ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Uncertainty driven Compensation of Multi-Stream MLP Acoustic Models for Robust ASR

Ramón Fernandez Astudillo, Alberto Abad, João Paulo da Silva Neto

Spoken Language Laboratory, INESC-ID-Lisboa, Lisboa, Portugal
ramon@astudillo.com, {alberto.abad,Joao.Neto}@inesc-id.pt

Abstract

In this paper we show how the robustness of multi-stream multi-layer perceptron (MLP) acoustic models can be increased through uncertainty propagation and decoding. We demonstrate that MLP uncertainty decoding yields consistent improvements over using minimum mean square error (MMSE) feature enhancement in MFCC and RASTA-LPCC domains. We introduce as well formulas for the computation of the uncertainty associated to the acoustic likelihood computation and explore different stream integration schemes using this uncertainty on the AURORA4 corpus.

Index Terms: uncertainty propagation, observation uncertainty, MLP, multi-stream

10.21437/Interspeech.2012-17

1. Introduction

A simple method to attain robust automatic speech recognition (ASR) is to apply speech enhancement in the short-time Fourier transform (STFT) domain as an independent pre-processing step. As depicted in Fig. 1, left, speech enhancement in STFT domain provides a point estimate of the clean STFT of a speech signal given the observed noisy STFT. This is followed by the extraction of features, e.g. Mel-Frequency cepstral coefficients (MFCCs), from the point estimate. As depicted in Fig. 1, right, the ASR system determines the acoustic likelihood of the extracted features for a set of acoustic units, from which words are composed. The most likely word sequence for the given features is then determined in the decoding stage.

One method to improve the integration of speech enhancement and ASR systems is STFT uncertainty propagation (STFT-UP) [1]. Let $\mathbf{X}$ denote a single analysis frame of the STFT of the clean signal. Let $\hat{\mathbf{X}}$ be its point estimate obtained from the STFT frame of the observed noisy signal $\mathbf{Y}$. Finally, let $\mathbf{f}()$ denote a vector valued non-linear feature extraction. STFT-UP replaces the point estimate $\hat{\mathbf{X}}$ by a random variable distributed according to a posterior distribution $p(\mathbf{X}|\mathbf{Y})$, which reflects the residual uncertainty after enhancement. Applying the feature extraction $\mathbf{f}()$ to this variable results in a posterior distribution of the speech features attained by solving

$$p(\mathbf{x}|\mathbf{Y}) = \int_{\mathbb{C}^K} \boldsymbol{\delta}_I(\mathbf{x} - \mathbf{f}(\mathbf{X}))p(\mathbf{X}|\mathbf{Y})\mathbf{dx} \qquad (1)$$

where $\boldsymbol{\delta}_I$ is the multivariate delta, K is the number of frequency bins considered and I is the dimensionality of the features. STFT-UP provides various approximations to compute this posterior for MFCC and RASTA-LPCC feature extractions among other [1]. In particular, if the posterior after enhancement $p(\mathbf{X}|\mathbf{Y})$ is computed from a Wiener filter, it can be demonstrated that the mean of $p(\mathbf{x}|\mathbf{Y})$ is a minimum mean square error (MMSE) estimate directly in feature domain [2]. In principle, the acoustic likelihood can be directly computed from this MMSE estimate. However, unlike other MMSE estimators in feature domain e.g. [3], STFT-UP provides the whole posterior associated to that estimate $p(\mathbf{x}|\mathbf{Y})$. This is equivalent to the posterior obtained through ALGOQUIN [4] or other similar methods and can thus be combined with observation uncertainty techniques. The simplest of these techniques is to compute the expected likelihood with respect to the uncertain features posterior given by

$$E\{p(\mathbf{x}|c_n)\} = \int_{\mathbb{R}^I} p(\mathbf{x}|c_n)p(\mathbf{x}|\mathbf{Y})\mathbf{dx} \qquad (2)$$

if $p(\mathbf{x}|c_n)$ was to be a conventional Gaussian mixture acoustic model, solving Eq. 2 yields a simple variance compensation rule introduced in [5] which is also known as front-end uncertainty decoding (UD)[1].

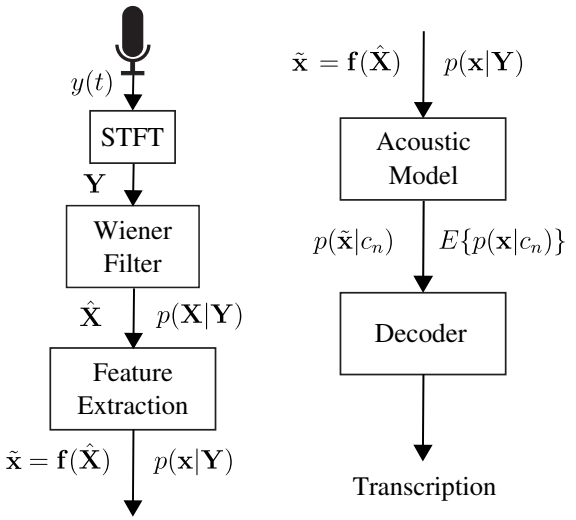

Figure 1: *Left: Speech enhancement and feature extraction. Right: Acoustic likelihood computation and decoding. Intermediate variables for conventional systems (left) and uncertainty propagation and decoding systems (right) provided.*

This work was partially funded by the DIRHA European project (FP7-ICT-2011-7-288121) and the Portuguese Foundation for Science and Technology (FCT) through the grant number SFRH/BPD/68428/2010 and the project PEst-OE/EEI/LA0021/2011

Preprint submitted to INTERSPEECH 2012.

[1]Note that UD is usually derived as an approximation of the modified Bayesian decoding rule for ASR [6], rather than as the expected likelihood.

In the case considered in this paper, however, multi layer perceptron (MLP) based acoustic models are used. The likelihood is computed in this case by using

$$p(\mathbf{x}|c_n) \propto \frac{p(c_n|\mathbf{x})}{p(c_n)} \tag{3}$$

where $p(c_n)$ is the phonetic class prior modeled by a categorical distribution and $p(c_n|\mathbf{x})$ the phonetic class posterior modeled with an MLP. In [7], two approximate solutions to compute the expected likelihood in Eq. 2 for MLPs were introduced and are here referred as MLP uncertainty decoding (MLP-UD). The experiments in [7] were however carried with a posterior determined from oracle information which casted some doubts over its application to real scenarios.

This paper demonstrates the applicability of MLP-UD to real scenarios by estimating the feature posterior $p(\mathbf{x}|\mathbf{Y})$ from a Wiener filter using STFT-UP. Apart from the MMSE-MFCC estimator introduced in [2] an MMSE-RASTA-LPCC estimator is here used as a second stream. Furthermore, solutions are introduced to compute the variance associated to the expected likelihood computation in Eq. 2. The use of this variance to weight stream integration is also explored. Results on the AURORA4 corpus show that MLP-UD consistently outperforms conventional point estimates. The integration of uncertain streams also provides promising results.

2. Piecewise Exponential Solution to MLP-UD

2.1. Reviewing the Expected Likelihood Computation

A multilayer perceptron is a non-linear function obtained by concatenation of successive layers of nodes. The n^{th} node of each layer computes first a weighted sum of all node outputs from the previous layer z_n. After this, a saturation non-linearity is applied, the most typical being the sigmoid. The MLP-UD solution proposed in [7] assumed that each node output after the weighted sum z_n could be modeled as an independent Gaussian variable due to the central limit theorem

$$z_n \sim \mathcal{N}(\mu^{z_n}, \Sigma^{z_n}). \tag{4}$$

This decomposes the solution of Eq. 2 into successively computing mean and variance after each layer of the MLP. The only pending problem is computing the first and second moments of a Gaussian variable transformed through the sigmoid function. For this purpose a piecewise exponential approximation of the sigmoid was used

$$\frac{1}{1+e^{-z_n}} \approx 2^{z_n-1}u(-z_n) + (1-2^{-z_n-1})u(z_n) \tag{5}$$

where $u(x)$ is the unit step function. For this approximation, exact closed form solutions can be derived for the propagation of the mean

$$\begin{aligned} E\{\mathrm{sig}(z_n)\} \approx & \frac{1}{2}\cdot\Omega(\ln(2)\mu^{z_n}, \ln(2)^2\Sigma^{z_n}) \\ & - \frac{1}{2}\cdot\Omega(-\ln(2)\mu^{z_n}, \ln(2)^2\Sigma^{z_n}) \\ & + \phi(0, -\mu^{z_n}, \Sigma^{z_n}) \end{aligned} \tag{6}$$

and second order central moment

$$\begin{aligned} E\{\mathrm{sig}(z_n)^2\} \approx & \frac{1}{4}\cdot\Omega(2\ln(2)\mu^{z_n}, 4\ln(2)^2\Sigma^{z_n}) \\ & + \frac{1}{4}\cdot\Omega(-2\ln(2)\mu^{z_n}, 4\ln(2)^2\Sigma^{z_n}) \\ & - \Omega(-\ln(2)\mu^{z_n}, \ln(2)^2\Sigma^{z_n}) \\ & + \phi(0, -\mu^{z_n}, \Sigma^{z_n}) \end{aligned} \tag{7}$$

where ϕ is the cumulative density function (CDF) of the Gaussian variable and Ω is the partial expectation of the exponential of a Gaussian variable, which also depends on ϕ.

One of the advantages of this propagation approach is that its computational cost is low and scales linearly with the number of nodes, leading to real time performance on a Matlab implementation.

2.2. Variance Associated to the Expected Likelihood

The objective of the approach presented in this paper is to derive the variance associated to the MLP posterior computation implicit in Eq. 2 when using the piecewise exponential approximation. That is

$$\lambda_n = \mathrm{Var}\{p(c_n|\mathbf{x})\}. \tag{8}$$

This variance can then be used in posterior steps like decoding, although in this paper we explore its use only for stream integration. For this purpose it is first necessary to consider an additional step which was ignored in [7]. In order to provide a normalized output, the sigmoid of the last layer of an MLP is usually replaced by a soft-max transformation

$$p(c_n|\mathbf{x}) = \frac{e^{z_n}}{\sum_{n'}^{J} e^{z_{n'}}} = \exp\left(z_n - \log(m)\right) \tag{9}$$

where J is the total number of layer nodes and m the normalizing denominator. Since in [7] only the expected likelihood was needed, soft-max was applied directly to the node means μ^{z_n}. However, it is also possible to approximate the transformation of a Gaussian variable through the soft-max transformation. Transforming a Gaussian variable through the exponential leads to the well known log-normal distribution. The distribution of $z_n - \log(m)$ is however unknown. It is tempting to think that since m is a sum of independent log-normal distributions it is approximately log-normal, and thus z_n and $\log(m)$ are jointly Gaussian. This approximation was implemented but led to high computational costs with little improvement. The approach used here instead a simplification which neglects the uncertainty of the normalization denominator m, or equivalently considers the term subtracted to the exponent in the leftmost expression in Eq. 9 as a deterministic constant. Since subtracting a constant from a Gaussian variable does only alter its mean the resulting variance is the variance of a log-normal distribution given by

$$\begin{aligned} \mathrm{Var}\{p(c_n|\mathbf{x})\} \approx & \left(\exp\left(\Sigma^{z_n}\right) - 1\right) \\ & \cdot \exp\left(2\mu^{z_n} - \log(m) + \Sigma^{z_n}\right). \end{aligned} \tag{10}$$

The equivalent assumption can be here used to compute the mean of the propagation through the soft-max. This led however to little improvements over using the soft-max of the mean and was left out of the experimental results.

3. Multi-stream speech recognition

Multi-stream processing is a successful approach to enhance the generalization capability of speech recognizers. In HMM/MLP

hybrid recognition systems multi-stream combination can be very efficiently implemented at the probability level combining the posterior probabilities obtained with several MLPs.

3.1. MLP posterior combination

Our in-house speech recognizer implements posterior multiple-stream combination based on the product rule like in [8] except for the normalization terms. The product rule can be generalized to a geometric mean combination rule, that is:

$$p(c_n|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} p^{w_s}(c_n|\mathbf{x}_s)}{p^{S-1}(c_n)} \tag{11}$$

where $\mathbf{x}_{1:S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_s, \ldots, \mathbf{x}_S\}$ is the set of S feature stream observations, $p(c_n|\mathbf{x}_{1:S})$ is the result of combining the posterior probabilities $p(c_n|\mathbf{x}_s)$ of every stream, $p(c_n)$ is the class prior, and w_s are exponential weights for each stream such that $\sum_{s=1}^{S} w_s = S$. These weights w_s represent the reliability or confidence of the posterior estimations provided by each feature stream. Several approaches to stream weight estimation can be found in the literature. For a review of some of these methods refer to [8].

3.2. Uncertainty based stream combination

For a set of multi-stream feature vectors $\mathbf{x}_{1:S}$, we hypothesize that the variance of each stream $\lambda_s = \text{Var}\{p(c_n|\mathbf{x}_s)\}$ derived in previous section may be a good indicator of the reliability of that stream at estimating the posterior probability of class c_n. In order to obtain a class independent confidence measure, we compute the average variance $\bar{\lambda}_s = \sum_n \text{Var}\{p(c_n|\mathbf{x}_s)\}/J$. Then, given the average variance $\bar{\lambda}_s$, an uncertainty score of the stream s can be defined as:

$$r_s = \frac{\bar{\lambda}_s^{\gamma}}{\sum_{s'=1}^{S} \bar{\lambda}_{s'}^{\gamma}} \tag{12}$$

where $\gamma \in [0, \infty)$ is an arbitrary exponential factor that allows different ways of computing stream uncertainty scores. These uncertainty scores r_s can be used to derive stream exponential weights for Eq. 11 that incorporate variance posterior information, for insnance:

$$w_s = \tfrac{S}{S-1}(1 - r_s) \tag{13}$$

where the $\frac{S}{S-1}$ constant term assures that $\sum_{s=1}^{S} w_s = S$. Given that the simple equal weight solution usually provides excellent stream fusion results [8], it is convenient to define w_s as a distortion that depends on the uncertainty score around $w_s = 1$. Thus, for the particular case of $S = 2$, it is possible to re-write w_s to introduce a $\beta \in [0, 1]$ term to balance the influence of r_s with respect to the equal weight solution as follows (for $S > 2$ a different normalization is needed):

$$w_{s=\{1,2\}}^{\beta} = \tfrac{S}{S-1}(0.5 + \beta(0.5 - r_s)) \tag{14}$$

Notice that for shake of clarity the time frame index was omitted in the derivations of this section. However, the stream uncertainty score of Eq. 12 is in fact time dependent, since it is computed for every time frame, and consequently w_s are time varying weights. Thus, it is possible to compute the time smoothed uncertainty score $r'_s(t)$ at time instant t as

$$r'_s(t) = (1 - \alpha)r'_s(t-1) + \alpha r_s(t) \tag{15}$$

where α is a smoothing factor, $r'_s(t-1)$ is the previous frame smoothed uncertainty score, and $r_s(t)$ is the instantaneous estimation of Eq. 12. It can be easily demonstrated that $\sum_{s=1}^{S} r'_s(t) = 1$, then previous expressions for the computation of stream weights stand.

4. Experiments and Results

4.1. Experimental set-up

In order to test the proposed techniques in a realistic scenario the AURORA4 framework [9], an artificially corrupted version of the the Wall Street Journal database, was used. The training data corresponds to the SI-84 corpora and the test set is based on the November 1992 ARPA WSJ evaluation set, contaminated with six different types of noise. The test set was divided into three groups, clean, stationary noise (car), for which speech enhancement is usually more effective, and non-stationary noises. The systems are evaluated with the WSJ standard 5K non-verbalized closed bi-gram language model. The recognition setup implements a version of our own in-house hybrid ASR system AUDIMUS [10] with multiple state sub-phoneme recognition units and a restricted set of phone transition units [11]. Two feature streams, corresponding to amplitude based MFCC features and RASTA-LPCC features were considered. These were complemented with delta and acceleration coefficients as well as cepstral mean subtraction. The posterior in STFT domain was computed with a Wiener filter with an IMCRA noise variance estimator [12] as described in [2]. The posterior propagation was computed using STFT-UP as described [1, Ch. 6]. The only change with respect to the usual methods was using the log-normal assumption for the amplitude MFCCs, rather than the Unscented transform, since it is numerically more stable. Diagonal covariance was used for the MMSE-MFCC estimator and full covariance for the MMSE-RASTA-LPCC estimator, since it affects the accuracy of the method.

4.2. Results

Four speech enhancement cases have been considered: no enhancement (baseline), conventional Wiener filter[2], Wiener propagated posterior mean (MMSE-feature estimate) and MMSE-feature estimate with MLP-UD. In Table. 1 word error rate (WER) results are provided for each individual stream (top), conventional equal weight stream fusion (middle) and uncertainty based stream fusion (bottom).

Single stream results show complementary performances for MFCC and RASTA-LPCC features. MFCCs show a particularly better performance for stationary noise whilst RASTA-LPCC outperform MFCCs in clean speech and non stationary noises. The use of a simple speech enhancement pre-processing provides a reduction of WER, with the Wiener filter outperforming MMSE estimators in feature domain in some cases. The additional use of MLP-UD however, improves the use of MMSE estimators in feature domain in all scenarios also outperforming all other techniques for noisy speech.

Regarding conventional stream fusion, results do notably improve for all conditions. Wiener and the MMSE estimator in feature domain show similar performance although the MMSE point estimator yields a very good suppression of stationary noise. This is coherent with the fact that the propagated vari-

[2]Wiener filter was chosen instead of other well known methods like the Ephraim-Malah filters since it gave better results in the initial experiments.

Table 1: *Top-down: Word error rates for MFCC and RASTA individual streams, conventional and uncertainty driven fusion.*

	Clean	Stat.	Non-St.
Single Stream			
MFCC Baseline	11.1	23.8	50.5
Wiener	**10.2**	21.7	49.4
MMSE-MFCC	12.0	21.5	48.5
MMSE-MFCC+MLP-UD	10.8	**19.8**	**46.4**
RASTA-LPCC Baseline	**10.8**	28.0	48.4
Wiener	12.4	22.2	45.6
MMSE-RASTA	11.2	22.1	45.8
MMSE-RASTA+MLP-UD	11.0	**21.0**	**43.1**
Two Streams			
MFCC+RASTA Baseline	9.2	19.4	41.7
Wiener	8.8	17.1	38.6
MMSE-MFCC+RASTA	8.9	15.5	39.0
+MLP-UD	**8.6**	**15.4**	**37.8**
Uncertain Stream Fusion (MFCC+RASTA+MLP-UD)			
$\gamma = 1, \beta = 1, \alpha = 1$	8.3	15.9	38.5
$\gamma = 0.5, \beta = 1, \alpha = 1$	8.2	15.3	38.0
$\gamma = 0.1, \beta = 1, \alpha = 1$	8.4	15.2	37.8
$\gamma = 0.05, \beta = 1, \alpha = 1$	8.5	15.2	**37.7**
$\gamma = 1, \beta = 0.75, \alpha = 1$	8.2	15.5	38.2
$\gamma = 1, \beta = 0.5, \alpha = 1$	8.3	15.3	38.0
$\gamma = 1, \beta = 0.25, \alpha = 1$	8.3	**15.1**	37.9
$\gamma = 1, \beta = 1, \alpha = 0.75$	**8.1**	15.7	38.6
$\gamma = 1, \beta = 1, \alpha = 0.5$	8.3	15.7	38.6
$\gamma = 1, \beta = 1, \alpha = 0.25$	8.2	15.8	38.6

ance of the Wiener filter only takes into account the residual mean square error and not the errors in the estimation of noise variances [2]. Such errors appear often in the case of non stationary noise, when the IMCRA speech probability estimator fails, and thus decrease the efficiency of the approach. As in the single channel case, the best results for all conditions and equal weight stream fusion are achieved when using MLP-UD together with the MMSE point estimates. Although consistent, these improvements are small compared to the ones attained with oracle uncertainties [7]. This shows that there is room for improvement if uncertainties are better estimated.

Regarding uncertainty driven stream fusion, results for the default scheme ($\gamma = \beta = \alpha = 1$), together with some representative results varying these parameters are provided. Comparing to the equal weight combination, the most relevant fact is that improvements are always obtained in the clean condition case independently of the fusion parameters. Moreover, modest improvements are achieved with some configurations in stationary noise conditions, while no improvements or even worse performance is generally attained in non-stationary noise conditions. In the latter case, the inaccuracies produced in the estimation of the variance commented previously, affects negatively to the computation of the stream weights. Thus, it might result surprising that propagating the variance of a speech enhancement system improves the results mainly when no noise is present. It must be however taken into account that speech enhancement does not work perfectly and thus modifies the clean speech signal, which generates an uncertainty that is accounted for by the variance. Anyway, it must be also considered that there is not reason to believe that the method proposed here for uncertainty based multi-stream fusion is the best possible approach. Alternative methods to the geometric mean combination rule to incorporate uncertainty (see for instance [13]), other formulas for computing the uncertainty scores and different stream weights derivations may likely result in enhanced combination schemes.

5. Conclusions

It has been shown how MLP uncertainty decoding yields consistent improvements over using MMSE feature enhancement when the uncertainties are non-ideally estimated. Inclusion of an additional MMSE-RASTA-LPCC stream notably increases performance. An approximate solution to derive the variance associated to MLP-UD has also been proposed. Initial experiments using this variance for multi-stream fusion shows performance improvements although the technique must be further explored.

6. References

[1] R. F. Astudillo, "Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, Technische Universität Berlin, 2010.

[2] R. F. Astudillo and R. Orglmeister, "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in *Proc. Interspeech*, 2010.

[3] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *ICASSP 2008*, 2008, pp. 4041–4044.

[4] B. Frey, L. Deng, A. Acero, T. T., and Kristjansson, "Iterating laplaces method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[5] N. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Trans. Speech, Audio Processing*, vol. 10 (3), pp. 158–166, 2002.

[6] L. Deng, "Front-end, back-end, and hybrid techniques to noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Berlin, Germany: Springer, 2011, ch. 4, pp. 67–99.

[7] R. F. Astudillo and J. P. Neto, "Propagation of uncertainty through multilayer perceptrons for robust automatic speech recognition," in *Proc. Interspeech*, 2011, pp. 461–464.

[8] A. Hagen, *Robust speech recognition based on multi-stream processing*. PhD thesis, École Poly. Fédérale de Lausanne, 2001.

[9] G. Hirsch, *Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task*, Niederrhein University of Applied Sciences, Nov. 2002.

[10] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: A broadcast news speech recognition system for the european portuguese language," in *Computational Processing of the Portuguese Language, LNCS*, 2003, vol. 2721, pp. 196–196.

[11] A. Abad and J. Neto, "Incorporating acoustical modeling of phone transitions in an hybrid ann/hmm speech," in *Proc. Interspeech*, 2008, pp. 2394–2397.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[13] F. Valente and H. Hermansky, "Combination of acoustic classifiers based on dempster-shafer theory of evidence," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.