Uncertainty decoding for DNN-HMM hybrid systems based on numerical sampling

Christian Huemmer\textsuperscript{1}, Roland Maas\textsuperscript{1}, Andreas Schwarz\textsuperscript{1}, Ramón Fernández Astudillo\textsuperscript{2}, and Walter Kellermann\textsuperscript{1}

\textsuperscript{1}Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Erlangen, Germany
\textsuperscript{2}Spoken Language Systems Laboratory, INESC-ID-Lisboa, Lisboa, Portugal

\{huemmer,maas,schwarz,wk\}@int.de, ramon@astudillo.com

Abstract

In this article, we propose an uncertainty decoding scheme for DNN-HMM hybrid systems based on numerical sampling. A finite set of samples is drawn from the estimated probability distribution of the acoustic features and subsequently passed through feature transformations/extensions and the deep neural network (DNN). Then, the nonlinearly-transformed feature samples are averaged at the output of the DNN in order to approximate the posterior distribution of the context-dependent Hidden Markov Model (HMM) states. This concept is experimentally verified for the Reverberation robust DNN-HMM hybrid system: The numerical sampling is performed in the logmel-spectra domain, where we estimate the posterior distribution of the acoustic features by combining coherence-based Wiener filtering and uncertainty propagation. The experimental results highlight the good performance of the proposed uncertainty decoding scheme with significantly increased recognition accuracy even for a small number of feature samples.

Index Terms: robust speech recognition, observation uncertainty, numerical sampling, uncertainty decoding

1. Introduction

A few years ago, the state-of-the-art acoustic models in automatic speech recognition (ASR) systems have started to integrate deep neural networks (DNNs) for discriminating between context-dependent phonetic units. For instance, the so-called DNN-HMM hybrid systems include one single DNN which directly maps the observed feature vector sequence to the posterior probabilities of the context-dependent Hidden Markov Model (HMM) states [1]. Although DNN-based acoustic modeling has been shown to outperform classical Gaussian mixture models (GMMs) [2, 3], the ASR performance is still degraded by environmental distortions in adverse acoustic environments [4, 5], which is why a variety of different robustness approaches have been proposed which adapt either the feature vectors or the acoustic-model parameters. The adaptation schemes for DNN-HMM hybrid systems can be classified as follows [5, 6]: First, front-end processing techniques adjust the input features of the DNN. This includes speech enhancement methods to estimate clean from distorted data [7, 8, 9] and feature transformations to reduce irrelevancy while increasing discriminability [10, 11, 12]. Second, back-end processing techniques adjust the DNN parameters to account for speaker or environmental characteristics [13, 14, 15].

As bridge between these two categories, the concept of observation uncertainty considers feature vectors as random variables to account for missing information caused by environmental distortions or estimation errors of front-end algorithms [16, 17]. The underlying probabilistic distortion model can be combined with the acoustic model by applying uncertainty decoding, which has been widely used for GMM-HMM ASR systems [18, 19]. For DNN-based acoustic modeling, it has been shown that random feature variables can be propagated through a nonlinearity (i.e. the probability distribution at the output of the nonlinearity is approximated) by applying piece-wise approximation techniques or the unscented transform [20].

In this article, we follow the concept of observation uncertainty based on a probabilistic distortion model in the logmel-spectra domain. As main innovation, we propose an uncertainty decoding scheme based on numerical sampling which can be summarized in the following way: A finite set of samples is drawn from the estimated probability density function (PDF) of the uncertain features and subsequently passed through feature transformations/extensions and the DNN. Then, these nonlinearly transformed feature samples are averaged to approximate the posterior probabilities of the context-dependent HMM states. The proposed uncertainty decoding scheme is experimentally verified using a two-channel DNN-HMM hybrid system: Both microphone signals are transformed into the short-time Fourier transform (STFT) domain and processed by a coherence-based Wiener filter [21]. Next, the STFT uncertainty propagation (STFT-UP) approach [22] is applied to propagate the posterior distribution of the Wiener filter output into the lower-dimensional logmel-spectra domain. From this distribution, we draw a finite set of samples, pass them through feature transformations (normalization, creation of delta parameters and context extension) as well as the DNN and approximate the posterior probability by averaging over the nonlinearly-transformed samples at the output of the DNN. This reverberation-robust DNN-HMM hybrid system is evaluated on the Reverberation challenge task for both clean and multi-condition DNN training. It is shown that the recognition accuracy in terms of word error rate (WER) scores improves even for a small number of feature samples.

This article is organized as follows: In Section 2, we review the structure of a DNN-HMM hybrid system and propose numerical sampling for uncertainty decoding. This is followed by the description of the reverberation-robust ASR system in Section 3. Finally, experimental results for the Reverberation challenge task are shown in Section 4.
2. Numerical sampling for uncertainty decoding in DNN-HMM hybrid systems

2.1. DNN-HMM hybrid system with speech enhancement in the STFT domain

In most robust ASR systems, the front-end approaches for compensating environmental distortions can be classified as speech enhancement techniques (to estimate clean from distorted speech) or feature transformations (to reduce irrelevancy or enhance discriminability). As illustrated in Fig. 1(a), we focus on a DNN-HMM hybrid system with speech enhancement in the STFT domain, where the input vector

$$y_n = [y_{1,n}, y_{1,n}, ..., y_{M,n}]^T$$

(1)

with complex-valued coefficients $y_{v,n}$ (at frequency bin $v = 1, ..., M$ and time instant $n$) is enhanced to remove the distortions caused by the acoustic environment. The resulting point estimate of the clean STFT-domain signal is passed through a feature extraction step to create the lower-dimensional feature vector $\hat{z}_n$. In case of ideal speech enhancement, this estimate $\hat{z}_n$ equals the clean feature vector $z_n$, so that the nonlinear function $f_j(\cdot)$ (modeling feature transformations/extensions and DNN) produces the posterior likelihood

$$c_j = p(s_j|z_n) = f_j(z_n)$$

(2)

of the $j$th context-dependent HMM state $s_j$ at the respective output of the DNN. For recognition, the posterior likelihood is rescaled [23]:

$$\log(p(z_n|s_j)) \sim \log(p(s_j|z_n)) - \log(p(s_j))$$

(3)

with the prior probability $p(s_j)$ estimated in the training phase.

Figure 1: DNN-HMM hybrid system with speech enhancement in the STFT domain. The point estimate $\hat{z}_n$ is assumed to have no residual error in (a) and to be uncertain in (b), where we illustrate the proposed uncertainty decoding scheme.

2.2. Observation uncertainty techniques for robust ASR

The fundamental idea of observation uncertainty is to account for missing information by considering a probabilistic model which describes the mathematical relation between corrupted and undistorted feature vectors. Assuming the point estimate $\hat{z}_n$ to be a distorted version of the clean feature vector $z_n$ and the nonlinear mapping $f_j(\cdot)$ to be known, we can rewrite the posterior distribution of HMM state $s_j$ to

$$p(s_j|\hat{z}_n) = \frac{p(\hat{z}_n|s_j)p(s_j)}{p(\hat{z}_n)}$$

(4)

based on the assumption that the distorted vector $\hat{z}_n$ is conditionally independent from $s_j$ given $z_n$. As a consequence of this, the posterior distribution $p(s_j|\hat{z}_n)$ equals the mean value $E\{f_j(z_n)\}$ at the output of the DNN. For instance, the distortions of the point estimate $\hat{z}_n$ may be modeled by a normally distributed additive uncertainty $b_n$:

$$\hat{z}_n = z_n + b_n, \quad b_n \sim \mathcal{N}(0, \hat{U}_n),$$

(5)

with zero mean and estimated covariance matrix $\hat{U}_n$. This probabilistic model leads to the following mean value at the output of the DNN:

$$E\{c_j|\hat{z}_n\} = E\{f_j(z_n)|\hat{z}_n\} = E\{f_j(\hat{z}_n - b_n)\},$$

(6)

which can be approximated by applying the concept of uncertainty propagation [20, 24]. The random variable $z_n$ is propagated through the nonlinear function $f_j(\cdot)$ by layer-wise estimating the PDF of the uncertain features using piece-wise approximation techniques [25] or the unscented transform [26]. In the following, we propose a numerical sampling scheme to approximate $E\{c_j|\hat{z}_n\}$ (without explicitly propagating the PDF of the uncertain features) by averaging non-linear transformed samples at the output of the DNN.

2.3. Numerical sampling for uncertainty decoding

To estimate the posterior likelihood $p(s_j|\hat{z}_n)$ in (4), we approximate the mean value $E\{c_j|\hat{z}_n\}$ at the output of the DNN by drawing $L$ samples $z_n^{(l)}$ from the probability distribution of $z_n$ and averaging over the samples after they have been passed through the nonlinear function $f_j(\cdot)$. This concept is illustrated in Fig 1(b) for the exemplary model of (5), where we draw $L$ samples $b_n^{(l)}$ from the PDF of $b_n$:

$$p(s_j|\hat{z}_n) = E\{c_j|\hat{z}_n\} \approx \frac{1}{L} \sum_{l=1}^{L} f_j(z_n^{(l)})$$

(7)

Note that this numerical approximation is independent of the structure of the nonlinearity $f_j(\cdot)$ and not restricted to a specific feature extraction technique.
3. Reverberation-robust DNN-HMM hybrid system

In this part, we describe a DNN-HMM hybrid system which combines coherence-based dereverberation and noise suppression in the STFT domain with the concept of observation uncertainty in the logmelspec domain. This reverberation-robust ASR system is illustrated in Fig. 2 and will be later employed (in Section 4) for the experimental verification of the proposed numerical sampling scheme.

3.1. Speech enhancement in the STFT domain

As illustrated in Fig. 2, we realize a coherence-based Wiener filter by estimating the time- and frequency-dependent coherent-to-diffuse power ratio $CDR_{\nu,n}$ following (12) in [27]. In addition, the magnitudes of both microphone signals are averaged in the STFT domain (see [21] for more details) to determine the single-channel input vector $y_n$. The complex Gaussian posterior PDF at the output of the single-channel Wiener filter

$$p(x_n|y_n) = \mathcal{N}(\hat{x}_n, \bar{V}_n)$$

is calculated (substituting the time-frequency dependent signal-to-noise ratio $CDR_{\nu,n}$, e.g. in (11) and (12) of [28]) using

$$\hat{x}_n = \text{diag}\{g_{1,n}, ..., g_{M,n}\}y_n,$$

as estimate of the mean vector $x_n$ and

$$\bar{V}_n = \text{diag}\{\hat{v}_{1,n}, ..., \hat{v}_{M,n}\},$$

as estimate for the diagonal covariance matrix $\bar{V}_n$ $(M = 257$ for our experiments). Here, $\text{diag}\{\cdot\}$ creates a diagonal matrix with the input arguments as diagonal elements and $\text{var}\{\cdot\}$ denotes the variance of an input variable.

3.2. Observation uncertainty and numerical sampling

The Gaussian posterior distribution $p(x_n|y_n)$ in (8) is propagated (at each time step $\nu$) from the output of the Wiener filter into the logmelspec domain using the STFT-UP approach derived in [22]. Thus, we estimate the Gaussian PDF

$$p(x_n|y_n) = \mathcal{N}(\hat{z}_n, \bar{U}_n),$$

where $z_n$ is a real-valued vector of length $24$ and $\bar{U}_n$ a diagonal covariance matrix of size $24 \times 24$. Following the concept in [28], we assume $p(z_n|y_n)$ to model the observation uncertainty and draw $L$ samples $z_n^{(l)}$ from this estimated probability distribution.

3.3. Feature transformation and DNN

As shown in Fig. 1(a), the nonlinear function $f_\nu(\cdot)$ is modeling the feature transformation/extension as well as the DNN. The former can be summarized for our implementation as follows:

- Per-utterance mean and variance normalization.
- Dynamic extension: Delta and acceleration coefficients.
- Context extension using ±5 frame splicing (the size of the context window has been manually optimized).

These extended logmelspec feature vectors of length 792 are used as input of the DNN which is characterized by the following topology:

- 6 hidden layers, each with 2048 nodes and sigmoid activation functions.
- Output layer with softmax nonlinearity and 3463 elements (number of context-dependent HMM states).

Note that this complex structure of the nonlinear function $f_\nu(\cdot)$ emphasizes the generality and capability of the proposed numerical sampling scheme.

4. Experiments

We employ the Kaldi toolkit [29] as ASR back-end system using the WSJ0 trigram 5k language model provided by the REVERB challenge [30]. As first step, we train a GMM-HMM baseline system on the clean WSJCAM0 Cambridge Read News REVERB corpus [31] with feature extraction following the Type-I creation in [32] (which is state-of-the-art in the Kaldi recipe): The extraction of 13 mel-frequency cepstral coefficients (MFCCs) is followed by linear discriminant analysis (with splicing optimized to ±4 input frames), maximum likelihood linear transform and feature-space maximum likelihood linear regression. The state-frame alignment of the trained GMM-HMM baseline system is employed for training the DNN on extended logmelspec feature vectors (see Section 3.3 for more details on the feature extraction): A generative pretraining using the contrastive divergence algorithm (on restricted Boltzmann machines) is followed by discriminative fine-tuning using the mini-batch stochastic gradient descent approach (based on the cross-entropy criterion) [1]. We consider two setups of DNN training by using the clean and the multi-condition training sets (each of 7861 utterances) provided by the REVERB challenge [30].

The evaluation of the reverberation-robust DNN-HMM hybrid system (shown in Fig. 2) is realized using the two-channel task of the REVERB challenge [30]: The evaluation test set consists of ~5000 environmentally-distorted utterances and is split into two categories: First, the utterances of...
Table 1: WER scores for the REVERB challenge evaluation test set with DNN trained on clean data.

<table>
<thead>
<tr>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{60} \approx 0.25$ s</td>
<td>$T_{60} \approx 0.75$ s</td>
</tr>
<tr>
<td>No sampling</td>
<td></td>
</tr>
<tr>
<td>Near</td>
<td>Far</td>
</tr>
<tr>
<td>9.57</td>
<td>29.51</td>
</tr>
<tr>
<td>L = 10</td>
<td></td>
</tr>
<tr>
<td>L = 20</td>
<td></td>
</tr>
<tr>
<td>L = 30</td>
<td></td>
</tr>
<tr>
<td>L = 50</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: WER scores for the REVERB challenge evaluation test set with DNN trained on multi-condition data.

<table>
<thead>
<tr>
<th>SimData</th>
<th>RealData</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{60} \approx 0.25$ s</td>
<td>$T_{60} \approx 0.75$ s</td>
</tr>
<tr>
<td>No sampling</td>
<td></td>
</tr>
<tr>
<td>Near</td>
<td>Far</td>
</tr>
<tr>
<td>5.81</td>
<td>10.25</td>
</tr>
<tr>
<td>L = 10</td>
<td></td>
</tr>
<tr>
<td>L = 20</td>
<td></td>
</tr>
<tr>
<td>L = 30</td>
<td></td>
</tr>
<tr>
<td>L = 50</td>
<td></td>
</tr>
</tbody>
</table>

The clean WSJCAM0 Cambridge Read News REVERB corpus are artificially corrupted (“SimData”) using measured impulse responses ($T_{60} \approx 0.25$, 0.5 s and 0.7 s), recorded noise sequences (added to the microphones signals with a signal-to-noise ratio of 20 dB) and source-microphone spacings of 0.5 m (“Near”) and 2 m (“Far”). Second, multichannel recordings (“RealData”) in a reverberant ($T_{60} \approx 0.7$ s) and noisy environment are considered with source-microphone spacings of 1 m (“Near”) and 2.5 m (“Far”). In both cases, two microphones with a spacing of 8 cm are selected out of an 8-channel circular microphone array to realize the two-channel task which is evaluated in the following.

Table 1 shows the WER scores for the evaluation test set, where the DNN has been trained on clean data. We observe a scenario-independent improvement of the recognition accuracy for artificially corrupted evaluation data as well as for real recordings. Note that even a small number of $L = 10$ samples is sufficient to achieve a significant performance gain compared to the conventional DNN-HMM hybrid system without uncertainty decoding (“No sampling”). In summary, the results in Table 1 emphasize the consistent performance gain achieved by the proposed numerical sampling scheme consistently reduces the WER scores of the DNN-HMM hybrid system if trained on clean (Table 1) or multi-condition (Table 2) data. In the practically relevant case of multi-condition DNN training, we gain significant improvements especially in the challenging task of large speaker-microphone distances.

5. Conclusions

We proposed an uncertainty decoding scheme for DNN-HMM hybrid systems based on numerical sampling. The missing information caused by environmental distortions or front-end estimation errors is reflected by modeling acoustic features as random variables. From the probability distribution of the features, we draw a finite set of samples which are nonlinearly transformed and subsequently averaged at the output of the DNN to approximate the posterior probability of the context-dependent HMM states. This numerical sampling scheme is experimentally verified for the REVERB challenge task based on a reverberation-robust DNN-HMM hybrid system, where we combine coherence-based Wiener filtering and uncertainty propagation to estimate the probability distribution of the acoustic features in the logmelspec domain. By drawing a finite number of samples from this estimated PDF, the proposed uncertainty decoding scheme is shown to increase the recognition accuracy for artificially corrupted evaluation data as well as for real recordings.

6. Acknowledgements

The authors would like to thank the Deutsche Forschungsgemeinschaft (contract number KE 890/4-2) and the Foundation for Science and Technology (grant number SFRH/BPD/68428/2010) for supporting this work.
7. References


