Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 30–46

www.elsevier.com/locate/csl

Uncertainty weighting and propagation in DNN–HMM-based speech recognition☆

José Novoa[a], Josué Fredes[a], Víctor Poblete[b], Néstor Becerra Yoma[a,]*

[a] *Speech Processing and Transmission Laboratory, Electrical Engineering Department, University of Chile, Santiago, Chile*
[b] *Institute of Acoustics, Universidad Austral de Chile, Valdivia, Chile*

Received 7 September 2016; received in revised form 14 June 2017; accepted 16 June 2017
Available online 30 June 2017

Abstract

In this paper an uncertainty weighting scheme for DNN–HMM-based speech recognition is proposed to increase discriminability in the decoding process. To this end, the DNN pseudo-log-likelihoods are weighted according to the uncertainty variance assigned to the acoustic observation. The results presented here suggest that substantial reduction in WER is achieved with clean training. Moreover, modelling the uncertainty propagation through the DNN is not required and no approximations for non-linear activation functions are made. The presented method can be applied to any network topology that delivers log-likelihood-like scores. It can be combined with any noise removal technique and adds a minimal computational cost. This technique was exhaustively evaluated and combined with uncertainty-propagation-based schemes for computing the pseudo-log-likelihoods and uncertainty variance at the DNN output. Two proposed methods optimized the parameters of the weighting function by leveraging the grid search either on a development database representing the given task or on each utterance based on discrimination metrics. Experiments with Aurora-4 task showed that, with clean training, the proposed weighting scheme can reduce WER by a maximum of 21% compared with a baseline system with spectral subtraction and uncertainty propagation using the unscented transform. The uncertainty weighting method reduced the gap between clean and multi-noise/multi-condition training. This can be useful when it is not easy to train a DNN–HMM system in conditions that are similar to the testing ones. Finally, the presented results on the use of uncertainty are very competitive with those published elsewhere using the same database as the one employed here.

Keywords: Automatic speech recognition; Deep neural network; Uncertainty weighting; Uncertainty propagation; DNN–HMM

1. Introduction

Uncertainty variance in noise removal was firstly proposed to weight the information provided by frames according to their reliability in DTW and HMM algorithms (Yoma et al., 1997a, 1997b, 1998). To this end, the enhanced features (e.g. MFCC or filter-bank log-energies) should be considered random variables with the corresponding

☆ This paper has been recommended for acceptance by Roger K. Moore.
* Corresponding author.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

http://dx.doi.org/10.1016/j.csl.2017.06.005

mean and variance. In Yoma and Villar (2002) it was proposed “the replacement of the ordinary output probability with its expected value if the addition of noise is modelled as a stochastic process, which in turn is merged with the hidden Markov model (HMM) in the Viterbi algorithm.” As a result, the new output probability for the generic case of a mixture of Gaussians can be regarded as the definition of a stochastic version of the weighted Viterbi algorithm. This is because the final variances of the Gaussians correspond to the sum of the HMM and uncertainty variances. If the uncertainty variances increase, the discriminability of the GMM observation probability decreases and the decoding process relies more on the language model (Yoma et al., 2003). The Viterbi decoding algorithm, which incorporates the uncertainty in noise cancelling is called Stochastic Weighted Viterbi (SWV) algorithm because the increase of the GMM variances leads to a discriminability reduction of those frames with high uncertainty. Results with GMM−HMM-based speaker verification (Yoma and Villar, 2002) and speech recognition (Yoma et al., 2003, 2004) suggested that SWV can lead to significant WER reductions when speech signals are corrupted with additive, convolutional and coding-decoding distortion.

In Droppo et al. (2002), a similar result was later obtained by marginalizing the joint conditional pdf of the original and corrupted cepstral features over all possible unseen clean speech cepstra. Instead of using a model for additive noise, as in Yoma and Villar (2002), the pdf of the noisy features, given the clean coefficients, was assumed to be as a Gaussian distribution. However, this result employed the same idea of uncertainty proposed in Yoma et al. (1997a, 1997b, 1998). Additionally, in Droppo et al. (2002), the weighting nature of the use of uncertainty was not analysed. In Deng et al. (2005), a new classification rule was presented by proposing an integration over the feature space instead of over the model-parameter space. It was tested with connected speech recognition. The enhancement uncertainty variances were estimated by using a probabilistic and parametric model of speech distortion. In Trausti and Kristjansson (2002), two adaptation schemes were proposed to preserve the observation uncertainty. The results were obtained with connected digits. It is worth noting that in Yoma et al. (2003, 2004) a generalization of the model presented in Yoma and Villar (2002) was successfully applied to a continuous speech recognition task.

The uncertainty estimation of speech features was also later addressed in Liao and Gales (2005); Benítez et al. (2004); Tran et al. (2014); Astudillo and Orglmeister (2013). Particularly in Astudillo and Orglmeister (2013), it was shown that short-term Fourier transform (STFT) uncertainty propagation can be combined with the Wiener filter to compute minimum mean square error (MMSE) estimations in the feature domain for various parameter extraction methods. In contrast, despite the noise cancelling uncertainty being presented only for band-pass filters and MFCC coefficients, the proposed modelling employed in Yoma et al. (1997a, 1997b, 1998), Yoma and Villar (2002) does not require consideration of a Gaussian distribution for the additive noise in the STFT domain. Moreover, the non-linear log function is included by definition in the uncertainty estimation with spectral subtraction.

As mentioned above, in the context of band-pass filter bank analysis based features, the uncertainty in noise cancelling was proposed initially in Yoma et al. (1997a, 1998), and further developed in Yoma and Villar (2002). According to Yoma and Villar (2002) the uncertainty variance in noise cancelling in a band-pass filter is expressed as:

$$Var\left[\log\left(\overline{s_m^2}|\overline{y_m^2}\right)\right] = \begin{cases} \dfrac{2\cdot c_m \cdot \mathrm{E}\left[\overline{n_m^2}\right]}{\overline{y_m^2}\ -\mathrm{E}\left[\overline{n_m^2}\right]}, & \text{if } \overline{y_m^2}\ -E[\overline{n_m^2}] \geq 10\cdot c_m \cdot E[\overline{n_m^2}] \\ -\dfrac{\overline{y_m^2}\ -\mathrm{E}[\overline{n_m^2}]}{50\cdot c_m \cdot \mathrm{E}\left[\overline{n_m^2}\right]} + 0.4, & \text{else} \end{cases} \tag{1}$$

where $\overline{s_m^2}$, $\overline{y_m^2}$ and $\mathrm{E}[\overline{n_m^2}]$ are the estimated original clean energy, observed noisy energy and estimated noise energy at filter m, respectively. In addition, c_m is a correction coefficient that considers the short-term correlation between the clean and noise signals. According to Yoma et al. (1998), $\mathrm{E}[\log(\overline{s_m^2}|\overline{y_m^2})] = \log(\overline{y_m^2}\ -\mathrm{E}[\overline{n_m^2}])$, where $\overline{y_m^2}\ -\mathrm{E}[\overline{n_m^2}]$ can be seen as the spectral subtraction (SS) estimate of the clean signal. As shown in Yoma et al. (1998) and Yoma and Villar (2002), the uncertainty variance of the Mel filter bank and MFCC can be obtained with (1). The uncertainty variance of delta and delta−delta features can also be estimated as in Yoma and Villar (2002). This uncertainty variance is a key component of the SWV algorithm, which can lead to significant improvements in HMM-based speaker verification and speech recognition tasks.