen and Kalle Palomaki. Uncertainty measures for improving exemplarbased source separation. In Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), pages 469-472, Florence, Italy, August 2011

Uncertainty measures for improving exemplar-based source separation

Heikki Kallasjoki[1]*, Ulpu Remes*[1]*, Jort F. Gemmeke*[2]*, Tuomas Virtanen*[3]*, Kalle J. Palomäki*[1]

[1] Adaptive Informatics Research Centre, Aalto University, Finland
[2] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[3] Department of Signal Processing, Tampere University of Technology, Finland

Abstract

This work studies the use of observation uncertainty measures for improving the speech recognition performance of an exemplar-based source separation based front end. To generate the observation uncertainty estimates for the enhanced features, we propose the use of heuristic methods based on the sparse representation of the noisy signal in the exemplar-based source separation algorithm. The effectiveness of the proposed measures is evaluated in a large vocabulary noisy speech recognition task. The best proposed measure achieved relative error reductions up to 18 % over the baseline feature enhancement method without uncertainty measures.

Index Terms: robustness, speech recognition, source separation, observation uncertainties

1. Introduction

Feature enhancement is a common way to improve the performance of automatic speech recognition (ASR) in the presence of noise. In general, the goal is to produce estimates of the underlying clean speech features, discarding the effects of noise on the signal. Feature enhancement can be seen from a source separation point of view, where the task is to extract the original signals, in particular the uncorrupted speech signal, from a mixture of speech and noise.

A feature enhancement method for improving noise robust speech recognition based on source separation was recently proposed in [1]. The feature enhancement method is based on representing the noisy observations using a sparse, non-negative linear combination of a predefined collection of speech and noise samples called *exemplars*. Non-negative matrix factorization (NMF) methods are used to obtain this sparse representation [2].

The reconstruction of the clean speech signal, however, is never exactly identical to the original. Taking into account the reliability, or conversely, the uncertainty of the reconstructed features using either observation uncertainty techniques [3, 4] or uncertainty decoding [5] generally improves ASR performance in noisy conditions. In order to utilize these methods, uncertainty estimates need to be produced for the observed features.

In this work, we investigate the use of observation uncertainty measures in improving the speech recognition performance of the feature enhancement front-end of [1]. As the source separation algorithm does not directly produce estimates for the variance of the enhanced features, we propose heuristic methods to characterize the uncertainty of the observations. Four measures are constructed based on the structure of the sparse representation in the exemplar-based source separation approach and evaluated in a large vocabulary continuous noisy speech recognition task.

2. Source separation

2.1. Exemplar-based representation of noisy speech

The source separation technique employed in this paper was presented in [2]. It operates on magnitude mel-spectrograms $\mathbf{Y}$ of the noisy speech signal, described as a $B \times T$ dimensional matrix (with B frequency bands and T time frames), which is a linear addition of underlying clean speech $\mathbf{S}$ and noise $\mathbf{N}$ magnitude spectrograms. To simplify the notation, the columns of each matrix are stacked into the vectors $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{n}$, respectively, each of length $D = B \cdot T$.

We model $\mathbf{s}$ as a sparse, non-negative linear combination of example speech spectrograms *exemplars*, which have been extracted from the training data. The exemplars are denoted as $\mathbf{a}^{\mathrm{s}}_j$, with $j = 1, \ldots, J$ denoting the exemplar index. Accordingly, the noise spectrogram is modelled using K noise exemplars $\mathbf{a}^{\mathrm{n}}_k$, with $k = 1, \ldots, K$. This can be expressed as

$$\mathbf{y} \approx \mathbf{s} + \mathbf{n} \tag{1}$$

$$\approx \sum_{j=1}^{J} x^{\mathrm{s}}_j \mathbf{a}^{\mathrm{s}}_j + \sum_{k=1}^{K} x^{\mathrm{n}}_k \mathbf{a}^{\mathrm{n}}_k \tag{2}$$

$$= [\mathbf{A}^{\mathrm{s}} \mathbf{A}^{\mathrm{n}}] \begin{bmatrix} \mathbf{x}^{\mathrm{s}} \\ \mathbf{x}^{\mathrm{n}} \end{bmatrix} \tag{3}$$

$$= \mathbf{A}\mathbf{x} \quad \text{s.t.} \quad \mathbf{x}^{\mathrm{s}}, \mathbf{x}^{\mathrm{n}}, \mathbf{x} \geq 0 \tag{4}$$

with $\mathbf{x}^{\mathrm{s}}$ and $\mathbf{x}^{\mathrm{n}}$ sparse representations of the underlying speech and noise, respectively. In order to obtain $\mathbf{x}$, we minimize the cost function:

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + ||\boldsymbol{\lambda} .* \mathbf{x}||_1 \quad \text{s.t.,} \quad \mathbf{x} \geq 0 \tag{5}$$

where d is the generalized Kullback-Leibler (KL) divergence and the second term a sparsity inducing L-1 norm of the activations weighted by element-wise multiplication (operator $.*$) with vector $\boldsymbol{\lambda} = [\lambda_1 \;\; \lambda_2 \ldots \lambda_L]$. The cost function (5) is minimized using a multiplicative updates routine as in [1].

2.2. Feature enhancement

Let us denote the noisy speech spectrum in frame t as $\mathbf{y}_t$. Similarly, let us denote the spectra of speech exemplar j and noise exemplar k in frame t as $\mathbf{a}^{\mathrm{s}}_{j,t}$ and $\mathbf{a}^{\mathrm{n}}_{k,t}$, respectively. Models for the clean speech and noise are then given as

$$\tilde{\mathbf{s}}_t = \sum_{j=1}^{J} x^{\mathrm{s}}_j \mathbf{a}^{\mathrm{s}}_{j,t}, \tag{6}$$

$$\tilde{\mathbf{n}}_t = \sum_{k=1}^{K} x^{\mathrm{n}}_k \mathbf{a}^{\mathrm{n}}_{k,t}. \tag{7}$$

In order to decode utterances of arbitrary lengths, we adopt a sliding time window approach as in [1]. In this approach, we represent a noisy utterance as a number of fixed-size, overlapping speech segments, each of length T. For each segment, we calculate clean speech estimates $\tilde{\mathbf{s}}_t$ and noise estimates $\tilde{\mathbf{n}}_t$ as described above. To get a single clean speech and noise estimate for each frame of the original utterance, the segment-wise estimates are then averaged over all the segments overlapping the frame.

Instead of using the above clean speech estimates directly, we further process the the original noisy features $\mathbf{y}_t$ as follows. A spectral domain filter $\mathbf{h}_t$ is designed for each frame t as

$$\mathbf{h}_t = \tilde{\mathbf{s}}_t \,./ (\tilde{\mathbf{s}}_t + \tilde{\mathbf{n}}_t), \tag{8}$$

with $./$ denoting element-wise division. The enhanced features are then obtained as $\hat{\mathbf{s}}_t = \mathbf{h}_t \,.* \mathbf{y}_t$. The above feature enhancement procedure was first proposed and analyzed in more detail in [1]. Finally, the enhanced features are transformed into the acoustic model domain features $\hat{\boldsymbol{\varsigma}}_t$ of the recognition system described in Section 4.2.

3. Observation uncertainties

3.1. Use of observation uncertainties

In any feature enhancement task, the accuracy of estimated features varies depending on e.g. the noise level in the original features. Therefore, observation uncertainties [6] were proposed in [3] to characterize the expected error i.e. variance in the clean speech estimates $\hat{\boldsymbol{\varsigma}}_t$ based on which the acoustic model likelihoods are calculated. With estimated uncertainties, the point estimates $\hat{\boldsymbol{\varsigma}}_t$ can be replaced with the complete estimated posteriors $p(\boldsymbol{\varsigma}_t|\boldsymbol{\theta}, \mathbf{y}_t)$, where $\boldsymbol{\theta}$ denotes the parameters applied in feature enhancement and $\mathbf{y}_t$ the observed noisy speech features. The likelihood of the q-th state of the clean speech acoustic models $\mathcal{M}$ is then calculated as

$$L(q) = \int p(\boldsymbol{\varsigma}|\boldsymbol{\theta}, \mathbf{y}_t)\, p(\boldsymbol{\varsigma}|\mathcal{M}, q)\, d\boldsymbol{\varsigma}. \tag{9}$$

Decoding with observation uncertainties is closely related to so-called *uncertainty decoding* and has been successfully used with several feature enhancement techniques; see [5] for a review and discussion.

Assuming the states q are modelled as Gaussian mixtures, the state likelihoods are calculated as a weighted sum over the likelihoods of each mixture component l. Furthermore, assuming a Gaussian posterior $p(\boldsymbol{\varsigma}_t|\boldsymbol{\theta}, \mathbf{y}_t)$ as proposed in [3], the likelihood of the l-th Gaussian component is calculated as

$$\begin{aligned} L(l) &= \int \mathcal{N}(\boldsymbol{\varsigma}; \hat{\boldsymbol{\varsigma}}_t, \boldsymbol{\Sigma}_{\hat{\varsigma},t})\, \mathcal{N}(\boldsymbol{\varsigma}; \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)})\, d\boldsymbol{\varsigma} \\ &= \mathcal{N}(\hat{\boldsymbol{\varsigma}}_t; \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)} + \boldsymbol{\Sigma}_{\hat{\varsigma},t}), \end{aligned} \tag{10}$$

where $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ are the mean and covariance of the l-th Gaussian in the uncompensated clean speech model and $\hat{\boldsymbol{\varsigma}}_t$ and $\boldsymbol{\Sigma}_{\hat{\varsigma},t}$ are the mean and covariance of the clean speech posterior estimate at frame t. Thus, decoding with observation uncertainties reduces to adding the estimated uncertainties $\boldsymbol{\Sigma}_{\hat{\varsigma},t}$ to the model covariances $\boldsymbol{\Sigma}^{(l)}$. In this work, the covariances are assumed diagonal, with the notation $\boldsymbol{\Sigma}_{\hat{\varsigma}} = \mathrm{diag}(\boldsymbol{\sigma}_{\hat{\varsigma}})$.

3.2. Uncertainty measures from source separation

The exemplar-based source separation approach produces an estimate of the clean speech features but does not directly allow the estimation of feature variances that could be used as observation uncertainties. In this work, we consider several heuristic measures based on the source separation algorithm for estimating the uncertainties of the enhanced features. Here we denote by $\mathbf{y}_t$ the noisy, observed mel-spectral features in frame t, and by $\mathbf{s}_t$ and $\hat{\mathbf{s}}_t$ the underlying clean speech and feature enhancement outputs, respectively. The following uncertainty measures $\hat{\boldsymbol{\sigma}}_t$ are considered:

H1 Under the assumption that the enhanced features that differ the most from the observed noisy mixture are most unreliable, the uncertainty related to $\mathbf{y}_t$ can be set proportional to the relative energy difference between $\mathbf{y}_t$ and $\hat{\mathbf{s}}_t$ as proposed in [7]. Hence we define the uncertainty vector as $\hat{\boldsymbol{\sigma}}_t = \log\left[(\mathbf{y}_t - \hat{\mathbf{s}}_t)\,./\mathbf{y}_t\right]$. Logarithmic compression is used for better fit with the mixture of Gaussians used for mapping the uncertainties; the mapping is discussed in Section 3.3.

H2 If the exemplars that are chosen for the sparse representation of a particular observed feature are mostly selected from the noise dictionary, the signal is likely to have been relatively noisy. The enhanced features can therefore be considered to be more uncertain than in the case when the observation is represented primarily using clean speech exemplars.

Accordingly, we propose to set the observation uncertainty estimates for each frame t proportional to the ratio of the summed weights of the noise and speech exemplar activations: $\hat{\sigma}_t = g\big(\big(\sum_\tau \sum_i x_i^{\tau,n}\big) / \big(\sum_\tau \sum_i x_i^{\tau,s}\big)\big)$, where $x_i^{\tau,n}$ and $x_i^{\tau,s}$ are the noise and clean speech sparse representations of Equation (3) for speech segment τ, and the summation index τ ranges over the speech segments that contain the frame t.

The normalization function g is an affine transformation chosen independently for each utterance to scale the uncertainty estimates to the interval $[0, 1]$, as the range of the proposed measure varies widely between utterances.

H3 If an observation does not match well with the clean speech dictionary $\mathbf{A}^s$, there is no single exemplar that would be sufficiently similar to the underlying clean speech observation and the sparse representation of the clean speech will consist of multiple different exemplars. In this case, also, the uncertainty of the reconstructed features for that frame can be considered relatively large.

We therefore propose setting the uncertainties (inversely) proportional to the number of clean speech exemplars used: $\hat{\sigma}_t = g\big(\sum_\tau \sum_i f(p : x_i^{\tau,s} > 0.01)\big)$, where notation from H2 has been used, and $f(p) = 1$ if the proposition p is true, otherwise zero.

H4 Similar to H3, we can use the combined speech and noise activation vector to consider also how well the noise is being modeled by the noise exemplars: $\hat{\sigma}_t = g\big(\sum_\tau \sum_i f(p : x_i^\tau > 0.01)\big)$, where x_i^τ is the combined sparse representation in Equation (4).

H* A combined uncertainty measure denoted by H* is defined as a simple concatenation of the uncertainty estimates for methods H1–H4.

Finally, if the clean speech features $\mathbf{s}_t$ are known, an oracle uncertainty estimate in the mel-spectral domain can be computed as the squared error between the known clean speech features and the enhanced features: $[\hat{\boldsymbol{\sigma}}_t]_i = \log\left[(\hat{\mathbf{s}}_t - \mathbf{s}_t)\right]_i^2$,

where log-compression is used for a better fit with the mixture of Gaussians used in mapping the uncertainties. The mel-spectral domain oracle uncertainties are denoted as OS in this work. The oracle uncertainties computed in the acoustic model domain as squared difference between $\boldsymbol{\varsigma}_t$ and $\hat{\boldsymbol{\varsigma}}_t$ are denoted as OA.

3.3. Mapping uncertainties between domains

The uncertainty measures proposed in Section 3.2 characterize the uncertainty of either the produced mel-spectral feature components or the entire frame. The decoding process described in Section 3.1, however, requires observation uncertainty estimates in the acoustic model domain. In this work, the uncertainty estimates $\hat{\boldsymbol{\sigma}}$ are transformed to the acoustic model domain observation uncertainties $\boldsymbol{\sigma}_{\hat{\varsigma}}$ using a Gaussian mixture model (GMM). The joint distribution of the estimated and acoustic model domain uncertainties is modelled as

$$p(\boldsymbol{z}) = \sum_k P(k)\,\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}), \tag{11}$$

where $\boldsymbol{z}^\mathsf{T} = [\hat{\boldsymbol{\sigma}}^\mathsf{T}\ \log \boldsymbol{\sigma}_{\hat{\varsigma}}^\mathsf{T}]$ are the concatenated uncertainty vectors, k is a mixture component index, $P(k)$ are the component weights, and $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ the means and covariances. The acoustic model domain uncertainties are logarithmically compressed to improve fit with the model. The component index k is assumed a hidden variable.

Given the uncertainty measure $\hat{\boldsymbol{\sigma}}_t$ calculated from the t-th frame and the joint distribution from Equation (11), the minimum mean square error (MMSE) estimate for the corresponding acoustic model domain uncertainties is calculated as

$$E\{\log \boldsymbol{\sigma}_{\hat{\varsigma}} \mid \hat{\boldsymbol{\sigma}}_t, \Lambda\} = \sum_k P(k \mid \hat{\boldsymbol{\sigma}}_t, \Lambda)\, E\{\log \boldsymbol{\sigma}_{\hat{\varsigma}} \mid \hat{\boldsymbol{\sigma}}_t, \Lambda, k\}, \tag{12}$$

where $\hat{\boldsymbol{\sigma}}_t$ denotes $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_t$ and Λ the model parameters. The posterior probabilities for clusters k are calculated from $P(k)$ and likelihoods $p(\hat{\boldsymbol{\sigma}}_t \mid \Lambda, k)$ which are calculated using diagonal covariances. The cluster-conditional estimates are calculated as

$$E\{\log \boldsymbol{\sigma}_{\hat{\varsigma}} \mid \hat{\boldsymbol{\sigma}}_t, \Lambda, k\} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}(\hat{\boldsymbol{\sigma}}_t - \boldsymbol{\mu}_1), \tag{13}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the means of heuristic and acoustic model domain uncertainties, $\boldsymbol{\Sigma}_{12}$ the cross-covariance, and $\boldsymbol{\Sigma}_{11}$ the covariance of the heuristic uncertainties. The means and covariances are subsets of $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ in Equation (11).

4. Experiments

4.1. Experimental setup

The proposed uncertainty measures were evaluated in a speech recognition task using artificially corrupted clean speech data. The clean speech data was taken from the Finnish SPEECON corpus and the noise samples from the NOISEX-92 database. Separate segments of factory noise recorded near plate-cutting and electrical welding equipment were used as training set noise and matching test set noise. In addition, a factory noise sample recorded in a car production hall was used as mismatched test set noise.

Training data for the acoustic models consisted of approximately 30 hours of clean speech from 293 speakers. The evaluation set contained 564 sentences (57 minutes) of clean speech from 40 speakers. Three variants of the evaluation set were used: two with matching noise at SNR of 10 dB and 5 dB, and one with mismatched noise at SNR of 10 dB.

The exemplar-based source separation method (Section 2.2) was applied on 21-dimensional mel-spectral features calculated in 25 ms windows with a 10 ms frame shift and processed in $T = 15$ frame segments. The speech exemplar dictionary $\mathbf{A}^s$ in Equation (3) had $J = 8000$ samples randomly selected from read sentences in the clean speech training data, and the noise exemplar dictionary $\mathbf{A}^n$ had $K = 4000$ random samples of the training set noise.

The sparsity coefficient vector $\boldsymbol{\lambda}$ in Equation (5) was of the form $[\lambda^s \cdots \lambda^s\ \lambda^n \cdots \lambda^n]$, with λ^s and λ^n corresponding to the clean speech and noise exemplars. Based on speech recognition experiments on separate development datasets, we chose $\lambda^s = 0.7,\ \lambda^n = 0.5$ for matching noise and $\lambda^s = 0.65,\ \lambda^n = 0$ for mismatched noise.

The GMM parameters in Equation (11) were trained using the expectation-maximization algorithm implemented in the GMMBAYES Matlab toolbox[1]. The dataset used for parameter estimation was a subset of 500 utterances (52 minutes) randomly selected from the read sentences of the SPEECON training set, corrupted with the training set noise at SNR of 10 dB. The target acoustic model domain uncertainties were computed as the OA measure in Section 3.2.

Finally, two baseline systems were provided for comparison. One was trained on the clean speech training data and one trained on multicondition training set i.e. the clean speech training data where each utterance is either used as clean speech or corrupted with the training set noise, at SNR of 5 dB, 10 dB or 15 dB, with uniform probability for each alternative. The baseline systems are as described in Section 4.2.

4.2. Speech recognition system

Input speech is represented with 12 MFCC coefficients and logarithmic frame energy, and their first and second differentials. Cepstral mean subtraction (CMS) and a maximum likelihood linear transformation (MLLT) are applied as post-processing steps. The acoustic models are state-clustered hidden Markov models that use cross-word triphones as units. State feature distributions are modeled as Gaussian mixtures, and Gamma distributions are used for state duration modeling; see [8] for details. Language modeling is based on a variable-length, growing n-gram model of statistical morphemes learned with an unsupervised method [8]. 145-million word Finnish book and newspaper corpus was used in training the model. The decoder employes a one-pass time-synchronous Viterbi beam search algorithm. Letter error rate (LER) is used to measure the recognition performance as it is better suited for agglutinative languages such as Finnish than word error rate (WER).

5. Results

Table 1 presents the speech recognition performance of the evaluated methods. The baseline recognizer, which does not use observation uncertainties, was evaluated using the noisy features (BL) and with the speech separation feature enhancement method (SS). The recognizer using observation uncertainties was evaluated with the oracle uncertainties calculated in the acoustic model domain (OA) and in the mel-spectral domain (OS) and with the uncertainty heuristics (H1–H4, H*). Results obtained with a multicondition trained baseline recognizer (MC) are provided for reference.

All the uncertainty measures proposed in this work improve recognition performance over the SS baseline. The overall best

[1] Available in http://www.it.lut.fi/project/gmmbayes/

Table 1: Letter error rates (LER) for the compared systems.

	Matching noise		Mismatched	Clean
	SNR 10	SNR 5	SNR 10	speech
BL	21.2	58.7	57.9	3.4
SS	10.4	22.8	21.8	3.4
OA	9.4	16.3	13.6	3.3
OS	9.3	18.0	18.3	3.3
H1	9.6	18.7	20.4	3.4
H2	10.1	19.6	20.5	3.4
H3	9.9	19.6	20.1	3.3
H4	9.9	19.7	20.5	3.4
H*	9.6	19.0	20.0	3.4
MC	6.5	12.9	17.2	4.4

heuristic, H1, achieves relative error rate reductions of 8–18 % for matching noise and 6 % for mismatched noise. The corresponding relative improvements with OS are 11–21 % and 16 %, respectively. Finally, with OA relative improvements of 9–38 % are obtained.

The letter error rates for the matching noise at SNR 10 dB are in general substantially lower than for mismatched noise at SNR 10 dB, for all methods including the clean speech trained baseline system. This is mostly due to the different frequency characteristics of the noise samples, which causes cross-comparison of the linear SNR values to be misleading. The matching noise sample contains a strong low-frequency component that dominates in a linear SNR measurement. A-weighted SNR values for the nominal 10 dB and 5 dB matching noise sets are 13.3 dB and 8.3 dB, respectively, whereas the SNR 10 dB mismatched noise set has an A-weighted SNR of 9.3 dB.

6. Discussion

All the proposed uncertainty heuristics improved the recognition performance of the exemplar-based source separation feature enhancement method. The observed error reductions are comparable with results reported for observation uncertainties derived from the variance of a feature enhancement process [3] and for similar heuristics evaluated with a sparse imputation front-end [4].

The overall best performing method was the H1 measure proposed in [7]. This is likely because the relative difference of the enhanced and original features used in H1 has a direct correspondence with the amount of attenuation in the spectral domain filter $\boldsymbol{h}_t$ employed by the feature enhancement method. In contrast, when heuristics similar H3 and H4 which are based on the number of active exemplars were evaluated in the context of a sparse imputation, they were found to measure well the reliability of the clean speech estimate $\tilde{\boldsymbol{s}}_t$ [4]. In sparse imputation, the clean speech estimate is directly used to compute the enhanced features $\hat{\boldsymbol{\varsigma}}$.

The difference in recognition results for acoustic model domain uncertainties (OA) and the transformed mel-spectral domain uncertainties (OS, H1–H4) is significantly larger for the mismatched noise case. As the matching noise type was also used in training the mapping between uncertainty domains described in Section 3.3, the difference likely reflects imperfections in the uncertainty mapping. Various approaches could be investigated for improving the mapping of the mel-spectral observation uncertainties into the acoustic model domain. In particular, the uncertainties for the first and second order differential features could be more accurately approximated by utilizing the time context in the mel-spectral domain.

The combined heuristic measure H* did not achieve prominent improvement over the best-performing individual heuristics. This suggests that the proposed heuristic estimates make similar errors. In order to improve generalization to the mismatched noise case, a multi-view learning algorithm instead of simple concatenation could be used to produce acoustic model domain uncertainty estimates based on multiple heuristics.

In most conditions, the recognition performance of the feature enhancement method did not reach the multicondition trained model performance reported for reference. However, for the clean speech test set the letter error rate of the multicondition trained model was notably higher than the clean speech baseline. In contrast, this is not the case for the feature enhancement method and the use of observation uncertainties, where the results were comparable to the baseline model. Moreover, in the mismatched noise case, the recognition performance achieved by the speech separation feature enhancement method with acoustic domain oracle uncertainties was higher than that of the multicondition trained reference model, and notably higher than with the uncertainty heuristics. This suggests that further recognition performance improvements are possible by improving the uncertainty estimates. In future work, activation variances produced by a probabilistic NMF model could be used as a source for uncertainty estimates.

7. Acknowledgements

The work was supported by the Hecse graduate school (HK, UR), and by the Academy of Finland projects 136209 (KJP) and AIRC (HK, UR, KJP). The research of JFG was supported by the Dutch-Flemish STEVIN project MIDAS and by IWT project ALADIN.

8. References

[1] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, accepted for publication.

[2] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. of Interspeech*, 2010, pp. 717–720.

[3] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[4] J. F. Gemmeke, U. Remes, and K. J. Palomäki, "Observation uncertainty measures for sparse imputation," in *Proc. of Interspeech*, 2010, pp. 2262–2265.

[5] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Communication*, vol. 50, no. 4, pp. 265–277, 2008.

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

[7] J. A. Arrowood, "Using observation uncertainty for robust speech recognition," Ph.D. dissertation, Georgia Institute of Technology, 2003.

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