Conditional Emission Densities for Combining Speech Enhancement and Recognition Systems

Armin Sehr¹, Takuya Yoshioka², Marc Delcroix², Keisuke Kinoshita², Tomohiro Nakatani², Roland Maas³, and Walter Kellermann³

¹Department VII, Beuth University of Applied Sciences Berlin, Germany ²NTT Communication Science Laboratories, Kyoto, Japan ³Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Erlangen, Germany

Abstract
A novel framework based on conditional emission densities for hidden Markov models (HMMs) is proposed in this contribution to integrate speech enhancement systems with automatic speech recognition systems. In the training phase, the observed feature vectors, corrupted by background noise and reverberation, together with estimates for the interference as provided by the speech enhancement system are used for training joint densities of the observations and the interference estimates. In the decoding phase, the joint densities are transformed to conditional densities of the observed features given the interference estimates. Thus, front end processing can be exploited for obtaining interference estimates, and the estimation errors can be modeled very effectively in a data-driven way. Connected digit recognition experiments in a simulated reverberant environment show the potential of the proposed approach: HMMs with the proposed conditional densities outperform various configurations of conventional HMMs in the logarithmic mel-spectral domain. This is a first step towards using conditional densities for creating synergies between front end and back end.

Index Terms: speech enhancement, robust speech recognition, dereverberation, conditional HMM emission densities, frame-by-frame model adaptation.

1. Introduction
There are many potential automatic speech recognition (ASR) applications where close-talking microphones are not desirable or even unacceptable. In such scenarios, the microphones must be placed at some distance from the users. Therefore, background noise, competing sources, and reverberation are recorded in addition to the desired speech signal. Since the acoustic models used in ASR systems are often trained on clean speech data, these interferences lead to a significant mismatch between the models and the speech signals to be recognized, resulting in increased error rates. To reduce this mismatch, either the acoustic models can be adjusted to the statistical properties of the impaired feature vectors, or the speech signal can be enhanced prior to recognition. For the former case, referred to as back end-based methods, training on corrupted data [1, 2], multi-style training [3–5], and model adaptation approaches [6–8] have been proposed. In the latter case, referred to as front end-based methods, speech enhancement algorithms [9–16] provide a point estimate for the clean speech signal, which can then be used in any conventional recognizer.

Front end-based methods can exploit cues and techniques for the estimation of the clean speech signal that are not available in the back end, e.g., phase information, spectral details, and multi-channel processing. However, the clean speech estimation inevitably introduces errors, which may degrade the recognition accuracy. These estimation errors and any other uncertainties, which are very hard to capture by signal-based models in the front end, can be described very effectively by data-driven techniques in the back end. To exploit both the strengths of the back end-based and front end-based methods, the acoustic models can be adjusted to the output of the speech enhancement system. Many approaches for such an adjustment have been suggested, including HMMs trained on enhanced data [4, 17, 18], HMMs adapted to the enhanced data [19] and uncertainty decoding [20–24]. The uncertainty decoding methods do not only use a point estimate of the clean speech signal, but also information about its reliability.

This paper proposes an alternative framework for a seamless integration of speech enhancement with ASR systems based on conditional HMM emission densities of the observed feature vector given an interference estimate. It aims at avoiding the potential loss of information that may result from transforming the observations in the front end. With the proposed framework, the interference signal is estimated in the front end. Then the back end uses the observed features and the interference estimate directly in the acoustic models. Thus, the front end does not have to transform the observation and the interference estimate into a clean speech estimate. The estimation errors and uncertainties are modeled by the back end in a data-driven way, providing modeling accuracy that could not be achieved in the front end. While conditional densities have been proposed for HMM-based speech recognition before [25, 26], we propose to use them explicitly for combining speech enhancement and recognition systems.

2. Observation model
The microphone signal \( x(t) \) consists of the desired clean speech signal \( s(t) \) and an unwanted interference signal \( u(t) \), i.e.,

\[
x(t) = s(t) + u(t) .
\]  

Typically, the interference signal \( u(t) \) contains background noise and competing speakers, modeled by the signal \( b(t) \), and the late reverberation \( r(t) \) of the desired signal, so that

\[ u(t) = b(t) + r(t) . \]  

Since the early reflections in a reverberant environment usually increase the speech intelligibility [27] and the ASR accuracy [28,29], the desired signal \( s(t) \) may not always be the pure clean speech signal, but may contain early reflections arriving up to 50 ms after the direct sound. For notational convenience, we will use the symbol \( s(t) \) and the term clean speech signal also for the speech signal containing early reflections.

The feature vectors extracted from the signals \( x(t), s(t), \) and \( u(t) \) are denoted as \( x_k, s_k, \) and \( u_k \), respectively, where \( k \)
is the frame index. Thus, the observed feature vector $\mathbf{x}_k$ can be expressed as a function of the desired features $\mathbf{s}_k$ and the interference features $\mathbf{u}_k$ according to

$$\mathbf{x}_k = f (\mathbf{s}_k, \mathbf{u}_k) .$$

(3)

Depending on the features used, approximations for the function $f$ can be given, e.g.,

$$f_m (\mathbf{s}_k, \mathbf{u}_k) \approx \mathbf{s}_k + \mathbf{u}_k ,$$

(4)

$$f_l (\mathbf{s}_k, \mathbf{u}_k) \approx \log (\exp (\mathbf{s}_k) + \exp (\mathbf{u}_k)) ,$$

(5)

in the mel-spectral and logmel-spectral domain, respectively. Equation (5) serves as observation model in the following.

### 3. Proposed approach

The proposed approach is inspired by the following idea: The interference signal $\hat{u}(t)$ is estimated in the front end by signal-based approaches, which are able to exploit the underlying signal generation mechanisms. Then the feature vectors calculated from $x(t)$ and $\hat{u}(t)$ are used directly in the back end. The remaining estimation and approximation errors and all other sources or uncertainty are modeled in a data-driven way in the back end. A straightforward way for directly modeling the statistical properties of $\mathbf{x}_k$ and $\hat{\mathbf{u}}_k$ in the back end is to employ joint HMM emission densities of the observed features and the interference estimate $p(\mathbf{x}_k, \hat{\mathbf{u}}_k) = p(\mathbf{x}_k | \hat{\mathbf{u}}_k) p(\hat{\mathbf{u}}_k)$, which can be written as the product of the conditional density and marginal density. While the marginal density of the interference estimate $p(\hat{\mathbf{u}}_k)$ does not provide any information on the speech signal, it causes the joint density to be strongly dependent on the interference. Thus, the joint density is not robust to non-matched conditions. To reduce the dependence of the acoustic models on the interference, we propose to use only the conditional $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ as HMM emission density.

Note that if an affine observation model as in (4) was fulfilled with equality (rather than just providing an approximation), the proposed conditional density would be equivalent to the density of the enhanced features $\mathbf{s}_k$, i.e.,

$$p(\mathbf{x}_k | \hat{\mathbf{u}}_k) = p(\mathbf{s}_k) .$$

This means that if the approximation errors of (4) are neglected, the proposed conditional densities are equivalent to conventional marginal densities of the enhanced features in the mel-spectral domain due to the linearity of (4). In contrast, for non-linear observation models as in (5), the proposed conditional densities are more specific than the marginal densities of the enhanced features as illustrated in Figure 1 for a simplified example with vector length one. In this example, $\mathbf{x}_k$ is obtained according to (5) from the normally distributed random vectors $\mathbf{s}_k$ and $\mathbf{u}_k$. The joint density $p(\mathbf{x}_k, \hat{\mathbf{u}}_k)$ for the example is shown in Figure 1 a), and Figure 1 b) compares the corresponding conditional densities $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ for different values of $\hat{\mathbf{u}}_k$ to the marginal density $p(\mathbf{s}_k)$ of the clean speech estimate. We can observe the following: 1) The density $p(\mathbf{s}_k)$ is independent of $\hat{\mathbf{u}}_k$. This means that the same density is applied to the enhanced feature vector $\mathbf{s}_k$, regardless of the severity of the interference, i.e., regardless of the uncertainty of the clean speech estimate. 2) The conditional density $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ varies depending on the value of $\hat{\mathbf{u}}_k$. This makes it possible to take to the severity of the interference into account. Therefore, a better discrimination capability can be expected from $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ compared to $p(\mathbf{s}_k)$.

![Figure 1](image1.png)

Figure 1: Illustration of different emission densities for a simplified example with feature vector length one: a) contour of the joint density $p(\mathbf{x}_k, \hat{\mathbf{u}}_k)$, b) comparison of conditional densities $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ for different values of $\hat{\mathbf{u}}_k$ ($\hat{\mathbf{u}}_k = 0$: red, $\hat{\mathbf{u}}_k = 2$: magenta, $\hat{\mathbf{u}}_k = 4$: blue) and the marginal density $p(\mathbf{s}_k)$ of the clean speech estimate (black).

### 3.1. Approach

For non-stationary interference signals $\mathbf{u}(t)$, the interference estimate $\hat{\mathbf{u}}_k$ has to be obtained for each frame $k$ by front end processing, and hence the conditional densities $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ have to be updated in each frame $k$. Therefore, we propose the GMM (Gaussian mixture model)-based approach illustrated in Figure 2, allowing for a frame-by-frame update with relatively low complexity:

1. **Training phase:** For a set of noisy and/or reverberant training signals $x(t)$, calculate the corresponding estimated interference signals $\hat{u}(t)$ and calculate the respective features $\mathbf{x}_k$ and $\hat{\mathbf{u}}_k$. Use these features to train HMMs with joint emission densities $p(\mathbf{x}_k, \hat{\mathbf{u}}_k)$.

2. **Decoding phase:** Given the interference estimate $\hat{\mathbf{u}}_k$ for the current frame $k$, transform the joint density $p(\mathbf{x}_k, \hat{\mathbf{u}}_k)$ to the conditional density $p(\mathbf{x}_k | \hat{\mathbf{u}}_k)$ and evaluate the conditional density for the current observed feature vector $\mathbf{x}_k$.

For training the HMMs with the joint densities $p(\mathbf{x}_k, \hat{\mathbf{u}}_k)$, conventional HMM training approaches and in particular ICEWIND [30] can be used if the stacked vector $\mathbf{y}_k = [\mathbf{x}_k, \hat{\mathbf{u}}_k]^T$ is considered as one feature vector. The resulting joint emission densities $p(\mathbf{x}_k, \hat{\mathbf{u}}_k) = p(\mathbf{y}_k)$ are given by the following GMMs with $M$ components:

$$p(\mathbf{y}_k) = \sum_{m=1}^{M} w_m N (\mathbf{y}_k; \mu_y(m), C_{yy}(m)) ,$$

(6)
where \( m \) is the mixture component index, and \( w_m \),

\[
\begin{align*}
\mu_y(m) &= \left( \mu_y(m) \right), \\
C_y(m) &= \left( C_y(m) \right),
\end{align*}
\] (7)

\[
\begin{align*}
C_y(m) &= \left( C_y(m) \right),
\end{align*}
\] (8)

are the weight, mean vector, and covariance matrix for component \( m \). Once the current interference estimate \( \hat{u}_k \) is available, the joint pdf \( p(y_k) \) can be transformed to the conditional pdf \( p(x_k | \hat{u}_k) \), which is described by the following GMM

\[
p(x_k | \hat{u}_k) = \sum_{m=1}^{M} \nu_m N(x_k; \mu_x(m), C_x(m)),
\] (9)

where the conditional weight \( \nu_m \), the conditional mean vector \( \mu_x(m) \) and the conditional covariance matrix \( C_x(m) \) for mixture component \( m \) can be obtained from the parameters of the joint density and the current interference estimate as

\[
\begin{align*}
\nu_m &= \frac{w_m N(\hat{u}_k; \mu_x(m), C_x(m))}{\sum_{m=1}^{M} w_l N(\hat{u}_k; \mu_x(l), C_x(l))}, \\
\mu_x(m) &= \frac{\mu_x(m) + C_x(m) \left( \mu_x(m) - \mu_x(m) \right)}{C_x(m)}, \\
C_x(m) &= \frac{C_x(m) - C_x(m) C_x(m)}{C_x(m) C_x(m)},
\end{align*}
\] (10)

where the frame index \( k \) has been dropped for brevity.

Due to the frame-by-frame update, the proposed approach appears to be particularly interesting for non-stationary interference, like impulsive noise, competing speakers, and reverberation if powerful interference estimation methods are available in the front end. Furthermore, if the interference estimation exploits previous observations, the frame-by-frame update relays the conditional independence assumption of the HMMs so that the strong long-term relations between frames introduced by reverberation [31, 32] can be modeled very accurately.

### 3.2. Comparison to related approaches

The proposed conditional densities may be compared to vector Taylor series (VTS) model adaptation [8], which is one of the most popular approaches for adjusting HMMs to noisy conditions. VTS obtains the mean vectors and covariance matrices of the noisy HMMs from the clean-speech HMMs, a noise model \( p(u_k) \), and a Taylor approximation of the observation model. While in this way, very little data is required to estimate the noise model, the resulting noisy HMMs inherit all approximation errors of the observation model. In contrast, the proposed conditional densities are based on a point estimate \( \hat{u}_k \) of the interference and model all remaining uncertainties in a data-driven way by the densities \( p(x_k | \hat{u}_k) \). Therefore, the proposed approach requires significantly more data for model estimation, but it is able to capture all inaccuracies of the observation model and the speech enhancement algorithm. For highly non-stationary interferences, like competing speakers and reverberation, the models should be adjusted at each frame [32]. Due to their significantly lower adaptation complexity (assuming block diagonal covariance matrices), the conditional densities are more suitable for a frame-wise model update.

### 4. Experimental evaluation

Experiments with the TI digits corpus [33] are carried out to evaluate the performance of the proposed conditional densities.

![Table 1: Summary of room characteristics: \( T_{60} \) is the reverberation time, \( d \) is the distance between speaker and microphone, \( SRR \) is the signal-to-reverberation ratio.](image)

<table>
<thead>
<tr>
<th>para- meter</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_{60} )</td>
<td>300 ms</td>
<td>700 ms</td>
<td>900 ms</td>
<td>780 ms</td>
<td>600 ms</td>
</tr>
<tr>
<td>( d )</td>
<td>2.0 m</td>
<td>4.1 m</td>
<td>4.0 m</td>
<td>2.0 m</td>
<td>2.0 m</td>
</tr>
<tr>
<td>( SRR )</td>
<td>-4.0 dB</td>
<td>4.0 dB</td>
<td>-4.0 dB</td>
<td>0.5 dB</td>
<td>0.5 dB</td>
</tr>
</tbody>
</table>

Due to the proposed method's particular suitability for modeling reverberation, the following experiments focus on ASR in reverberant environments. Therefore, background noise and competing speakers are neglected so that \( b_k = 0 \) and \( u_k = r_k \).

### 4.1. Experimental setup

The simulations are performed using room impulse responses (RIRs) measured in five different rooms. The most important details of the room characteristics are summarized in Table 1, and further details including a description of the room geometry can be found in [34]. Note that room A is a moderately reverberant environment while rooms B to E are highly reverberant environments. Since the frequency-response of room B has a strong low-pass character, its RIRs are significantly different from the RIRs of rooms A and C, even though the global reverberation characteristics in terms of reverberation time and signal-to-reverberation ratio of room B are quite similar to those of room C. The RIRs for the rooms D and E are taken from the RWCp sound scene database [35].

A subset of the TI digits training set with 4579 connected digit utterances, corresponding to 1.5 hours of speech, is used for training. The stereo version of the dereverberation method presented in [36] is employed for front end processing. To obtain the two-channel reverberant training signals \( x_1(t) = x(t) \) and \( x_2(t) \), the clean signals \( s(t) \) from the training set are convolved with two different RIRs measured in the corresponding room. The reverberant data \( x_k \) are obtained by extracting features from \( x(t) \). Interference signal estimates \( \hat{u}_k(t) \) are obtained from the reverberant signals \( x_1(t) \) and \( x_2(t) \) by [36], and then estimated interference feature vectors \( \hat{u}_k \) are extracted from \( \hat{u}(t) \). As features, logarithmic melspectral (logmelspec) coefficients are used, calculated with the following parameters: sampling frequency 20kHz, frame length 25 ms, frame shift 10 ms, and DFT length 512.

A 16-state word-level HMM with no skips over states is used for each of the 11 digits (‘zero’ to ‘nine’ and ‘oh’). Additionally, a three-state silence model with a backward skip from state three to state one is used. GMMs with three components are used as emission pdfs. The HMMs are trained using the ICEWIND training algorithm [30]. For the proposed approach, GMMs modeling the joint densities according to (6) are trained. Both room-specific HMMs for the rooms A to E and multi-style HMMs for the rooms A, B, C as well as for the rooms D, E are estimated. For comparison, also conventional clean-speech HMMs, conventional clean-speech HMMs with mean adaptation by supervised MLLR [7] using 44 adaptation utterances, conventional HMMs trained on reverberant data, and conventional HMMs trained on dereverberated data are used. While diagonal covariance matrices are used for the conventional HMMs, block diagonal covariance matrices \( C_{y}(m) \) are used for the joint densities of the proposed approach so that each of the four submatrices in (8) is modeled by a diagonal matrix.

A subset of 512 utterances randomly selected from the TI
both the observed feature vectors \( x \) are treated in the same way as for the training data. For the proposed approach, reverberant data and interference estimates are calculated in the log-melspec domain. Each observed feature vector \( x_k \), and the interference estimate \( \hat{u}_k \) are passed to the decoding algorithm, and in each frame, the joint density is transformed to the conditional density according to equations (10) to (12). Then the conditional density is evaluated for the current observed feature vector \( x_k \). To assess the performance both in matched and mismatched conditions, crosstests are performed, where each room-specific HMM and each multi-style HMM is tested against each room.

### 4.2. Complexity

In the following, the complexity of the proposed conditional densities is compared to that of conventional HMMs both with respect to the number of model parameters and in terms of decoding complexity, assuming block diagonal covariance matrices \( C_{yy}^{(m)} \). The joint densities needed for the proposed approach require estimation and storage of the mean vector (7) and covariance matrix (8) for the stacked vector \( y_k \). Neglecting the mixture weights and the transition probabilities, the number of parameters for the proposed approach is about 2.5 times the number of parameters for conventional HMMs. The transformation of the joint densities to the conditional densities in equations (10) to (12) and the evaluation of the conditional densities requires about 2.5 times the number of operations as evaluating the marginal densities in conventional HMMs.

### 4.3. Recognition results

Tables 2–5 show the word accuracies in % for the different approaches, where the results of Tables 2–4 are baseline results achieved with known approaches and Table 5 shows the results of the proposed conditional densities. While the columns correspond to the test data, the rows correspond to the data used for training the corresponding HMMs. A to E stands for rooms specific models, and ABC, DE stands for multi-style models trained on mixed data from rooms A,B,C and rooms D, E, respectively. To facilitate the comparison of the different approaches, average word accuracy values for each HMM, for each test condition and the total averages are also given for each table. To evaluate the robustness of the different approaches with respect to a mismatch between training and test data, each HMM is tested in each condition (crosstests). The matched-condition performance is highlighted with bold numbers.

With a total average word accuracy of 82.36 %, the conditional densities significantly outperform all other investigated approaches, which achieve average word accuracies between 51.70% (HMMs trained on clean data) and 73.95% (HMMs trained on reverberant data). Comparing for each HMM the result of the matched condition to the average result of the corresponding HMM indicates that the conditional densities also achieve the best robustness in mismatched conditions. Furthermore, the multi-style HMMs with conditional densities significantly outperform the multi-style HMMs of all other investigated approaches. In summary, the results in the logmelspec domain are very promising and underline the potential of the proposed approach. Future work should include testing the proposed approach with mel frequency cepstral coefficients (MFCCs) including delta features. Initial tests indicate that the block diagonal covariance matrices, which provide a good approximation in the logmelspec domain, are not sufficient in the cepstral domain.

### 5. Summary and conclusions

The novel framework described in this paper can be considered as a first step towards using conditional HMM emission densities for combining speech enhancement and recognition systems in a synergistic way. To this end, joint densities are estimated from the observed features and the estimated interference features in the training phase. During recognition, these joint densities are transformed to conditional densities for the current interference estimate provided by the speech enhancement system. By avoiding the calculation of a clean speech estimate, a potential loss of information due to distortions of the clean speech is avoided, and all estimation errors are very accurately modeled by the conditional densities in a data-driven way. Experiments using the TI digits corpus in simulated reverberant environments show that the conditional densities outperform conventional HMMs in the logmelspec domain, regardless whether the latter are trained on reverberant or on dereverberated data, thus confirming the potential of the proposed approach. Directions for future research include evaluation of full covariance matrices and alternative matrix structures for more accurately capturing the joint densities in the cepstral domain, investigation of alternative conditional densities, e.g., clean speech estimates given the interference estimate, and ways for direct estimation of the conditional densities from the data instead of converting joint densities to conditional densities.
6. References


