Estimating Uncertainty to Improve Exemplar-Based Feature Enhancement for Noise Robust Speech Recognition


Abstract—We present a method of improving automatic speech recognition performance under noisy conditions by using a source separation approach to extract the underlying clean speech signal. The feature enhancement processing is complemented with heuristic estimates of the uncertainty of the source separation, that are used to further assist the recognition. The uncertainty heuristics are converted to estimates of variance for the extracted clean speech using a Gaussian Mixture Model based mapping, and applied in the decoding stage under the observation uncertainty framework. We propose six heuristics, and evaluate them using both artificial and real-world noisy data, and with acoustic models trained on clean speech, a multi-condition noisy data set, and the multi-condition set processed with the source separation front-end. Taking the uncertainty of the enhanced features into account is shown to improve recognition performance when the acoustic models are trained on unenhanced data, while training on enhanced noisy data yields the lowest error rates.

Index Terms—Uncertainty estimation, observation uncertainty, exemplar-based, noise robustness, speech recognition.

EDICS Category: SPE-ROBU

I. INTRODUCTION

PERFORMANCE of conventional automatic speech recognition (ASR) systems is generally strongly diminished in the presence of noise. Feature enhancement methods are a popular approach of mitigating this performance degradation. The goal of a feature enhancement front-end is to produce an estimate of the features that would have been extracted from the underlying clean speech signal, had the noise not been present. From a source separation point of view, the task is to decompose the observed mixture of speech and noise sources, and extract the clean speech signal.

In general, it is not possible for feature enhancement to reconstruct a signal exactly identical to the original clean speech. The performance of noise robust ASR can be further improved by taking into account the varying reliability of the enhanced features, e.g., by using observation uncertainty techniques [1], [2] or uncertainty decoding [3]. In order to utilize these methods, information of the uncertainty of the enhanced features is required.

The uncertainty of observations is commonly modeled either as a binary mask indicating “reliable” and “unreliable” components [4], [5], or by Gaussian distributions for each observation frame [1], [2], [6]–[10], where the mean and covariance represent the observed signal and its uncertainty, respectively. In this latter category, variations of uncertainty approaches have been applied in several scenarios, such as in multitalker conditions [6], [7] or under environmental noise [8], [9]. The use of uncertainty information has also been investigated in conjunction with various feature enhancement methods, such as parametric distortion models [1], ICA source separation [6], auditory scene analysis [7] and multichannel beamforming [9]. In order to transform the uncertainty information from the feature enhancement domain into a domain more suitable for speech recognition, both transformations learned from data [2], [5], [10] and uncertainty propagation approaches based on the feature computation algorithms [6], [8], [9], [11] have been used. In addition to the actual recognition phase, the training of acoustic models for speech recognition has also been improved by taking uncertainty of features into account [8].

A novel feature enhancement method based on sparse source separation for improving noise robustness of ASR has been proposed in [12]. This method is based on representing the noisy observation as a sparse linear combination of atoms in a predefined dictionary of speech and noise samples (exemplars). However, the method does not inherently provide a way of measuring the variance of the produced estimates of the clean speech features. In the present study, we focus on the problem of estimating the reliability of the enhanced features in order to benefit from observation uncertainty techniques, and propose methods for estimating the required variances by using heuristic measures of the feature enhancement uncertainty.

The use of such heuristic measures for exemplar-based representations has been briefly explored in our preliminary work, in conjunction with two feature enhancement methods: the sparse source separation approach investigated in this work [10], as well as a sparse missing data imputation.
approach [2], where the observations considered unreliable are reconstructed based on the reliable values. These earlier investigations have been performed using artificial additive noise scenarios, and acoustic models trained on clean speech data.

For this more detailed account, we focus on the source separation approach and investigate its performance using a realistic large vocabulary speech corpus recorded in real-world noisy environments, as well as on acoustic model training with a multicondition training set. In addition, we introduce two further heuristics, propose using a channel normalization step to enhance the match between the predefined dictionaries and the observed signal, and provide a more detailed analysis of the proposed uncertainty heuristics.

The remainder of this paper is structured as follows. The feature enhancement approach used is described in Section II. The observation uncertainty model is presented in Section III-A, and the heuristic uncertainty estimates are enumerated in Section III-B, while Section III-C explains the selected method of mapping the heuristic estimates to the acoustic model feature domain. The speech recognition system and experiments are described in Section IV, and the results in Section V. The results are discussed in Section VI, and conclusions presented in Section VII.

II. FEATURE ENHANCEMENT OF NOISY SPEECH

A. Exemplar-based Representation of Noisy Speech

The sparse source separation approach used in this work was initially presented in [13]. It is based on representing a noisy magnitude mel-spectrogram as a sparse, non-negative linear combination of example speech spectrograms called exemplars.

We denote by $y$ a $T$-frame magnitude mel-spectrogram of $B$ frequency bands, stacked into a vector of length $TB$. In the case of additive noise, $y$ is approximated as the sum of clean speech spectrogram $s$ and a noise spectrogram $n$. Both the speech and noise are approximated as weighted linear combinations of exemplars from speech and noise dictionaries.

We denote the $J$ exemplars of the clean speech dictionary as $a_j^s$, where $j = 1, \ldots, J$, and the $K$ exemplars of the noise dictionary as $a_k^n$, $k = 1, \ldots, K$, giving the model

$$y \approx s + n$$

$$\approx \sum_{j=1}^{J} x_j^s a_j^s + \sum_{k=1}^{K} x_k^n a_k^n$$

$$= [A^s A^n] [x^s x^n]$$

where $x^s$ and $x^n$ are the activations, i.e., the sparse representations of the underlying speech and noise, respectively.

The complete activation vector $x$ is obtained by solving the constrained optimization problem

$$\min_{x} d(y, Ax) + \| \lambda \cdot x \|_1 \quad \text{s.t.} \quad x \geq 0,$$

where $d(\cdot, \cdot)$ is the generalized Kullback-Leibler divergence, and the second term is a L1 regularization term over the activation vector designed to induce sparsity in the representation. The activations are weighted by a sparsity coefficient vector $\lambda$ using an elementwise multiplication operator $\cdot .$. The optimization problem (5) is solved by using a multiplicative update routine [12].

B. Sparse Separation Feature Enhancement

In order to apply the source separation method to utterances of arbitrary length, we use a sliding time window approach as in [12]. In this approach, the magnitude mel-spectrogram of a noisy utterance is represented using a number of overlapping speech segments, each $T$ frames long.

For each segment $\tau$, we solve the optimization problem described in Section II-A, and obtain corresponding activation vector $x_\tau$. From the sparse representation, we can derive estimates for the clean speech and noise spectrograms as

$$\tilde{s}_\tau = A^s x_\tau^s,$$

$$\tilde{n}_\tau = A^n x_\tau^n,$$

where $x_\tau^s$ and $x_\tau^n$ are the activation vector components corresponding to the clean speech and noise dictionaries, respectively. In order to obtain estimates for each frame $t$, the corresponding frames of the estimates of all windows overlapping frame $t$ are averaged.

Finally, we process the original noisy features $y_\tau$ of frame $t$ as follows. We define a mel-spectral domain Wiener filter $h_t$ as

$$h_t = \tilde{s}_t / (\tilde{s}_t + \tilde{n}_t),$$

where $\cdot / \cdot$ denotes elementwise division. The enhanced magnitude mel-spectral features are then obtained as

$$\hat{s}_t = h_t \cdot \cdot y_\tau.$$
of the noise. The approach is an adaptation of a related method shown to counteract spectral deviation and scaling differences in a missing data ASR scheme [14].

First, each frequency band of the logarithmic mel-scale spectrogram is smoothed by using an unweighted averaging filter over a $T_{\text{smooth}}$-frame neighborhood. Then, the five highest-valued local maxima of each frequency band are located, and their mean is used as the value of the normalization factor corresponding to that frequency band. Choosing local maxima is based on the assumption that they are more likely to correspond strong speech regions and are less biased by the noise component than less energetic regions. The normalization is performed in the logarithmic mel-spectral domain, by subtracting the per-utterance normalization vector from each observation frame. After the feature enhancement processing, the normalization vector is added back to cancel the effects of the normalization. The channel normalization is also applied to data used to construct the speech and noise exemplar dictionaries.

III. OBSERVATION UNCERTAINTIES

The reliability of the clean speech estimates produced by any feature enhancement system varies depending on, e.g., the dynamic noise conditions in the observed signal. Observation uncertainties [1], [15] provide a framework for taking the expected variance of the enhanced features into account in the speech recognition process.

A. Decoding with Observation Uncertainties

The decoding process of a conventional HMM-based speech recognition system combined with feature enhancement is typically based on obtaining an estimate of the underlying clean speech features in the acoustic model domain, $\hat{\varsigma}_t = \text{FE}(y_t, \theta)$, where $y_t$ is the noisy observation, and FE and $\theta$ denote the feature enhancement method and its parameters, respectively. The decoder then produces the output based on the state likelihoods $L(q) = p(\hat{\varsigma}_t | M, q)$, where $q$ denotes a particular state of the acoustic model $M$. From a probabilistic point of view, $\hat{\varsigma}_t$ can be seen as a point estimate of the clean speech features.

When observation uncertainties are used, the point estimates $\hat{\varsigma}_t$ are replaced with posterior distributions $p(\varsigma | y_t, \theta)$ defined by the feature enhancement process. In decoding, the state likelihoods are then calculated by marginalizing over the acoustic model features $\varsigma$, yielding

$$L(q) = \int p(\varsigma | y_t, \theta) p(\varsigma | M, q) \, d\varsigma. \quad (10)$$

Observation uncertainties, and the closely related technique of uncertainty decoding, have been successfully used with various feature enhancement approaches [1]–[3], [6], [7], [9].

When the acoustic model $M$ contains a Gaussian mixture model (GMM) for each state $q$, the total likelihood $L(q)$ is a weighted sum of likelihoods $L(l)$ for each mixture component $l$. If the posterior probability $p(\varsigma | y_t, \theta)$ given by the feature enhancement method is further restricted to be Gaussian as proposed in [1], the likelihood computation reduces to

$$L(l) = \int N(\varsigma | \hat{\varsigma}_t, \Sigma_l) N(\varsigma | \mu^{(l)} + \Sigma^{(l)} \sigma_t) \, d\varsigma$$

where $\mu^{(l)}$ and $\Sigma^{(l)}$ are the mean and covariance of mixture component $l$ in the acoustic model, and $\hat{\varsigma}_t$ and $\Sigma_l$ denote the estimated mean and covariance of the clean speech posterior for frame $t$.

The heuristic uncertainty estimates presented in Section III-B produce, for each speech frame, either a single scalar value estimating the overall uncertainty, or a vector in the mel-spectral domain. In order to apply (12), however, a covariance estimate $\Sigma_l$ in the feature domain used by the acoustic models is required. In this work, we use diagonal covariances $\Sigma_l = \text{diag}(\sigma_l)$ and, following [5], a supervised learning approach to transform the heuristic values into variance estimates $\sigma_l$ in the acoustic model domain, described in Section III-C.

B. Uncertainty Heuristics

The feature enhancement system used in this work produces only a point estimate of the clean speech features. In our preliminary studies, heuristic measures to estimate the feature reliability have been successfully applied to the sparse separation feature enhancement method considered in this work [10] as well as sparse imputation based methods [2]. Empirical measures have also been considered in the context of uncertainty propagation [11].

In this work, six heuristics are considered. Heuristics 1, 4, 5 and 6 also appear in our earlier work [10], while the present study proposes the additional heuristics 2 and 3 as variants based on the magnitude of changes between the noisy and enhanced features.

Notation: In the descriptions, $H_n(t, m)$ is the estimated uncertainty for heuristic number $n$, frame $t$ and mel-spectral band $m$. For those heuristics that only produce a single scalar for each frame, $m$ has been omitted. The scaling function $S(x)$, used in heuristics H4–H6, denotes a linearly scaled version of $x$, so that over a single utterance, the resulting values range from zero to one. This scaling is used with heuristics whose absolute values vary widely for different utterances.

Relative difference (H1): If the enhanced features differ much from the observed noisy features, the uncertainty of the reconstruction is also likely to be relatively large. Therefore we propose to use the relative difference between noisy and enhanced features in each mel-spectral band as an uncertainty measure:

$$H1(t, m) = \frac{|y_t - \hat{y}_t|_m}{|y_t|_m} = 1 - \frac{|\hat{y}_t|_m}{|y_t|_m} \quad (13)$$

Given that $|\hat{y}_t|_m = |h_t|_m |y_t|_m$, H1 can also be expressed in terms of the Wiener filter, as

$$H1(t, m) = 1 - |h_t|_m, \quad (14)$$

and is related to using the variance of the Wiener filter as the uncertainty estimate, as proposed in [11].
Absolute difference (H2): Analogously to H1, we define the second heuristic by considering the squared absolute difference of the noisy and enhanced features:

$$H2(t, m) = \log ([y_t]_m - [\hat{s}_t]_m)^2. \quad (15)$$

The difference between noisy and enhanced features has been used directly as an empirical uncertainty estimate in the STFT domain, in [6] with a single scaling factor determined from oracle data, and in the uncertainty training framework of [8] with a scaling factor varying as a function of the frequency band. In this work, the absolute differences are additionally log-compressed in order for their distribution to be easier to model using Gaussian mixtures, as needed for the mapping to the acoustic model domain described in Section III-C.

Logarithmic difference (H3): As a final variant of the heuristics based on the magnitude of changes caused by the feature enhancement, we define the third heuristic as the difference between the noisy and enhanced log-scale mel-spectral features:

$$H3(t, m) = \log y_t - \log \hat{s}_t. \quad (16)$$

The log-scale enhanced features are directly used in the computation of the MFCC features used by the acoustic modeling of the speech recognition system. Again, given that $[\hat{s}_t]_m = [h_u]_m[y_t]_m$, H3 reduces to

$$H3(t, m) = -\log [h_u]_m. \quad (17)$$

Number of active exemplars (H4): When there is significant mismatch between the observed signals and the speech dictionary or the noise dictionary, there are no single exemplars that would be close to the speech or noise sources of the observation, and the sparse representation will consist of multiple exemplars. The uncertainty of the reconstruction in this case can be expected to be relatively large. Accordingly, we can make the uncertainty proportional to the number of significantly non-zero components in the activation vector:

$$H4(t) = S \left( \sum_{\tau} \sum_i I(x_{\tau,i} > T_{H4} \sum_i x_{\tau,i'}) \right), \quad (18)$$

where the summation index $\tau$ ranges over the speech segments that contain frame $t$, indices $i$ and $i'$ sum over all dictionary exemplars, $x_{\tau,i}$ are the activation vector components, and the indicator function $I(p) = 1$ when the proposition $p$ is true, and otherwise zero. The term $T_{H4} \sum_i x_{\tau,i}$ is used to make the threshold value $T_{H4}$ relative to the overall activation magnitudes at frame $\tau$. The value of $T_{H4}$ was chosen by small-scale testing, and is used both to provide a tunable free parameter, as well as to take into account that the iterative matrix factorization solution never results in truly zero activation vector components.

Number of active clean speech exemplars (H5): Similarly to H4, when using the sparse separation approach, we can consider how well the underlying clean speech is being modeled by restricting the summation to the part of the activation vector corresponding to the clean speech dictionary:

$$H5(t) = S \left( \sum_{\tau} \sum_{i} I(x^c_{\tau,i} > T_{H5} \sum_{i} x^c_{\tau,i'}) \right). \quad (19)$$

Noise and clean speech exemplar weights (H6): When the activation vector weights corresponding to noise exemplars are large compared to those corresponding to clean speech exemplars, the observed signal is likely to have been relatively noisy, and as a result the clean speech estimate can be considered to be more uncertain. Therefore we propose to set the observation uncertainty proportions to the ratio of total weight given to noise and speech exemplars:

$$H6(t) = S \left( \frac{\sum_{\tau} \sum_{i} x^t_{\tau,i}}{\sum_{\tau} \sum_{i} x^c_{\tau,i}} \right), \quad (20)$$

where $\tau$ again ranges over the segments containing frame $t$. This quantity acts as a type of an estimate for the SNR within a frame, though it is also influenced by how well the exemplars in the noise dictionary match the noise component of the observations.

Oracle uncertainty: When knowledge of the true underlying clean speech signal is available, it is possible to derive an oracle uncertainty estimate in the acoustic model feature domain as the squared error between the clean features and the enhanced features:

$$ORACLE(t, i) = (M(y_t, i) - M(s_t, i))^2, \quad (21)$$

where $M(x, i)$ denotes the $i$th component of the acoustic model features given by processing the observation $x$. This oracle estimate is used in experiments performed on artificial noisy data as a comparison method for the heuristic approaches.

C. Uncertainty Mapping

The uncertainty measures proposed in Section III-B characterize the uncertainty of either the produced mel-spectral feature components or the entire frame. In order to use the observation uncertainty technique described in Section III-A, however, variance estimates in the acoustic model domain are required. Following [10] and [16], in this work a GMM is used to transform the heuristic value, denoted $H$, into the estimated acoustic model domain observation uncertainty $\sigma$.

The GMM models the joint distribution of $H$ and $\sigma$ as

$$p(z) = \sum_k P(k)N(z | \mu^{(k)}, \Sigma^{(k)}), \quad (22)$$

where $z = [H \ log \sigma]^T$ are the concatenated uncertainty values, $k$ is a mixture component index, $P(k)$ is the mixture component weight, and $\mu^{(k)}$ and $\Sigma^{(k)}$ the mixture component mean and covariance. The acoustic model domain uncertainties $\sigma$ are logarithmically compressed in order for their distribution to better fit the Gaussian assumption.

Given an uncertainty estimate $H$, calculated using one of the heuristics for frame $t$, the minimum mean square error (MMSE) estimate given by the GMM for the corresponding acoustic model domain uncertainties is calculated as

$$E\{\log \sigma_t | H_t, \Lambda\} = \sum_k P(k | H_t, \Lambda)E\{\log \sigma_t | H_t, \Lambda, k\}, \quad (23)$$

where $\Lambda$ denotes the GMM parameters. The posterior probabilities $P(k | H_t, \Lambda)$ for cluster $k$ are calculated using the
component weights $P(k)$ and the likelihoods $p(H_t | \Lambda, k)$ based on diagonal covariances. The cluster-conditional estimates of (23) are calculated as

$$E\{\log \sigma_t | H_t, \Lambda, k\} = \mu_\sigma + \Sigma_{Ht}^{-1} \Sigma_{HH}^{-1} (H_t - \mu_H),$$

(24)

where $\mu_H$ and $\mu_\sigma$ are the means of heuristic and acoustic model domain uncertainties, $\Sigma_{Ht}$ the cross-covariance, and $\Sigma_{HH}$ the covariance of the heuristic values. These means and covariances are subsets of $\mu(k)$ and $\Sigma(k)$ of (22).

The GMM parameters $\Lambda$ for a particular heuristic are obtained with expectation-maximization (EM) training on an artificial noisy data set using the oracle estimate of the variances in the acoustic model domain as the target $\sigma$ vector.

IV. EXPERIMENTAL SETUP

A. Data

The speech recognition experiments were performed on material from the SPEECON [17] Finnish language corpus. Two training sets consisting of approximately 21 hours of speech were defined. The clean speech training set contained read speech from 293 speakers, recorded in a relatively quiet office environment with a close-talking headset microphone. The multicondition training set was constructed by replacing half of the clean speech utterances with noisy data. The noisy samples included recordings in a moving car, both indoors and outdoors in public places, as well as indoor environments with television or music playing in the background. Close-talking, lavalier and medium-distance (approximately 1 meter from the speaker) microphones were used to record the noisy utterances.

For artificial noisy mixture experiments, headset microphone recordings from 40 speakers in the office environments were used as the source of clean speech data. To obtain the desired SNR, the clean speech sentences were mixed together with suitably scaled SPEECON recordings containing no speech activity, from both the car and public place environments. For both environments, data sets with nominal SNRs of 5, 10 and 15 dB were constructed. Each data set was based on the same clean speech recordings, but used different randomly selected segments of noise. Additionally, a separate data set was constructed from 500 utterances randomly selected from the clean speech training set, mixed with noise recordings from both environments, at SNR levels of 4–20 dB. This data set, denoted “Map”, was used for training the mapping of uncertainty heuristics to the acoustic model domain. Both unweighted and A-weighted SNR measurements for the artificial noisy data sets are shown in Table I. As the spectral properties of the two noise types are very different, using the same chosen unweighted SNR value results in significantly different A-weighted SNR values.

Speech recognition performance on real noisy utterances was evaluated using noisy speech test sets from the moving car (“Car”) and public place (“Public”) environments. The car and public place evaluation sets contained phonetically rich sentences of read speech from 20 and 30 individual speakers, with total lengths of 57 and 94 minutes, respectively. The SPEECON read speech text material was collected from text files found in the Internet, and was not restricted to any limited vocabulary.

Each utterance was simultaneously recorded using a close-talking headset microphone (“mic 1”), a lavalier microphone positioned between the chin and shoulder of the speaker (“mic 2”) and a medium-distance microphone (“mic 3”), in both the car and public place conditions, forming a total of six separate evaluation data sets. Average SNR estimates as provided by the SPEECON recording platform for the three microphones were 14 dB, 5 dB and 8 dB for the car recordings, and 24 dB, 14 dB and 9 dB for the public place recordings.

For each evaluation set, there was a corresponding development data set of approximately half the length, but equivalent in other respects. No speakers were shared between the training, evaluation and development data sets. The development sets were used for small-scale experiments to determine parameters for the feature enhancement and speech recognition systems.

B. Feature Enhancement

The feature enhancement front-end operated on 21-dimensional mel-scale spectral features computed from 16 kHz audio signals, using Hamming window frames of length 16 ms (256 samples) and a frame step of 8 ms. Throughout the experiments, a window size of $T = 15$ frames and a window step of a single frame was used for the sliding-window source separation algorithm described in Section II-B. The window size was selected based on small-scale experiments on the development sets.

The clean speech exemplar dictionary for the source separation feature enhancement system was constructed by extracting 20000 random windows of speech from the SPEECON clean speech acoustic model training set. Similarly, a single fixed noise dictionary used was constructed by taking 4000 random windows from the non-speech regions in microphone 2 and 3 recordings in the car and public place environments contained in the SPEECON multicondition training set. As the SPEECON utterances contain relatively long leading and trailing segments with no speech activity, the first and last five windows of the utterance currently being processed were also dynamically added to the noise dictionary. Both dictionary sizes were selected based on small-scale experiments.

A simple voice activity detector (VAD) based on a background noise estimate formed from the leading and trailing non-speech parts of the recordings was used in locating

| TABLE I - SIGNAL-TO-NOISE RATIOS OF ARTIFICIAL NOISY DATA SETS |
|--------------------------------|------------------|------------------|
| Car, 5 dB                 | 5.0              | 3.4              |
| Car, 10 dB                | 10.0             | 8.3              |
| Car, 15 dB                | 15.0             | 13.3             |
| Public, 5 dB              | 5.0              | 9.9              |
| Public, 10 dB             | 10.0             | 15.0             |
| Public, 15 dB             | 15.0             | 19.9             |
| Map                       | 7.8              | 11.6             |
probable speech and non-speech frames in the training data for constructing the dictionaries. Frames were classified as speech if the Euclidean distance of the observed mel-spectral feature from the mean of the first and last 60 frames, which were assumed to have no speech activity, was greater than a threshold value. The threshold was set to be \( 0.15D_{\text{max}} \), where \( D_{\text{max}} \) was the maximum distance over the entire utterance.

The sparsity coefficient vector \( \lambda \) in (5) was constrained to be of the form \( [\lambda^s \cdots \lambda^n \lambda^n \cdots \lambda^n] \), with \( \lambda^s \) and \( \lambda^n \) corresponding to exemplars of the speech and noise dictionaries, respectively. Values of \( \lambda^s = 1, \lambda^n = 0.5 \) were selected based on small-scale experiments on the development set.

The three microphones used for the SPEECON recordings had significantly different frequency response characteristics, so the normalization scheme described in Section II-C was used. Both the utterances used for dictionary construction and for speech recognition experiments were normalized with the same method. The smoothing filter parameter \( T_{\text{smooth}} \) was set to 10 frames.

The recorded utterances in the SPEECON corpus have substantial segments containing only leading and trailing noise, often for more than a third of the total utterance length. When the source separation feature enhancement is applied to these utterances, the parts with no voice activity will be heavily attenuated. However, small deviations in the Wiener filter coefficients caused by weak, transient noises will show up as highly visible structures in the logarithmic spectral domain, leading the speech recognition system to assign spurious phonemes for the sections with no real speech content. To alleviate this problem, the Wiener filter coefficients were thresholded so that \( h_t \geq 0.05 \) for all frames and frequency bands.

Threshold values \( T_{H4} \) and \( T_{H5} \) used in (18) and (19) for determining the number of significantly nonzero activations for uncertainty heuristics H4 and H5 were set to \( T_{H4} = 0.02 \) and \( T_{H5} = 0.001 \), based on development set experiments.

C. Speech Recognition System

Three sets of acoustic models were trained for the speech recognition experiments. A baseline clean speech model was trained on the SPEECON clean speech training set, and a baseline multicondition model was trained on the multicondition training set. Finally, a feature enhancement model was trained on the multicondition training set processed using the source separation feature enhancement front-end.

A large vocabulary continuous speech recognition system was used to perform the speech recognition experiments. The acoustic models were based on cross-word triphones modeled as hidden Markov models using Gaussian mixtures for observation probabilities. On average mixtures of 16 Gaussians were used to model the speech feature space. Additionally, a separate Gamma probability distribution was used for state duration modeling [18].

Language model used by the recognizer was a variable-length n-gram model trained with a growing method on a Finnish language book and newspaper data set of approximately 145 million words. The language modeling units were statistical morphemes learned from the text corpus with an unsupervised method [19]. A single-pass time-synchronous Viterbi beam search algorithm was used by the decoder to combine the acoustic and language model scores. Scaling factor parameters of the language model scores for each of the evaluation sets were based on recognition results on the corresponding development data sets.

V. RESULTS

The beneficial effect of the channel normalization method proposed in Section II-C was verified by performing a small-scale speech recognition experiment. Results of this experiment are presented in Section V-A. The proposed uncertainty heuristics are analyzed in detail in Section V-B, primarily by considering the correlations of both the predicted variances as well as errors of the predicted variances for different heuristics. Finally, results of the speech recognition performance evaluation are reported in Section V-C.

Letter error rate (LER) percentages have been used as the primary performance metric for speech recognition accuracy in all the recognition experiments in this work. As Finnish is an agglutinative language, Finnish words are often long and consist of several morphemes, such as the word ‘kahvinjuokselkin’, representing the entire phrase “also for a coffee drinker”. The LER metric is therefore seen to correspond better to the conventional word error rate (WER) used for languages like English, whereas using the WER metric with Finnish would be closer to measuring sentence or phrase error rates.

A. Channel Normalization

The recognition performance experiment for the channel normalization method was performed using the development (parameter-tuning) data set of real car and public place noise. Results of the experiment are summarized in Table II. In the table, abbreviation “BL” denotes the baseline method, which does not use any feature enhancement, while “FE” and “FE+N” denote the sparse source separation feature enhancement without and with the described channel normalization, respectively. Performing the normalization step improved the recognition accuracy for all test conditions except the least noisy car recording channel, where using the feature enhancement in general had a detrimental effect on recognition.

B. Analysis of Uncertainty Heuristics

In order to compare the behavior of the uncertainty heuristics under car and public place noise, histograms of the
distribution of the individual uncertainty heuristic samples, for the noisiest (microphone channel 3) real noise development set data are presented in Fig. 1. In general, the distributions for the car and public place noise types do not differ from each other qualitatively.

For heuristics H1 and H3, there are large peaks at heuristic values 0.95 and $-\log 0.05 \approx 3$, respectively, where the fraction of occurrences is approximately 0.28. These peaks are artificial, and are due to the filter thresholding constraint $h_{t} \geq 0.05$, described in Section IV-B, which inflates the amount of time-frequency components $t, m$ for which $[h_{t}]_{m} = 0.05$. The locations of the peaks follow directly from expressing the heuristic values of H1 and H3 in terms of the Wiener filter coefficients, as seen in (14) and (17).

Uncertainty heuristic H6, based on the ratio of speech and noise dictionary activation weights, shows a bimodal structure. Based on a manual examination of a sampling of frames with H6 values in both modes, the modes seem to correspond with whether the observed frame contains any speech activity at all.

To determine whether individual heuristics are relatively better suited for different conditions, pairwise Pearson’s correlation coefficients were computed between the errors made by the uncertainty heuristics, i.e., the absolute difference between the estimated and oracle observation uncertainties in the acoustic model domain. The pairwise correlation matrices are presented in Fig. 2. The correlations have been computed over the combined artificial noisy data sets of all three SNR levels (5 dB, 10 dB and 15 dB) of a particular noise type. Corresponding correlations between the heuristic predictions themselves are shown in Fig. 3.

The oracle variances contain occasional large outliers that would have biased the correlation coefficients, as they are not predicted by any of the heuristics. Therefore, only acoustic model feature components where the oracle variance was within two standard deviations from the mean, 97% of the total number, were included in the correlation analysis of the errors.

In order to assess the effectiveness of considering the different mel frequency bands separately, the heuristics H1, H2 and H3 were compared against variants which used an average across the frequency bands as a scalar uncertainty estimate for a single frame. Table III compares the obtained letter error rate percentages, when using the source separation front-end in conjunction with acoustic models trained on clean speech. In general, there were no significant differences between the original heuristics and the averaged versions.

### C. Speech Recognition Experiments

The speech recognition performance of all six evaluated uncertainty heuristics, in the case of acoustic models trained with clean speech, is shown in tables IV and V for the artificial and real noisy data sets, respectively. As the underlying clean speech signal is known in the case of artificial noisy data, Table IV also includes the speech recognition results obtained
TABLE IV
RECOGNITION ERROR RATES OF UNCERTAINTY HEURISTICS WITH CLEAN SPEECH MODELS FOR ARTIFICIAL NOISY DATA

<table>
<thead>
<tr>
<th>SNR (dB):</th>
<th>Map</th>
<th>Car</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>55.1</td>
<td>29.3</td>
<td>64.2</td>
</tr>
<tr>
<td>SS</td>
<td>25.8</td>
<td>11.7</td>
<td>22.6</td>
</tr>
<tr>
<td>SS + H1</td>
<td>26.4</td>
<td>11.0</td>
<td>22.4</td>
</tr>
<tr>
<td>SS + H2</td>
<td>25.7</td>
<td>10.8</td>
<td>21.6</td>
</tr>
<tr>
<td>SS + H3</td>
<td>26.3</td>
<td>11.0</td>
<td>22.1</td>
</tr>
<tr>
<td>SS + H4</td>
<td>25.3</td>
<td>10.9</td>
<td>21.4</td>
</tr>
<tr>
<td>SS + H5</td>
<td>25.2</td>
<td>10.5</td>
<td>21.1</td>
</tr>
<tr>
<td>SS + H6</td>
<td>25.7</td>
<td>10.8</td>
<td>21.5</td>
</tr>
<tr>
<td>SS + oracle</td>
<td>18.3</td>
<td>7.5</td>
<td>13.8</td>
</tr>
</tbody>
</table>

TABLE V
RECOGNITION ERROR RATES OF UNCERTAINTY HEURISTICS WITH CLEAN SPEECH MODELS IN REAL NOISE

<table>
<thead>
<tr>
<th>mic:</th>
<th>Car</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>4.3</td>
<td>29.5</td>
</tr>
<tr>
<td>SS</td>
<td>5.7</td>
<td>13.3</td>
</tr>
<tr>
<td>SS + H1</td>
<td>4.8</td>
<td>11.1</td>
</tr>
<tr>
<td>SS + H2</td>
<td>4.7</td>
<td>11.0</td>
</tr>
<tr>
<td>SS + H3</td>
<td>4.8</td>
<td>11.3</td>
</tr>
<tr>
<td>SS + H4</td>
<td>4.9</td>
<td>10.4</td>
</tr>
<tr>
<td>SS + H5</td>
<td>4.8</td>
<td>10.4</td>
</tr>
<tr>
<td>SS + H6</td>
<td>5.1</td>
<td>11.2</td>
</tr>
</tbody>
</table>

TABLE VI
RECOGNITION ERROR RATES ON ARTIFICIAL NOISY UTTERANCES

<table>
<thead>
<tr>
<th>SNR (dB):</th>
<th>Car</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>25.8</td>
<td>11.7</td>
</tr>
<tr>
<td>SS</td>
<td>26.4</td>
<td>11.0</td>
</tr>
<tr>
<td>SS + H1</td>
<td>26.3</td>
<td>11.0</td>
</tr>
<tr>
<td>SS + H2</td>
<td>25.7</td>
<td>10.8</td>
</tr>
<tr>
<td>SS + H3</td>
<td>25.3</td>
<td>10.9</td>
</tr>
<tr>
<td>SS + H4</td>
<td>25.2</td>
<td>10.5</td>
</tr>
<tr>
<td>SS + H5</td>
<td>25.7</td>
<td>10.8</td>
</tr>
<tr>
<td>SS + oracle</td>
<td>18.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>

TABLE VII
RECOGNITION ERROR RATES ON REAL NOISY UTTERANCES

<table>
<thead>
<tr>
<th>mic:</th>
<th>Car</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>4.3</td>
<td>29.5</td>
</tr>
<tr>
<td>SS</td>
<td>5.7</td>
<td>13.3</td>
</tr>
<tr>
<td>SS + H1</td>
<td>4.8</td>
<td>11.1</td>
</tr>
<tr>
<td>SS + H2</td>
<td>4.7</td>
<td>11.0</td>
</tr>
<tr>
<td>SS + H3</td>
<td>4.8</td>
<td>11.3</td>
</tr>
<tr>
<td>SS + H4</td>
<td>4.9</td>
<td>10.4</td>
</tr>
<tr>
<td>SS + H5</td>
<td>4.8</td>
<td>10.4</td>
</tr>
<tr>
<td>SS + H6</td>
<td>5.1</td>
<td>11.2</td>
</tr>
</tbody>
</table>

when using oracle uncertainty values. In both tables, the best-performing heuristic is indicated in bold.

Based on overall performance, uncertainty heuristics H4 and H5 were chosen to be used for comparisons between methods in real and artificial noisy data experiments, respectively. Obtained letter error rates for the three acoustic model training sets (clean, multicondition, and feature-enhanced multicondition) are presented in Table VI for the artificially mixed noise, and Table VII for the realistic noisy utterances. In both tables, results are given for the evaluation utterances both without and with the feature enhancement front-end processing, and furthermore in the latter case both without and with the heuristic observation uncertainties used in the decoding step. Each system is denoted by an abbreviation of the form “X-Y”, where X indicates the acoustic model training set (“CL” clean, “MC” multicondition”, “FE” feature enhanced multicondition) and Y indicates the feature enhancement frontend in use (“NO” none, “SS” sparse separation, “UC” sparse separation enhanced with observation uncertainties). The best-performing method for a particular acoustic model is indicated in bold, while the best performance overall is underlined.

The results of statistical significance testing of the results in Tables IV–VII are shown in Appendix A.

VI. DISCUSSION

A. General Discussion

All six proposed uncertainty heuristics improve the recognition performance of the feature enhancement front-end when used in conjunction with acoustic models trained on clean speech and in the presence of noise. Applying the source separation front-end in conjunction with acoustic models trained using the (unenhanced) multicondition training set is beneficial for the noisier scenarios: all except one of the artificial noisy data sets, as well as the “mic 3” recording channel of the real noisy utterances. However, additionally using the heuristic-derived observation uncertainties when decoding does improve recognition performance over the results of the feature enhancement alone, in all test settings.

Finally, the feature enhancement front-end combined with acoustic models trained on the enhanced features of the multicondition training set produces the overall best recognition results for all experimental conditions involving real noisy utterances. In the case of artificial noisy data, the situation is similar with the exception of the noisiest data sets with an SNR of 5 dB, where the combination of the baseline multicondition-trained model, the source separation front-end, and the heuristic observation uncertainties outperforms the models trained on enhanced features.

Use of the heuristic observation uncertainties during decoding does not improve the ASR performance of the model trained with enhanced features using the multicondition training set. The most likely explanation is that this is due to a conflict with the assumptions of the observation uncertainty approach. When observation uncertainties are used in the decoding, the acoustic model probability densities are assumed to model the unseen clean speech, and feature uncertainty is added to compensate the model variances [15]. If the model has been trained with enhanced features, the acoustic
model parameters already take into account the distortion introduced by the feature enhancement front-end. This leads to an overestimation of the noise variance, as the sum of the contributions from the observation uncertainties during decoding, and the setting of the acoustic model parameters when training with noisy data [8].

For proper handling of observation uncertainties in the case of training on noisy data, the acoustic model estimation process would need to be modified to consider the variances of the training features. This approach has been recently used in the uncertainty training framework presented in [8]. Making use of the proposed heuristic variances also during the model training process is a potential direction for future work.

The proposed channel normalization has a clear beneficial effect on recognition accuracy of noisy speech. The effect is especially notable in the case of recording channel 3 of the noisy car data set. The car environment has a large low-frequency noise component for which the three microphones used have very different frequency responses, leading to a larger degree of potential mismatch between the observations and the dictionary elements.

B. Comparing and Combining Heuristics

In earlier work with the sparse separation front-end using artificial factory noise mixtures [10], the overall best performing heuristic was the one denoted by H1 in this work. By contrast, in sparse imputation experiments using artificially added babble noise [2] heuristics based on the activation vectors, such as H4 and H5, have been the best performing. In this work, the best performance when considered over all the evaluation data sets is achieved by heuristics H4 and H5, suggesting that the difference between suitability of the heuristics depends more on the type of the noise involved than in the feature enhancement method. The strong performance of the H1 heuristic for some of the real-world test conditions ("Public, mic 1" and "Public, mic 2") over the otherwise generally better H4 and H5 is further suggestive of this. A similar effect of H1 performing relatively well compared to the other heuristics is also seen for the public place artificial noisy data at the higher SNRs.

Overall, the heuristics make highly similar errors, as evidenced by generally high correlation coefficients in the analysis of the correlations between errors in Section V-B. The three heuristics directly based on the exemplar activations (H4, H5 and H6) in particular produce very similar errors in the acoustic model domain estimates, especially in the case of car noise. The correlations between the actual estimated variances show a similar structure, though less strongly, as the correlations between the errors. In both, heuristics H1 and H3, as well as the group of H4, H5 and H6 are each seen relatively similar.

The simple combination of all six uncertainty heuristics by concatenation was investigated using a small-scale recognition experiment. The combined heuristic did not, however, achieve any performance improvement over the best-performing single heuristics, as was also the case in previous work [10]. As mentioned, the errors made by the individual heuristics correlate strongly, which suggests that combining them does not yield much new information.

Similarly, no difference in recognition performance was observed between the heuristics H1–H3 and their corresponding versions averaged across frequency bands. While averaging the values discards information on the relative uncertainty of different frequency bands, the effect is likely diminished by the MFCC processing of the acoustic model features, as the individual feature vector components do not depend on single frequency bands.

Individual GMMs for each heuristic estimate were trained for the mapping from heuristic values to acoustic model domain variance estimates. For all the heuristics, however, the training set and the target oracle variances were the same. This identical postprocessing is a potential cause for the similarity of the heuristics.

As heuristic H6 is based on the ratio of the sums of the activation weights for the noise and clean speech dictionaries, it acts as a type of an SNR estimate. Its value is, however, also affected by how well the dictionary examples match the observed signal. A large mismatch in the noise dictionary would result in lower noise activations, causing a bias for overestimating the SNR. The use of a conventional SNR estimator as an uncertainty heuristic is a potential topic for future work.

C. Uncertainty Propagation

In the GMM based transformation of heuristic values into acoustic model domain variance estimates, described in Section III-C, all variance components of the final observation considered by the acoustic models are estimated from a single observation frame. However, the acoustic model feature extraction involves the computation of first and second-order time derivatives (delta features) across multiple consecutive frames. If the distributions of the static MFCC features of each frame are assumed to be independent and Gaussian, the final acoustic model domain features also have Gaussian distributions, and can be easily computed, as the involved operations are all linear.

Preliminary experiments within the context of a system described in [16], using the sparse imputation approach on the CHiME [20] artificial noisy corpus, yielded results suggesting that using the GMM transformation to obtain variance estimates for static MFCC features and deriving the final observation variances from these estimates may improve recognition performance. Early experiments in using this scheme with the SPEECON real noise data, however, have failed to yield consistent improvements.

Deriving the acoustic model domain variances from heuristic variance estimates of the static MFCC features can be seen as a variant of model-based uncertainty propagation, which has also been applied directly from time-frequency domain to MFCC features [8], [11]. The scheme could therefore be expanded by applying uncertainty propagation to heuristic-based uncertainty estimates in the mel-spectral domain.

In addition, the non-negative matrix factorization (NMF) algorithm used by the feature enhancement system can be
formulated in a Bayesian probabilistic framework [21], potentially allowing an alternative method of deriving the posterior distributions, and hence observation uncertainties, of the enhanced features. Combining the proposed heuristic estimates or Bayesian NMF variances with different levels of propagation of uncertainty information will be investigated in future work.

VII. CONCLUSIONS

In this study, we present a combination of a source separation feature enhancement front-end with heuristic measures of the reliability of the feature enhancement. The proposed system was evaluated under both artificial and real-world noisy conditions, and with acoustic models trained on clean speech data, multicondition noisy data as well as multicondition data processed using the feature enhancement algorithm.

Use of the heuristic observation uncertainty estimates generally improves the recognition performance of noisy speech, when the acoustic models have been trained on unenhanced data. The differences between the individual proposed heuristics are relatively minor.

Best overall results are obtained with a multicondition training set that has been processed using the same feature enhancement algorithm. Taking advantage of the heuristic observation uncertainties also in the acoustic model training is a possible avenue for future work.

APPENDIX A

STATISTICAL SIGNIFICANCE ANALYSIS

Statistical significance of the recognition letter error rate differences between each system pair was analyzed using the Wilcoxon signed-rank test, with a confidence level of 95%.

Tables VIII and IX contain the results of all pairwise tests pertaining to the uncertainty heuristic performance evaluations presented in Tables IV and V, for the artificial and real noise scenarios, respectively. For each pair, the better system is indicated, if the difference between systems was found statistically significant. The analysis was performed over the combined data sets of both noise types and all SNR levels or recording channels. The baseline result is denoted by the abbreviation “BL”, and the oracle heuristic by “OR”.

Summaries of the pairwise tests performed on the final recognition results on artificial and real noise, presented in Table VI and Table VII, are listed in Table X and Table XI, respectively. The analysis has been performed separately for each data set, and a statistically significant difference between systems is indicated with a “+” sign.

REFERENCES


Heikki Kallasjoki

Heikki Kallasjoki received the M. Sc. (tech.) degree in computer science and engineering from the Helsinki University of Technology, Espoo, Finland, in 2009.

He is currently pursuing the Ph.D. degree at the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research field is automatic speech recognition, in particular noise robust ASR.

Jort F. Gemmeke

Jort Florent Gemmeke is a postdoctoral researcher at the KU Leuven, Belgium. He received the M.Sc. degree in physics from the Universiteit van Amsterdam (UvA) in 2005. In 2011, he received the Ph.D. degree from the University of Nijmegen on the subject of noise robust ASR using missing data techniques.

He is known for pioneering the field of exemplar-based noise robust ASR. His research interests are automatic speech recognition, source separation, noise robustness and acoustic modeling, in particular exemplar-based methods and methods using sparse representations.

Kalle J. Palomäki

Kalle Palomäki received the D.Sc. (Ph.D.) in technology degree in electrical and communications engineering from the Helsinki University of Technology, Espoo, Finland, in 2005.

He is currently an Academy Research Fellow (by the Academy of Finland) in the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests are in speech technology, automatic speech recognition (ASR), noise-robust ASR, hearing inspired models in ASR, and audio information access.