On the Estimation and Use of Feature Reliability Information for Noise Robust Speech Recognition

Volker Leutnant and Reinhold Haeb-Umbach
University of Paderborn, Germany, Email: {leutnant,haeb}@nt.uni-paderborn.de

Abstract

In this paper we present an Uncertainty Decoding rule which exploits feature reliability information and inter-frame correlation for noise robust speech recognition. The reliability information can be obtained either from conditional Bayesian estimation, where speech and noise feature vectors are tracked jointly, or by augmenting conventional point estimation methods with heuristics about the estimator's reliability. Experimental results on the AURORA2 database demonstrate on the one hand that Uncertainty Decoding improves recognition performance, while on the other hand it is seen that the severe approximations needed to arrive at computationally tractable solutions have their noticable impact on recognition performance.

Introduction

Today's speech recognizers are notorious for their lack of robustness towards a mismatch between training and testing environments, be it due to additive or convolutional noise, or any other type of distortion. While a plethora of approaches has been proposed to mitigate the detrimental effect of environmental noise, a systematic treatment in a rigorous probabilistic framework, which in some sense reconciles front-end to back-end techniques, has gained attention only fairly recently under the name Uncertainty Decoding (UD), see e.g. [1–6]. This term has been phrased for a class of robustness enhancing algorithms in automatic speech recognition that replace point estimates of the clean speech signal and plug-in rules by posterior densities and optimal decision rules. In doing so, the imperfections of the enhancement stage are accounted for in the recognizer by placing more emphasis on those parts of the signal which have been restored more reliably. It is therefore also able to cope with non-stationary distortions, which pose a major challenge to many of the more conventional approaches to robust speech recognition.

In this paper we first rederive an Uncertainty Decoding rule which takes advantage of a relaxed conditional independence assumption and which we have previously introduced in [6], however with a somewhat more involved derivation. Then we take a conditional Bayesian estimation point of view and illustrate how the posterior of the clean speech feature vector, given the observed corrupted feature vectors can be estimated. Emphasis is placed on the importance of the variance of the posterior, which, in the case of Gaussians, equals the estimation error variance. Next, a heuristics is introduced how this variance can be estimated even if the aforementioned posterior is not available. As of today, however, this approach is barely able to compete with conventional sophisticated noise robustness enhancing techniques, such as the ETSI Advanced Front End (AFE) [7]. However, any speech feature enhancement scheme can be improved by UD, as is demonstrated with an oracle experiment employing the AFE.

Uncertainty Decoding for ASR

Given a sequence of feature vectors $\mathbf{x}_1^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of length T extracted from an utterance, the classification task amounts to finding that sequence of words $\hat{\mathbf{W}}$ from a given vocabulary which maximizes the joint probability $p(\mathbf{W}, \mathbf{x}_1^T)$ or, equivalently,

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{x}_1^T|\mathbf{W}) \cdot P(\mathbf{W}). \tag{1}$$

The a priori probability of the word sequence, $P(\mathbf{W})$, is provided by the language model, while the acoustic model is concerned with computing $p(\mathbf{x}_1^T|\mathbf{W})$. In a HMM-based speech recognizer this is accomplished by introducing the sequence of hidden states $q_1^T = (q_1, \ldots, q_T)$ underlying the sequence of observations:

$$p(\mathbf{x}_1^T|\mathbf{W}) \quad = \quad \sum_{\{q_1^T\}} p(\mathbf{x}_1^T|q_1^T) \cdot P(q_1^T|\mathbf{W}) \tag{2}$$

where the summation is carried out over all state sequences within $\mathbf{W}$.

In many practical situations there exists a mismatch between training and testing conditions. This can be expressed by the fact that the sequence of test features $\mathbf{x}_1^T$, which are representative of the training conditions, and which are denoted as "clean" features in the following, is not observable. A corrupted version $\mathbf{y}_1^T$ is observed instead, where the corruption is caused by e.g. acoustical environmental noise.

The recognition task is stated now as finding the most probable word sequence given $\mathbf{y}_1^T$:

$$\hat{\mathbf{W}} \quad = \quad \arg\max_{\mathbf{W}} p(\mathbf{y}_1^T|\mathbf{W}) \cdot P(\mathbf{W}). \tag{3}$$

Taking $\mathbf{y}_1^T$ as if they were the "clean", uncorrupted data, i.e. interpreting $\mathbf{y}_1^T$ as an estimate of $\mathbf{x}_1^T$ to be used in (1) results in the well-known poor performance of speech recognition in the presence of a mismatch between training and testing conditions.

Basically two classes of approaches for the task of recognizing corrupted speech exist. The first class comprises back-end methods. Starting from eq. (3), the

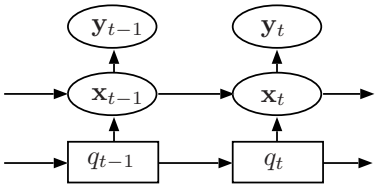

Figure 1: Bayesian network considering temporal correlation between features.

likelihood of $p(\mathbf{y}_1^T|\mathbf{W})$ is computed based on a modified or adapted model (e.g. by Maximum Likelihood Linear Regression, MLLR) for $p(\mathbf{x}_1^T|\mathbf{W})$. The second class is made up of front-end approaches which use the model for clean speech $p(\mathbf{x}_1^T|\mathbf{W})$ and attempt to estimate the clean speech feature sequence $\mathbf{x}_1^T$ from noisy input data $\mathbf{y}_1^T$. An example of the latter is the ETSI Advanced Front End [7].

In an optimal decoding rule, however, the reliablility of the clean feature estimates would have to be taken into account. To incorporate the reliablity information into the speech recognition system while still using the acoustic model trained under uncorrupted conditions, the unobservable clean feature sequence $\mathbf{x}_1^T$ is introduced as a hidden variable:

$$p(\mathbf{y}_1^T|\mathbf{W}) = \int_{\{\mathbf{x}_1^T\}} p(\mathbf{y}_1^T|\mathbf{x}_1^T)p(\mathbf{x}_1^T|\mathbf{W})d\mathbf{x}_1^T, \tag{4}$$

The marginalization has to be carried out over all possible clean feature sequences of length T, indicated by $\{\mathbf{x}_1^T\}$.

The Bayesian network of Fig. 1 depicts the assumed statistical dependencies among the random variables under consideration. Note that the observed feature vectors are statistically independent of the HMM states, if the clean features are given; and further note that a direct statistical dependency among successive clean feature vectors is assumed, thus relaxing the well-known conditional independence assumption.

Introducing the HMM state sequence q_1^T, replacing any condition on $\mathbf{W}$, the search now has to compute

$$p(\mathbf{y}_1^T|q_1^T) = \int_{\{\mathbf{x}_1^T\}} p(\mathbf{y}_1^T|\mathbf{x}_1^T)p(\mathbf{x}_1^T|q_1^T)d\mathbf{x}_1^T, \tag{5}$$

where

$$p(\mathbf{x}_1^T|q_1^T) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_1^T). \tag{6}$$

Using (6) in (5) we obtain

$$\begin{aligned} &p(\mathbf{y}_1^T|q_1^T) \\ &= \int_{\{\mathbf{x}_1^T\}} p(\mathbf{y}_1^T|\mathbf{x}_1^T) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_1^T)d\mathbf{x}_1^T \\ &= C_1 \int_{\{\mathbf{x}_1^T\}} \frac{p(\mathbf{x}_1^T|\mathbf{y}_1^T)}{p(\mathbf{x}_1^T)} \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_1^T)d\mathbf{x}_1^T \\ &= C_1 \int_{\{\mathbf{x}_1^T\}} \prod_{t=1}^{T} \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_1^T)}{p(\mathbf{x}_t|\mathbf{x}_{t-1})} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_1^T)d\mathbf{x}_1^T \end{aligned} \tag{7}$$

where C_1 is a constant. In order to further simplify (7) we first assume that $\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_1^T) \approx \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_t)$, an approximation that can reasonably be justified by the fact that the dependency between $\mathbf{x}_t$ and q_t is stronger than between $\mathbf{x}_t$ and past/future HMM states once $\mathbf{x}_{t-1}$ is given. Disregarding also the dependency on $\mathbf{x}_{t-1}$ allows us to interchange the order of integral and product:

$$\begin{aligned} p(\mathbf{y}_1^T|q_1^T) &\approx C_1 \int_{\{\mathbf{x}_1^T\}} \prod_{t=1}^{T} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t)d\mathbf{x}_1^T \\ &= C_1 \prod_{t=1}^{T} \int_{\{\mathbf{x}_t\}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t)d\mathbf{x}_t, \end{aligned} \tag{8}$$

i.e. the integral over the feature space $\{\mathbf{x}_1^T\}$ can now be evaluated as the product of T integrals, each over the space of an individual feature vector $\mathbf{x}_t$, $t = 1, \ldots, T$.

A comparison of (8) with the computation of the acoustic model likelihood in a conventional speech recognizer

$$p(\mathbf{x}_1^T|q_1^T) = \prod_{t=1}^{T} p(\mathbf{x}_t|q_t) \tag{9}$$

reveals that the optimization problem (3) differs from the optimization problem for clean speech in the classical HMM framework (1) solely with respect to the likelihood:

$$p_{\mathrm{LH}}(\mathbf{y}_1^T|q_t) = \int_{\{\mathbf{x}_t\}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)p(\mathbf{x}_t|q_t)}{p(\mathbf{x}_t)} d\mathbf{x}_t. \tag{10}$$

This variant of Uncertainty Decoding has been first published in [6], where however a more involved derivation was given. In UD, instead of evaluating the likelihood $p(\mathbf{x}_t|q_t)$ for a point estimate $E[\mathbf{x}_t|\mathbf{y}_1^T]$, i.e. setting $p(\mathbf{x}_t|\mathbf{y}_1^T) = \delta(\mathbf{x}_t - E[\mathbf{x}_t|\mathbf{y}_1^T])$ in eq. (10), the entire posterior, i.e. also its variance, is taken into account.

UD Based on Posterior Estimation

The key element of the Uncertainty Decoding rule is the clean feature posterior $p(\mathbf{x}_t|\mathbf{y}_1^T)$, and the success of UD crucially depends on how well it can be determined.

Knowledge of the posterior density enables one to compute an optimal estimate with respect to any criterion. For example the minimum mean squared error (MMSE) estimate equals its mean. Furthermore, a measure of accuracy of the estimate can be obtained form the variance of the posterior. In the Gaussian case the variance of the posterior is even identical to the estimation error variance.

Conceptually, the posterior can be estimated recursively via the following equations[1]:

$$p(\mathbf{x}_t|\mathbf{y}_1^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_1^{t-1})d\mathbf{x}_{t-1} \tag{11}$$

$$p(\mathbf{x}_t|\mathbf{y}_1^{t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_1^{t-1})}{\int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_1^{t-1})d\mathbf{x}_t} \tag{12}$$

[1]Here we restrict ourselves to causal processing, i.e. rather than computing $p(\mathbf{x}_t|\mathbf{y}_1^T)$ we compute $p(\mathbf{x}_t|\mathbf{y}_1^t)$.

However, in practice these equations are often computationally intractable. In the case of speech feature enhancement the key components have been chosen as follows to arrive at a realizable solution:

a) For the dynamical model of the clean speech feature trajectory, $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, switching linear dynamical models (SLDM) have been employed [8].

b) Different approximations of the highly nonlinear observation model have been proposed [9] to arrive at a tractable $p(\mathbf{y}_t|\mathbf{x}_t)$.

c) Extended or Unscented Kalman Filtering as well as Particle Filtering approaches have been used as inference algorithm to compute $p(\mathbf{x}_t|\mathbf{y}_1^T)$. If SLDMs are employed to model the speech dynamics they have to be embedded in an algorithm which infers the switching variable, such as the Generalized Pseudo-Bayesian Estimator (GPB) of first or second order and the Interacting Multiple Model algorithm [10].

If the posterior and the other densities in (10) are approximated by Gaussians, the integral can be solved analytically and the application of the UD rule becomes computationally tractable [6].

UD based on a Variance Compensation Scheme

While the above is a theoretically well motivated approach, in practice many difficulties arise due to the complicated dynamics of clean speech and the extremely nonlinear observation model.

An alternative, which less radically breaks with conventional approaches to noise-robust speech recognition, is to adjoin the point estimate $\hat{\mathbf{x}}_t(\mathbf{y}_t)$ of a speech feature enhancement scheme with an estimate of its estimation error variance by some heuristics, see e.g. [11].

The observation likelihood is then given by

$$p_{\text{LH}}(\mathbf{y}_t|q_t=n) = C_2 \sum_{m=1}^{M} c_{n,m}\,\mathcal{N}\left(\hat{\mathbf{x}}_t(\mathbf{y}_t); \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m}+\boldsymbol{\Sigma}_t\right) \quad (13)$$

Here, $c_{n,m}$, $\boldsymbol{\mu}_{n,m}$ and $\boldsymbol{\Sigma}_{n,m}$ are the weight, mean and (diagonal) covariance matrix of the m-Gaussian of the (clean speech) acoustic observation model of the n-th HMM state, and $\boldsymbol{\Sigma}_t = E[(\mathbf{x}_t - \hat{\mathbf{x}}_t(\mathbf{y}_t))(\mathbf{x}_t - \hat{\mathbf{x}}_t(\mathbf{y}_t))']$ is the estimation error covariance, while C_2 is again a constant.

Let $\sigma_{t,i}^2$ be the (i,i)-th element of the (diagonal) covariance $\boldsymbol{\Sigma}_t$. In [11] it has been proposed that the estimation error variance $\sigma_{t,i}^2$ is proportional to the amount of noise reduction:

$$\sigma_{t,i}^2 = \alpha_i (y_{t,i} - \hat{x}_{t,i})^2. \quad (14)$$

Here, $y_{t,i}$ and $\hat{x}_{t,i}$ are the i-th component of the noisy speech and estimated clean speech feature vector, respectively, and α_i is a scaling factor.

The assumption underlying this heuristics is that the speech enhancement introduces more distortions when a great amount of noise is removed. As the noise is unknown it is replaced by its estimate $\mathbf{y}_t - \hat{\mathbf{x}}_t$. This so-called *dynamic variance compensation* led to significant error rate reduction in the context of recognition of reverberant speech.

The scaling factor α can either be estimated on adaptation data using the EM algorithm [11] or set to unity for simplicity.

Experimental Results

The experiments were performed on test set A of the AURORA2 database. The AURORA2 database is a subset of the TI Digits recognition task to which noise was artificially added at different SNR levels. The test set consists of four different noise types. Training has been carried out on clean speech. We modified the ETSI standard front-end extraction in the same manner as in [8] by replacing the energy feature with c_0 and using the squared power spectral density rather than the spectral magnitude as the input of the Mel-frequency filter-bank.

Oracle Experiment

To show how speech recognition systems can benefit from incoorporating uncertainty information, an oracle experiment is carried out, first. Given an estimate $\hat{\mathbf{x}}_t^{AFE}$ of the clean speech feature obtained from the ETSI advanced front-end (AFE) [7], the posterior $p(\mathbf{x}_t|\mathbf{y}_1^t)$ will be modelled as a Gaussian centered at $\hat{\mathbf{x}}_t^{AFE}$, the estimate of the clean speech feature vector provided by the AFE. The covariance $E[(\mathbf{x}_t - \hat{\mathbf{x}}_t^{AFE})(\mathbf{x}_t - \hat{\mathbf{x}}_t^{AFE})']$ of the posterior is calculated as the empirical covariance of the deviation of the estimated from the true clean speech features. Thus, perfect knowledge of the true feature is assumed. To take into account that such perfect knowledge will never be available, the true error is averaged over the sliding window of length $2L+1$ ($L = 5$) centered at the current frame, simulating an estimator with limited accuracy. While the word accuracy obtained by the AFE is 88.45%, Table 1 shows that Uncertainty Decoding with the oracle covariance improves performance to 91.31%.

SNR	Sub.	Bab.	Car	Exh.	AVG
20dB	98.77	98.34	99.22	99.01	98.84
15dB	97.70	97.34	98.57	97.99	97.90
10dB	94.96	94.86	97.64	96.05	95.88
5dB	89.10	90.51	93.20	88.43	90.31
0dB	74.06	70.98	77.60	71.83	73.62
AVG	90.92	90.41	93.25	90.66	91.31

Table 1: AFE with UD using the oracle covariance calculated over a sliding window of size $2L+1$ ($L=5$)

Comparison of Variance Estimates

Next we compare different estimates of the estimation error variance with respect to their effect on the recognition accuracy. The baseline system is similar to the one described in [8], differing however in the fact that we jointly track both the clean speech and noise feature

vector [12]. Further characteristics of the system are:

- An SLDM with 16 individual linear dynamical models to describe the clean speech feature trajectories.
- An observation model according to $\mathbf{y}_t = \mathbf{x}_t + \log(1 - e^{\mathbf{n}_t - \mathbf{x}_t})$, where the SNR variable $\mathbf{r}_t := \mathbf{x}_t - \mathbf{n}_t$ is improved by iterated Taylor Series approximation [9].
- Inference is carried out using the GPB algorithm of order one.

In addition, Cepstral Mean Normalization (CMN) was applied to the enhanced features. If only the mean of the posterior is used in the recognizer, the baseline recognition accuracy is 83.7%.

The inference algorithm estimates the variance of the clean speech posterior in addition to its mean. It turned out that the variance is overestimated and needs to be upper bounded by 0.05 times the variance of the prior $p(\mathbf{x}_t)$ [3]. Recognition results are given in Table 2.

SNR	Sub.	Bab.	Car	Exh.	AVG
20dB	98.77	97.43	99.25	97.47	98.23
15dB	96.93	94.80	98.69	95.50	96.48
10dB	93.77	89.42	96.42	89.79	92.35
5dB	83.73	76.09	87.50	77.66	81.25
0dB	58.89	43.02	63.88	53.32	54.78
AVG	86.42	80.15	89.15	82.75	84.62

Table 2: Uncertainty Decoding based on clean speech posterior estimation.

In Table 3 the effect of using the variance estimate according to (14) can be studied. The recognition accuracy is improved to 85.06%.

SNR	Sub.	Bab.	Car	Exh.	AVG
20dB	98.68	98.55	99.19	98.55	98.74
15dB	97.42	97.04	98.51	96.64	97.40
10dB	93.71	93.20	96.78	91.92	93.90
5dB	82.35	81.59	85.74	80.38	82.52
0dB	54.62	49.61	54.46	52.18	52.72
AVG	85.36	84.00	86.94	83.93	85.06

Table 3: Uncertainty Decoding with heuristics (14) to obtain estimation error variance.

Conclusions

In this paper we have first motivated Uncertainty Decoding by deriving the decoding rule from an optimization problem of recognizing speech in the presence of corrupted features. While the estimation of the clean speech feature posterior in a Bayesian framework and its use in Uncertainty Decoding is a theoretically appealing approach, the achieved recognition performance is often worse than that of conventional approaches such as the ETSI Advanced Front End. This is probably due to the severe approximations that are needed to arrive at a computationally tractable solution. However, the potential of using feature uncertainty as demonstrated by the oracle experiment and the power of conditional Bayesian estimation known from many other applications call for further exploration of this approach to robust speech recognition.

Acknowledgement

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) unter contract no. Ha3455/6-1.

References

[1] A. Morris, J. Barker, and H. Bourland, "From missing data to maybe useful data: Soft data modelling for noise robust asr," in *WISP*, vol. 6. IEEE, 2001, pp. 153–164.

[2] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with splice for noise robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, A. Acero, Ed., vol. 1, 2002, pp. I–57–I–60 vol.1.

[3] H. Liao and M. Gales, "Issues with uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, vol. 1. ISCA, 2006, pp. 1627–1630.

[4] V. Stouten, H. Van Hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust asr," *Speech Communication*, vol. 48, no. 11, pp. 1616–1624, 2006.

[5] J. Deng, M. Bouchard, and T. Yeap, "Feature enhancement for noisy speech recognition with a time-variant linear predictive hmm structure," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 5, pp. 891–899, 2008.

[6] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 1047–1060, 2008.

[7] ETSI ES 202 050, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI Std. ES 202 050, Rev. 1.1.5, January 2007.

[8] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, A. Acero, Ed., vol. 1, 2004, pp. I–953–6 vol.1.

[9] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *Proc. EUROSPEECH*, vol. 1. ISCA, 2003, pp. 681–684.

[10] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. John Wiley and Sons, Inc., 2001.

[11] M. Delcroix, T. Nakatani, and S. Watanabe, "Combined static and dynamic variance adaptation for efficient interconnection of speech enhancement pre-processor with speech recognizer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 4073–4076.

[12] S. Windmann and R. Haeb-Umbach, "Modelling the dynamics of speech and noise for speech feature enhancement in asr," in *Proc. ICASSP*, vol. 1. IEEE, 2008, pp. 4409–4412.