Answer to decision on NLE-ARTC-15-0019 (Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children)

Romain Serizel, Diego Giuliani

July 24, 2015

1 Reviewer 1

Comments 1-3, 5-7, 9, 11, 13-14, 31

This has been corrected

4. Page 7. 3rd equation. How is the quantity $p(X)$ calculated?

$p(X)$ should actually be $p(S)$. This has been corrected.

8. Page 10, the penultimate paragraph, starting "The VTLN procedure..." is not clear. In the DNN that learns the warping factor, are the inputs individual feature vectors or whole utterances? The figure suggests individual feature vectors, but the sentence "Then training utterances and corresponding warping factors" suggests utterances. In the second part of this paragraph "This DNN is then used to produce the posterior probabilities of the VTLN warping factors..." needs to be clarified. Specifically you need to say that the outp[ut of this DNN is a vector, and the dimension of the vector corresponds to the number of discrete VTLN normalisation factors that are considered. This is clarified later, but it is confusing at this point.

The paragraph has been modified into:

The VTLN procedure is first applied to generate a warping factor for each utterance in the training set. Each acoustic feature vector in the utterance is labeled with the utterance warping factor. Then, training acoustic feature vectors and corresponding warping factors are used to train a DNN classifier. Each class of the DNN correspond to one of the discrete VTLN factors and the dimension of the DNN output corresponds to the number of discrete VTLN factors. The DNN learns to infer the VTLN warping factor from the acoustic feature vector

(Figure **??***) or more precisely the posterior probability of each VTLN factors knowing the input acoustic feature vector. This DNN will be referred to as DNN-warp.*

10.1 Page 12. The first sentence is not grammatical and needs attention.

The sentence has been modified into:

The ultimate goal here is not to estimate the VTLN warping factors but to perform robust speech recognition on heterogeneous corpora. To this end, the DNN-warp and the DNN-HMM can be optimised jointly (Figure 3).

10.2 Also, there is more discussion of the "posteriors of the warping factor" and this has still not been explained, and the phrase "posteriors of the warping factor" suggests that there is only one warping factor.

The posteriors probabilities are now explained p10 (see also comment 8). "Posterior probabilities of the warping factor" has been replaced everywhere by "Posterior probabilities of the warping factors" to avoid confusion.

Diego: 12. Page 13, 4.1.1 given that VTLN is likely to be most effective for younger children, it would be informative to see the distribution of ages in the ChildIt corpus.

13. 4.1.1 and 4.1.2 say something about the transcriptions of the data that are available - word-level or phone-level?

It is now mentionned for each corpus only word-level transcription is available. Also, at the begining of 24.2 we explain how the phone level transcription are obtained:

The approaches proposed in this paper have been first tested on small corpora (ChildIt + APASCI) for phone recognition to explore as many set-ups as possible in a limited amount of time. The reference phone-level transcriptions are obtained with Viterbi forced-alignement performed with our best HMM-GMM system at the moment of the experiments.

Diego: 18. Page 16. 3rd line. Please explain what is meant by "... single consonants and their germinant counterparts..." I have no idea what this means.

19. Page 16, 4.3. Does the HMM set used for word recognition use the same set of tied states as the previous HMM set?

The set of tied states are different. This is now mentionned explicitly.

Paragraph 4.2.1:

Acoustic models are 3039 tied-state triphone HMM based on a set of 48 phonetic units derived from the SAMPA Italian alphabet.

Paragraph 4.3.1:

The HMM-GMM are similar to those used for phone recognition except that they use more Gaussian densities to benefit from the extensive training data.

Acoustic models are 5021 tied-state triphone HMM based on a set of 48 phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modelled with a mixture of 32 Gaussian densities having a diagonal covariance matrix. In addition, "silence" is modelled with a Gaussian mixture model having 32 Gaussian densities.

20. line 2. Be more precise. What exactly do you mean by saying that the DNN was trained on a different set of Gaussians? This is too imprecise.

This has been corrected, paragraph 4.3.2 is now: *The DNN are similar to those used for phone recognition except that they are trained on a different set of targets. The targets of the DNN are the 5021 tied-states obtained from the word recognition HMM-GMM training on the mixture of adults' and children's speech (ChildIt + IBN). The DNN has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can be summarised as follows: 208 x 1500 x 1500 x 1500 x 1500 x 5021.*

21. Page 17, 4.4. In adaptation are all of the other DNN training parameters the same as in 4.2.2

Yes all the training parameters are similar to 4.2. Paragrazph 4.4 is now:

One option is to adapt an already trained general DNN to group specific corpora. The data architecture is the same as described above. The initial DNN weights are the weights obtained with a pre-training/training procedure applied on all training data available (ChildIt+APASCI, respectively ChildIt + IBN). The DNN is then trained with back propagation on a group specific corpora (ChildIt, adult female speech in APASCI and adult male speech in APASCI, respectively IBN). The training parameters are the same as during the general training (4.2.2 and 4.3.2, respectively) and the learning rate follows the same rule as above. The mini-batch size is 512 and a first-order momentum of 0.5 is applied.

22. Page 17, last line. This is the first time that the number of VTLN factors is specified, or even that it is acknowledged that only a discrete set of factors is considered and therefore it is possible to create a vector of posterior probabilities. This basic principle needs to be introduced much earlier to understand how the DNN for estimating VTLN factor posteriors works. This is now mentionned more explicitly in paragraph 3.1:

[...] A well known method for estimating the scaling factor is ***based on a grid search over a discrete set of possible scaling factors*** *by maximizing the likelihood of warped data given a current set of HMM-based acoustic models (Lee and Rose, 1996). Frequency scaling is performed by warping the power spectrum during signal analysis or, for filter-bank based acoustic front-end, by changing the spacing and width of the filters while maintaining the spectrum unchanged (Lee and Rose, 1996). In this work we adopted the latter approach*

***considering a discrete set of VTLN factors**. Details on the VTLN implementation are provided in Section 4.5.*

Also an explanation about the link between VTLN factors and posterior probabilities has been been given as answer to comment 8.

Diego: 23. Page 18. 1st paragraph. Why do you use HMMs with just 1 Gaussian component per state to optimise the VTLN factor?

24. Page 18, paragraph 3. In the specification of the DNN, what is the meaning of the (5021) in brackets? Does this mean that there are different numbers of tied states in different systems? If so, please explain.

The paragraph has been modified into:

[. . .] The new DNN acoustic model has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can then be summarized as follows: 233 x 1500 x 1500 x 1500 x 1500 x 3039 for phone recognition (233 x 1500 x 1500 x 1500 x 1500 x 5021 for word recognition).

25. Page 18. Where does the learning rate of 0.0002 come from?

The learning rate was chosen empirically. For higher learning rates, the training accuracy would improve but not the cross-validation accuracy. The DNN obtained were not saved by the training script until both training and cross-validation accuracy progressed (at learning rate 0.0002). Setting the learning rate directly at 0.0002 is just a way to speed up the training process. Paragraph 4.6 is now:

The DNN-warp and DNN-HMM can be fine-tuned jointly with back-propagation. In such case, the starting learning rate is set to 0.0002 in the first 4 hidden layers (corresponding to the DNN-warp) and to 0.0001 in the last 4 hidden layers (corresponding to the DNN-HMM). The learning rate is chosen empically as the highest value for which both training accuracy and cross-validation accuracy imrove. Setting a different learning rate in the first 4 hidden layers and the last 4 hidden layers is done in a attempt to overcome the vanishing gradient effect in the 8 layers DNN obtained from the concatenation of the DNN-warp and the DNN-HMM. The learning rates are then adapted following the same schedule as described above. The joint optimisation is done with a modified version of the TNet software package (Vesely et al., 2010).

32. Page 28. Lines 5-7. Augmenting features that have already been VTLN normalised with thge posteriort probabilities of the VTLN factors seems an odd thing to do. What is the motivation/justification?

VTLN-normalisation operates at utterances level whereas posterior probabilities are obtained at frame level. While estimating VTLN factors on a longer time unit (utterance) should allow for a more accurate average estimation, the ”true” warping factor might be fluctuating in time [ref]. We belive that combining VTLN normalisation at utterance level and posterior probabilities estimated at frame level should help overcoming this problem. According to results, this seems to be true. A second paragraph stating this has been added in 5.2.2:

The approaches combining VTLN-normalised features and posterior probabilties aim at testing the complementary between VTLN-normalisation that operates at utterances level and posterior probabilities that are obtained at frame level. While estimating VTLN factors on a longer time unit(utterance) should allow for a more accurate average estimation, the "true" warping factor might be fluctuating in time [ref]. Combining VTLN normalisation at utterance level and posterior probabilities estimated at frame level should help overcoming this problem.

2 Reviewer 2

Diego: **1. In section 4.5, the authors conduct a grid search based on a set of 25 warping factors via the standard maximum likelihood using a triphone single Gaussian system. My question is that are there any difficulties in using a more robust triphone GMM system to search the warping factors? Single Gaussian is definitely not robust enough for a reliable warping factor. The easiest argument is that at least two Gaussians are needed to model both genders. In addition, what if the warping factor is per speaker rather than per utterance?**

3 Reviewer 3

Comments III-V

This has been corrected