Bounded Conditional Mean Imputation with an Approximate Posterior

Ulpu Remes

Department of Information and Computer Science, Aalto University School of Science, Finland

ulpu.remes@aalto.fi

Abstract

Missing-feature imputation or reconstruction is used in noise-robust automatic speech recognition to recover the unobserved clean speech information. Reconstruction methods often use the noise-corrupted observations and a clean speech prior to calculate a point estimate for the unobserved clean speech features, whereas the approach proposed in this work associates the unobserved clean speech features with a full posterior distribution. The posterior mean can be used as a clean speech estimate in bounded conditional mean imputation and the posterior variance can be included as observation uncertainties. The proposed method is evaluated in a large-vocabulary noise-robust speech recognition task with speech data recorded in real noisy environments.

Index Terms: missing data, observation uncertainty, speech recognition

1. Introduction

Using missing-feature methods in noise-robust speech recognition is based on the assumption that additive noise can be modelled as a masker rather than distortion to the clean speech features [1, 2, 3]. The masked i.e. missing features can be marginalised or substituted with clean speech estimates calculated based on a clean speech prior. The missing-feature reconstruction or imputation methods that substitute masked features with feature estimates include several that model clean speech features as samples from a Gaussian mixture model (GMM). These include conditional mean imputation [1] and cluster-based imputation [3]. Both methods assume that the GMM has full covariance matrices.

In conditional mean imputation, the masked feature values are assumed independent from the observed noise [1] whereas bounded methods assume that the masked values could not have exceeded the observed feature value [2]. Bounded estimation generally improves the reconstruction performance, but the bounded feature estimates do not have analytical solutions in the conditional mean imputation framework. Previous works [3, 4] have proposed feature estimates that approximate the bounded maximum a posteriori (MAP) or the bounded minimum mean square error (MMSE) estimate. The approximate estimates are calculated based on the clean speech prior and the observed features.

In this work, an approximate posterior model is derived for the bounded clean speech features. The approximate posterior can be used to calculate point estimates for conventional missing-feature reconstruction or the complete posterior can be used for speech recognition as proposed in [5, 6]. In this work, both approaches are investigated. The method performance is evaluated in a large-vocabulary continuous speech recognition task with speech data recorded in real noisy environments, and performance rates with the conditional mean imputation [1] and cluster-based imputation [3] methods are reported for reference.

2. Methods

2.1. Mask estimation

The missing-feature approach [2] is motivated by studies on auditory scene analysis (ASA) which demonstrate that the most intense component in an auditory scene dominates the combined neural response to the scene. The effect is replicated on machines when speech is processed in compressed spectral domain: when speech is corrupted with additive noise, the observed feature component $Y(\tau, d)$ in channel d in frame τ in the compressed spectral domain can be approximated as $Y(\tau, d) \approx \max\{X(\tau, d), N(\tau, d)\}$, where $X(\tau, d)$ denotes the corresponding clean speech component and $N(\tau, d)$ the noise component. The speech-dominated feature components $Y_r(\tau, d)$ are referred to as reliable and the noise-dominated $Y_u(\tau, d)$ as unreliable components.

In this work, the binary mask which divides observations into reliable and unreliable components is estimated based on the negative energy criterion [2] in the log-compressed mel-spectral domain. An observed feature component $Y(\tau, d)$ is considered reliable when

$$\frac{\exp(Y(\tau, d))}{\exp(\hat{N}(\tau, d))} > \gamma, \tag{1}$$

where $\hat{N}(\tau, d)$ denotes a noise estimate in the log-mel domain and γ is a threshold parameter. The noise estimate is calculated based on feature vectors $\boldsymbol{Y}(\tau)$ in frames τ which are classified as non-speech which are temporally smoothed to produce the noise estimate $\hat{\boldsymbol{N}}(\tau)$. The estimated masks are further processed to remove connected regions with less than 5 reliable components as proposed in [7]. Setting a minimum span for the reliable regions simulates the glimpse model proposed in [8].

2.2. Feature reconstruction

The missing-feature methods in noise-robust speech recognition include marginalisation and reconstruction approaches [2]. Marginalisation approaches decode the speech signal based on the reliable observations and often function in the compressed spectral domain whereas reconstruction methods use the observed feature vector $\boldsymbol{Y}(\tau)$ and a clean speech model to recover the clean speech feature vector $\boldsymbol{X}(\tau)$.

Assume the reliable and unreliable components in a noisy observation $\boldsymbol{Y} = \boldsymbol{Y}(\tau)$ are rearranged so that $\boldsymbol{Y} = [\boldsymbol{Y}_r^\mathsf{T} \boldsymbol{Y}_u^\mathsf{T}]^\mathsf{T}$, where $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_u$ denote the reliable and unreliable subvectors in frame τ. We model the corresponding clean speech feature vector as a random variable $\boldsymbol{X}$ with subvectors $\boldsymbol{X}_r$ and $\boldsymbol{X}_u$ that correspond to $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_u$. In this work, the clean speech feature vectors are modelled, as proposed in [3], as independent and identically distributed samples from a full-covariance

GMM,

$$P(\boldsymbol{X}) = \sum_i P(i)N(\boldsymbol{X}|i), \quad (2)$$

where $P(i)$ denotes the weight and $N(\boldsymbol{X}|i)$ the Gaussian distribution associated with the ith mixture component. The component mean vectors are denoted as $\boldsymbol{\mu}_i$ and covariance matrices as $\boldsymbol{\Sigma}_i$. The speech prior is trained on clean speech data as described in Section 3.3.

The missing-feature assumption discussed in Section 2.1 holds that the reliable components in the observed feature vector $\boldsymbol{Y}$ represent clean speech, $\boldsymbol{X}_r = \boldsymbol{Y}_r$, whereas the clean speech features that correspond to the unreliable components are effectively missing. If we make the further assumption that $\boldsymbol{X}_u$ is independent from the unreliable observations $\boldsymbol{Y}_u$, a posterior distribution for the clean speech features can be calculated as

$$P(\boldsymbol{X}|\boldsymbol{Y}_r) = \sum_i P(i|\boldsymbol{Y}_r)N(\boldsymbol{X}|\boldsymbol{Y}_r, i), \quad (3)$$

where $\boldsymbol{Y}_r$ denotes $\boldsymbol{X}_r = \boldsymbol{Y}_r$. $P(i|\boldsymbol{Y}_r)$ denotes the probability that a feature vector with the subvector $\boldsymbol{X}_r = \boldsymbol{Y}_r$ would be sampled from the ith component of the clean speech prior in Equation (2) and $N(\boldsymbol{X}|\boldsymbol{Y}_r, i)$ denotes the ith component-conditional posterior. The mean and covariance of the ith component-conditional posterior are calculated as

$$\boldsymbol{m}_i = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{i,r}\boldsymbol{\Sigma}_{i,rr}^{-1}(\boldsymbol{Y}_r - \boldsymbol{\mu}_{i,r}), \quad (4)$$

$$\boldsymbol{S}_i = \boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_{i,r}\boldsymbol{\Sigma}_{i,rr}^{-1}\boldsymbol{\Sigma}_{i,r}^{\mathsf{T}}, \quad (5)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the clean speech model parameters from Equation (2). $\boldsymbol{\Sigma}_{i,r}$ denotes the covariance between $\boldsymbol{X}$ and $\boldsymbol{X}_r$, $\boldsymbol{\Sigma}_{i,rr}$ the covariance between the reliable components $\boldsymbol{X}_r$, and $\boldsymbol{\mu}_{i,r}$ the mean of $\boldsymbol{X}_r$. Note that the subvector in $\boldsymbol{m}_i$ that corresponds to the reliable subvector $\boldsymbol{X}_r$ equals the observed features $\boldsymbol{Y}_r$, and the corresponding submatrix in $\boldsymbol{S}_i$ are zero. This is in accordance with the assumption that the components $\boldsymbol{Y}_r$ are taken as reliable.

The clean speech features $\boldsymbol{X}$ can be reconstructed based on the posterior $P(\boldsymbol{X}|\boldsymbol{Y}_r)$ but the reconstruction performance improves if the bounds that constrain the missing features, $\boldsymbol{0} \leq \boldsymbol{X}_u \leq \boldsymbol{Y}_u$, are taken into account [2]. The approach proposed in this work is to approximate the posterior with a simple distribution and calculate bounded estimates from the approximate model. A Gaussian approximation for the posterior $P(\boldsymbol{X}|\boldsymbol{Y}_r)$ is optimised based on the approximate Kullback–Leibler divergence between a mixture model and a single distribution. The approximate posterior has mean and covariance

$$\boldsymbol{m} = \sum_i P(i|\boldsymbol{Y}_r)\boldsymbol{m}_i, \quad (6)$$

$$\boldsymbol{S} = \sum_i P(i|\boldsymbol{Y}_r)(\boldsymbol{S}_i + (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^{\mathsf{T}}), \quad (7)$$

where $\boldsymbol{m}_i$ and $\boldsymbol{S}_i$ are the posterior model parameters from Equations (4)-(5). The covariance is further approximated with a diagonal covariance matrix whose diagonal elements correspond to $S(d,d)$ for all channels d. Note that the covariance submatrix and hence the variances that correspond to the reliable feature components remain zero.

The approximate posterior is truncated to constrain the random variable $\boldsymbol{X}$ to non-negative values which do not exceed the observed features $\boldsymbol{Y}$. MMSE estimates are then calculated as the posterior mean as proposed in [9]. The truncated distribution mean $\upsilon = \upsilon(d)$ in channel d is calculated as

$$\upsilon = m + \frac{f(L) - f(U)}{F(U) - F(L)}\sqrt{S}, \quad (8)$$

where $L = m/\sqrt{S}$ and $U = (Y(d) - m)/\sqrt{S}$ and $m = m(d)$ and $S = S(d,d)$ are the mean and variance of the approximate posterior in channel d from Equations (6)-(7). $f(X)$ denotes the probability density function and $F(X)$ the cumulative distribution function of the standard normal distribution. In missing-feature reconstruction, the observed features $\boldsymbol{Y}$ are substituted with the MMSE estimates $\boldsymbol{\upsilon}$.

2.3. Observation uncertainties

The posterior mean $\boldsymbol{\upsilon}$ can be used as a point estimate to substitute $\boldsymbol{Y}$ in the baseline automatic speech recognition system as discussed in the previous section, but the approximate posterior approach also allows for using the full posterior. The variance $\Delta = \Delta(d)$ of the truncated posterior in channel d in frame τ is calculated as

$$\Delta = S\left[1 + \frac{L \cdot f(L) - U \cdot f(U)}{F(U) - F(L)} - \left(\frac{f(L) - f(U)}{F(U) - F(L)}\right)^2\right], \quad (9)$$

where we assume definitions from Equation (8). The feature variances can be used as observation uncertainties [5] to conceal reconstruction errors and improve speech recognition performance as proposed in [6]. The feature posterior in the acoustic model domain is calculated from the posterior in compressed spectral domain as follows.

Features used in our baseline system are outlined in Section 3.2. The feature set includes a log-energy feature calculated as $\log Z = \log\left[\sum_d \exp X(d)\right]$. The mean and variance of the exponential features are calculated as expectation values

$$\nu(d) = E\{\exp X(d)\}, \quad (10)$$

$$\lambda(d) = E\{[\exp X(d)]^2\} - [E\{\exp X(d)\}]^2. \quad (11)$$

When the random variable $X(d)$ follows the truncated Gaussian distribution in Equations (8)–(9), the expectation values for the mean and variance have solutions

$$\nu(d) = \exp\left[m + \frac{S}{2}\right]\frac{F(A) - F(B)}{F(U) - F(L)}, \quad (12)$$

$$\lambda(d) = \exp(2S + 2m)\frac{F(C) - F(D)}{F(U) - F(L)} - \nu(d)^2, \quad (13)$$

where we assume definitions from Equation (8) and additionally define

$$A = (Y(d) - S - m)/\sqrt{S}, B = (S + m)/\sqrt{S},$$
$$C = (Y(d) - 2S - m)/\sqrt{S}, D = (2S + m)/\sqrt{S}.$$

The sum of the exponential components has mean and variance

$$\nu' = \sum_d \nu(d),\ \lambda' = \sum_d \lambda(d). \quad (14)$$

We assume that the sum Z follows a log-normal distribution. It follows from the log-normal assumption that the distribution of $\log Z$ is a Gaussian with mean and variance

$$\upsilon' = \log(\nu') - \frac{\Delta'}{2},\ \Delta' = \log\left[\frac{\lambda'}{\nu'^2} + 1\right]. \quad (15)$$

The other feature operations used in our baseline system can be implemented as linear transformations. Linear transformations in uncertainty propagation are discussed in [10]. The transformations introduce correlations i.e. non-diagonal components in the covariance matrix, but these are often ignored to save computation [10]. In this work, the correlations between feature vector components are ignored, and the differential features are not associated with uncertainties so that the inter-frame correlations introduced in cepstral mean subtraction need not be modelled. The posterior in the acoustic model domain is assumed Gaussian with diagonal-covariance. The posterior mean is used as a feature estimate and the posterior variance is introduced in the acoustic model as a dynamic variance component as proposed in [6].

2.4. Related methods

In this work, bounded feature estimates are calculated from a distribution that approximates the clean speech posterior conditioned on the reliable feature componnents $\boldsymbol{X}_r = \boldsymbol{Y}_r$. Feature reconstruction based on a GMM clean speech prior and the conditional distribution $P(\boldsymbol{X}|\boldsymbol{Y}_r)$ was proposed for state-conditioned imputation in [1]. Methods that use $\boldsymbol{m}$ from Equation (3) to substitute the observed feature vector $\boldsymbol{Y}$ are referred to as mean imputation when $\boldsymbol{\Sigma}_i$ are diagonal covariances and conditional mean imputation when $\boldsymbol{\Sigma}_i$ are full covariances [1]. The estimated clean speech values can be reset not to exceed the observed bounds, $\boldsymbol{m} = \min\{\boldsymbol{m}, \boldsymbol{Y}\}$. Estimates reset in this manner correspond to bounded maximum a posteriori (MAP) estimates calculated from the approximate posterior proposed in this work.

Bounded mean or bounded conditional mean imputation is often based on estimates calculated from the component-conditional posteriors $N(\boldsymbol{X}|\boldsymbol{Y}_r, i)$. When diagonal covariance models are used, the posteriors can be box-truncated to calculate component-conditional bounded MAP estimates [2] or bounded MMSE estimates [9], but in conditional mean imputation, $\boldsymbol{S}_i$ are full covariance matrices and the box-truncated component-conditional posteriors do not have a closed-form solution. Previous works focus on point estimates derived from the posterior. In cluster-based imputation [3], MAP estimation from the component-conditional truncated posteriors is formulated as a constrained optimisation task solved with an iterative method, and in approximate bounded conditional mean imputation [4], component-conditional MMSE estimates are calculated from the box-truncated posterior using the axis-parallel approximation.

Finally, while the previous works on bounded conditional imputation have focussed on point estimates, using observation uncertainties with conditional mean imputation was proposed in [11] and a full posterior approach was proposed for bounded mean imputation in [12]. The heuristic uncertainties proposed in [11] were mapped to the acoustic model domain with a non-linear transformation trained on paired speech and noisy speech data, whereas in [12], the posterior variances were summarised in a single weighting factor $\alpha \in [0, 1]$ for each frame.

3. Experiments

3.1. Evaluation data

Missing-feature reconstruction is evaluated in speech recognition experiments using samples from the Finnish SPEECON database. The samples are read sentences recorded in public or car environments using three microphones. The evaluation data recorded in public environments such as parks and cafeterias contains 878 utterances (94 min) from 30 speakers and the development data 578 utterances (60 min) from 20 speakers. SNR values estimated with the recording platform are on average 24 dB for the headset data (P0), 14 dB for the lavalier data (P1), and 9 dB for the far-field data (P2) which has been recorded 0.5–1 m from the speaker.

The evaluation data recorded in cars contains 575 utterances (57 min) from 20 speakers and the development data 288 utterances (29 min) from 10 speakers. The estimated SNR average values are 13 dB for the headset data (C0), 5 dB for the lavalier data (C1), and 8 dB for the far-field data (C2) which has been recorded with a microphone mounted on the rear-view mirror. The far-field microphone has been designed for a hands-free system and supresses frequencies below 300 Hz and above 6500 Hz. The microphone reduces engine and turbulence noise, which results in a better SNR but degrades the speech recognition performance due to the supressed lowband and highband frequencies.

3.2. Baseline system

The large-vocabulary continuous speech recognition (LVCSR) system used in this work has been described in [13]. The input speech signal is represented with 12 MFCC and a log-energy feature along with their first and second order differentials, and the features are normalised with cepstral mean subtraction (CMS) and maximum likelihood linear transformation (MLLT). The GMM-HMM based acoustic models for context-dependent triphones use gamma distributions to model state durations. The models were trained with the 30-hour SPEECON training set that contains clean speech recorded with a headset in quiet office conditions. The decoder is a time-synchronous beam-pruned Viterbi token-pass system and the language model a morph-based growing n-gram model trained on 145 million words of Finnish book and newspaper data. Since all words and word forms can be represented with the statistical morphs, the decoding vocabulary is in practice unlimited [13]. Speech recognition performance is measured in letter errors rather than word errors.

3.3. Missing-feature system

The speech–non-speech classifier used in mask estimation (Section 2.1) is a two-state HMM where the speech and non-speech state are modelled with 24 Gaussian components. The classifier is trained on features described in Section 3.2 using babble-noise corrupted television news data from the Finnish Broadcasting Company (YLE). The threshold parameter γ in Equation (1) was optimised in decibel scale, $\gamma' = 10 \log_{10} \gamma$, using grid search with 1 dB step. The threshold parameter was optimised based on speech recognition performance measured on the development datasets and the optimal value for both public and car environments is $\gamma' = 3$ dB. The same value is optimal for all the feature reconstruction approaches evaluated in this work.

Missing features are reconstructed in the 21-dimensional log-compressed mel-spectral domain. We compare conditional mean imputation [1], cluster-based imputation [3], and the method proposed in this work. The constrained optimisation problem in cluster-based imputation is solved with an active-set method from the QPC toolbox [14], and in conditional mean imputation, the feature estimates are reset not to exceed the observed upper bound. The clean speech prior used in this work is a 5-component GMM trained with 500 read sentences (52 min)

Table 1: Letter error rates on data recorded in public and car environments.

	P0	P1	P2	C0	C1	C2
(a) baseline	3.4	22.2	38.3	4.2	33.7	67.3
(b) CMI	3.4	13.7	25.3	3.7	22.5	48.3
(c) cluster-based	3.6	14.3	23.1	3.9	19.9	39.6
(d) BCMI	3.7	12.4	21.5	3.7	18.5	33.7

Table 2: Letter error rates obtained with the observation uncertainty system.

	P0	P1	P2	C0	C1	C2
(e) full posterior	3.3	11.8	20.8	3.4	16.7	29.6

from the SPEECON training set. The model was initialised with fuzzy c-means and trained with the EM algorithm implemented in the GMMBAYES toolbox [15].

3.4. Results

Noisy speech recognition results with the uncompensated baseline system (a) and the missing-feature reconstruction systems (b)–(d) are reported in Table 1. The systems compared in this work include conditional mean imputation (CMI) (b), cluster-based imputation (c), and bounded conditional mean imputation (BCMI) with the proposed method (d). Since the conditional mean estimates are reset not to exceed the observed upper bound, estimates (b) and (d) can be interpreted as MAP and MMSE estimates calculated from the approximate posterior proposed in this work. Public environments are indicated with P0–P2 and car environments with C0–C2, where 0–2 indicates the microphone distance as described in Section 3.1. Speech recognition performance is reported in letter error rate (LER) and statistical significance is tested in pairwise comparisons with the Wilcoxon signed rank test at significance level $\alpha = 0.05$.

Bounded conditional mean imputation (d) results in the best performance ($p < 0.02$) in conditions P2 and C2. The relative error reductions are up to 50 % compared to the baseline (a), up to 30 % compared to conditional mean imputation (b), and up to 15 % compared to cluster-based imputation (c). In the other test conditions, the differences between systems are not statistically significant in all pairwise comparisons, and the best system cannot be determined. We note that in the cleanest condition P0, CMI (b) results in a better performance ($p < 0.01$) than BCMI (d). The difference between BCMI (d) and the uncompensated baseline (a) performance in condition P0 is not statistically significant, but the average performance rate on the dataset used in this work is better with the uncompensated baseline system (a).

Results obtained when the posterior variances are used as observation uncertainties in the state-likelihood calculation are reported in Table 2. While the differences are not statistically significant, using the approximate posterior variances as observation uncertainties improves the average performance rate compared to BCMI (d) on datasets P0–P2 and C0–C2. There is, however, a notable difference in the relative error reductions on the datasets P0 and C0–C2 and datasets P1 and P2. The relative error reductions compared to BCMI (d) are around 10 % on datasets P0 and C0–C2, 5 % on dataset P1, and 3 % on dataset P2.

4. Discussion

The missing-feature reconstruction method proposed in this work extends conditional mean imputation [1] to bounded conditional mean imputation based on an approximate posterior. Evaluated in a noise-robust speech recognition task with speech data recorded in real noisy environments, the proposed MMSE estimates outperformed conditional mean imputation and cluster-based imputation on most datasets, but degraded the performance on the low-noise dataset P0. When the posterior variances were included as observation uncertainties, the proposed method outperformed the reference methods on all the datasets.

The results obtained on the low-noise dataset illustrate how using the posterior variance as observation uncertainties can conceal errors made in missing-feature reconstruction. The relative error reductions in this condition, and on the datasets C0–C2, are comparable to the reductions reported in [6]. Evaluated on the datasets P1 and P2, however, observation uncertainties result in relative error reductions less than 5 %. The difference between these datasets could be due to false reliables in the estimated mask: When observation uncertainties are used, the feature components that have been labelled as reliable are emphasised over reconstructed components even if there were mask estimation errors and the reconstructed components were a better representation for clean speech. The mask estimation method used in this work assumes that the noise remains stationary, which means that sudden noise events like clatter or chimes could be labelled as reliable speech.

5. Conclusions

A method for bounded conditional mean imputation with an approximate posterior was derived and evaluated in a missing-feature reconstruction task. The method performance is comparable to cluster-based imputation but the method is non-iterative and therefore requires less computation than cluster-based imputation. Further experiments with artificially noise-corrupted speech data and oracle masks are required to evaluate whether using the posterior variance as observation uncertainties emphasises mask estimation errors.

6. Acknowledgements

This work received financial support from the Academy of Finland under the grants no 140969 and 251170 and from Tekes under the FUNESOMO project.

7. References

[1] A. C. Morris, M. P. Cooke, and P. D. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust ASR," in *Proc. ICASSP*, 1998, pp. 737–740.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, 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.

[4] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features," in *Proc. ICASSP*, 2009, pp. 3869–3872.

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

[6] 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," *Trans. Speech and Audio Processing*, vol. 13, pp. 412–421, 2005.

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

[8] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[9] B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition," in *Proc. ASRU*, 2005, pp. 65–70.

[10] R. F. Astudillo and D. Kolossa, "Uncertainty propagation," in *Robust Speech Recognition of Uncertain and Missing Data*, D. Kolossa and R. Haeb-Umbach, Eds. Springer Verlag, 2011, pp. 35–64.

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

[12] J. A. González, A. M. Peinado, A. M. Gómez, N. Ma, and J. Barker, "Combining missing-data reconstruction and uncertainty decoding for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4693–4696.

[13] 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 and Language*, vol. 20, no. 4, pp. 515–541, 2006.

[14] A. Willis, "QPC - Quadratic programming in C," 2010, Matlab toolbox.

[15] J. Kämäräinen and P. Paalanen, "GMMBAYES - Bayesian classifier and Gaussian mixture model toolbox V1.0," 2005, Matlab toolbox.