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 30, 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 output 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 labelled 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

	Grade						
	2	3	4	5	6	7	8
N. Speakers	24	24	23	24	28	26	22

Table 1: Distribution of speakers in the ChildIt corpus per grade. Children in grade 2 are approximatively 7 years old while children in grade 8 are approximatively 13 years old.

vector (Figure 1) 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.

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.

A table reporting the age distribution has been added in Section 4.1.1 (see Table 1).

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 mentioned for each corpus only word-level transcription is available. Also, at the beginning 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 transcription of an utterance was derived from the corresponding word transcription by performing Viterbi decoding on a pronunciation network. This pronunciation network was built by *concatenation of the phonetic transcriptions of the words in the word transcription. In doing this alternative word pronunciations were taken into account and an optional insertion of the silence model between words was allowed.*

15. Page 15, 4.2.2. Please explain a little more about how the Hamming window is applied to the sequence of feature vectors. Is this normal in DNN training?

16. Page 15, 1 line later, what was the rationale for choosing DCT (on the previous page you applied HLDA)

It is common to reduce the dimensionality of the input vector in DNN. DCT is probably not the most common choice (PCA and HLDA are commonly used) but it offers the advantage to be rather simple and requires little computational resources (as opposed to HLDA for example). Hamming windowing is applied to each frequency band (31-dimensional context in each band) before the DCT compression. This is a rather standard procedure. The processing description in the paper was not accurate, this has been corrected. Paragraph 4.2.2 has been corrected to:

The DNN uses again 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. The context spans on a 31 frames window. For each frequency band, the 31 coefficients context is separately scale with Hamming and projected to a 16 dimensional vector using DCT. The 13 resulting vectors are concatenated to obtain a 208 dimensional feature vector which is normalised to have zero-mean and unit variance before being used as input to the DNN. [. . .]

Paragraph 4.5 has been modified to:

The DNN-warp inputs are the MFCC with a 61 frames context window, DCT projected to a 208 dimensional features vector (the procedure is similar as in 4.2.2)

17. Page 15, 2nd and 3rd paragraphs. How was the size of the network chosen? How were all of the parameters in DNN training chosen? Is performance sensitive to the precise values of these parameters?

These are rather standard parameters when using DNN for acoustic modelling in automatic speech recognition: context is usually between 9 frames and 40 frames, hidden layers size between 1000 and 2000 elements and the DNN usually have 3 to 8 hidden layers. We simply adjust these values to our specific corpus. For example the corpus is rather small therefore we have only 4 hidden layers to prevent over-fitting The performance are sensitive to these parameters but they will not be affected drastically by minor changes on the hyper-parameters value. This is now mentioned at the end of the 2nd paragraph in 4.2.2:

The values of the hyper-parameter (network topology and learning parameters) are standard values, in the range of the values commonly used for these parameters in the literature.

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

In phonetics, gemination or consonant elongation happens when a spoken consonant is pronounced for an audibly longer period of time than a short consonant. Gemination is distinct from stress and may appear independently of it. Gemination literally means "twinning", and is from the same Latin root as "Gemini".

Consonant length is distinctive in some languages including Italian. In these languages most consonants have two versions: the short ("single") version and the corresponding long version (its "geminate counterpart").

For ASR purposes all consonants are modelled, however, traditionally for the Italian language when computing the phone error rate "no distinction is made between single consonants and their geminate counterparts". The motivation is that the main acoustic difference between "single consonants and their geminate counterparts" is the duration so that a single consonant is highly confusable with its geminate counterpart. To avoid to inflate phone recognition results, confusion between a consonant and its geminate counterpart is not considered as an error.

We have not changed the text in the paper as it seems sufficiently clear that the phone error rate is computed on a reduced set of 28 phone labels.

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 mentioned 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. Paragraph 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 mentioned 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.

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

The approach for estimating the VTLN scaling factors making use of speaker-independent triphone HMMs with just 1 Gaussian density per state was proposed by Welling et al. in the reference paper:

- L. Welling, S. Kanthak and H. Ney, “Improved Methods for Vocal Tract Normalization”, in Proc. of IEEE ICASSP, 1999, Vol. 2, pp. 761-764.

In the paper it is empirically verified that having 1 Gaussian density per state is a good choice when compared with the use of many Gaussian densities per state.

In past works, we adopted this approach in the context of children's and adults' speech recognition with good results, see for example the reference papers:

- M. Gerosa, D. Giuliani and Fabio Brugnara, "Acoustic variability and automatic recognition of children's speech", Speech Communication, 2007, Vol. 49, N. 10-11, pp. 847-860.
- M. Gerosa, D. Giuliani and Fabio Brugnara, "Towards age-independent acoustic modeling", Speech Communication, 2009, Vol. 51, N. 6, pp. 499-509.

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 empirically as the highest value for which both training accuracy and cross-validation accuracy improve. 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).

26. A general point. The adapted system is close to optimal in all cases, but most of the discussion is biased towards the benefits

of VTLN. For me, one of the most interesting consequences of this paper is that adaptation appears to compensate for VTL differences without any explicit VTLN. Of course, one of the problems with the "adapted system" is that model selection is not done automatically. A simple way to achieve this would be to compute the probability of an utterance for all three adapted models and then apply the highest scoring one to recognition. Why didn't you do something like this? It may be that even if this approach makes an "error" and, say, classifies a particular child as a female adult, the female adult model may give best performance on this child's data.

This comment is partially answered in comment 29. The solution you propose for model selection seems indeed interesting. Yet, our goal in this paper was to experiment on features and acoustic modelling with DNN and we voluntarily decided to limit the scope of the paper on these particular points for clarity. The approach you mention involves a full decoding system and is therefore beyond the scope of this paper. This is now clearly stated in the abstract:

For clarity sake, the scope of this paper is voluntarily limited to approaches based only on acoustic modelling and features transform based on VTLN.

In the introduction:

This paper voluntarily focuses on age and gender groups problems and and approaches based only on acoustic modeling and VTLN based transform in order to focus on the effects of these particular approaches. Therefore state-of-the-art approaches based on speaker identity models such as I-vector (Dehak, Kenny, Dehak, Dumouchel, and Ouellet, 2011; Saon, Soltau, Nahamoo, and Picheny, 2013; Senior and Lopez-Moreno, 2014), speaker code (Abdel-Hamid and Jiang, 2013a), linear input network and linear output networks (Li and Sim, 2010) are beyond the scope of this paper. At this initial stage, the authors in decided to focus on approaches based only on acoustic therefore excluding approach requiring an entire decoding stream (or at least a language model), such as approaches based on confusion networks (Mangu, Brill, and Stolcke, 2000).

And in the conclusion:

This paper voluntarily focused on age and gender groups problems and and approaches based only on acoustic modeling and VTLN based transform in order to focus on the effects of these particular approaches. Extension of this work could consider approaches base based on speaker identity models such as I-vector based approaches, speaker code, linear input network and linear output networks approaches such as approaches based on confusion networks that require an entire decoding stream (or at least a language model).

27. Page 22. In the earlier overview of the systems, did you include the case of MFCC feature vectors augmented with VTLN warping factors obtained in the normal way?

We did try to augment the MFCC with the warping factors obtained in the standard way (in this case there is only one scalar factor per utterance) the result of these experiments are report in Table 5 row *Warp + MFCC*.

28. Page 22. Exactly how are the posterior probabilities averaged to utterance level, and why is this done?

We compute a vector of averaged posterior probability for each warping factor over utterances. This can be considered as an intermediate step between the standard way to obtain warping factors and using the DNN-warp. This way we can check what is the impact of having of hard or soft decision on the warping factors (*Warp + MFCC* vs. *Warp-post (utt) + MFCC*) without considering the effect of the time unit on which the warping factor are computed. The effect of the time unit are check with: *Warp-post + MFCC* vs. *Warp-post (utt) + MFCC*. This is now clarified in an additional paragraph after the 1st paragraph of 5.1.3:

To compute the vectors Warp-post (utt) + MFCC *the posterior probability of each warping factor is averaged over utterances to obtain a vector of averaged posterior probabilities. This experiment allow to study independently the effects of having a soft or hard decision on the warping factor selection and the effects of the time unit used to compute the warping factors. The impact of having of hard or soft decision on the warping factors is check comparing* Warp + MFCC *to* Warp-post (utt) + MFCC*. While the effects of the time unit used to compute warping factors are check comparing* Warp-post + MFCC *to* Warp-post (utt) + MFCC

29. Page 23. Final paragraph. This is an example of what I was indicating earlier. The first line of the final paragraph is very "pro-VTLN" and ignores the fact that better performance is obtained by simple adaptation.

We do agree with the reviewer comment on this particular sentence and the second half of the sentence has been removed to try to be less biased. However, we believe that the last paragraph before the conclusion and the conclusion itself are not particularly biased towards VTLN approach. Indeed we do mention that adaptation generally outperform VTLN approach, that they both have their advantages and drawbacks and that, ultimately, if you want the best performance you should combined these approaches (if you can allow it computational).

30. Just a comment, but this seems a lot of effort to obtain a VTLN-based system that outperforms adaptation/model selection, and for children the best performance is obtained with model selection augmented with VTLN!

See previous comment.

32. Page 28. Lines 5-7. Augmenting features that have already been VTLN normalised with the posterior 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 believe 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 probabilities 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

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?

The approach for estimating the VTLN scaling factors making use of speaker-independent triphone HMMs with just 1 Gaussian density per state was proposed by Welling et al. in the reference paper:

- L. Welling, S. Kanthak and H. Ney, “Improved Methods for Vocal Tract Normalization”, in Proc. of IEEE ICASSP, 1999, Vol. 2, pp. 761-764.

In the paper it is empirically verified that having 1 Gaussian density per state is a good choice when compared with the use of many Gaussian densities per state.

In past works, we adopted this approach in the context of children’s and adults’ speech recognition with good results, see for example the reference papers:

- M. Gerosa, D. Giuliani and Fabio Brugnara, “Acoustic variability and automatic recognition of children’s speech”, Speech Communication, 2007, Vol. 49, N. 10-11, pp. 847-860.

- M. Gerosa, D. Giuliani and Fabio Brugnara, “Towards age-independent acoustic modeling”, Speech Communication, 2009, Vol. 51, N. 6, pp. 499-509.

Although VTLN scaling factor selection on a speaker-by-speaker basis could be more robust, in this work we adopted a per utterance approach as this approach better matches with a realistic application scenario in which only little data is presented for decoding.

2. The speaker adaptation is achieved by retraining of an age/gender independent DNN with an age/gender specific corpus adaptation. How does this approach compare with the stateoftheart DNN adaptation schemes? For example, using ivectors or speaker codes as additional inputs to the DNN? What if the speakeradapted features (e.g., fMLLR) are used as the DNN inputs?

I-vectors and speaker code as additional input should probably perform better than the approaches presented here. Yet, they require speaker annotations which is out of the scope of the paper. For clarity sake, we decided to focus in a first stage on the age gender groups problem for which require only annotations about the age class (adult, children) and the gender class (male and female). These annotations are far more simple than speaker annotations required by the aforementioned approaches. However, in the future it would indeed be interesting to compare the approaches presented here and speaker based approaches. All this is clarified in introduction and conclusion (see also comment 26 from reviewer 1).

Regarding speaker adapted features such as fMLLR, they are widely used as input in many ASR tools available. However, at this stage the target of this paper is to compare VTLN to approaches only on DNN (that could be inspired by DNN) not to compare several features normalisation approaching which has already been done. Yet, we agree that in later stages it would be interesting to compare results obtained with our best approach to a wider range of normalisation or adaptation techniques.

3. How is system combination performed? I am not sure I understand what the authors mean by features level in section 5.1.4. More implementation details are appreciated. How about the recognition level combination schemes, e.g., confusion network combination? Bayesian risk decoding? etc.

The paragraph about system combination has been modified into:

System combination is a common way to improve systems performance and robustness. It is decided here to combine the different approaches introduced until here to exploit their potential complementarity. In ASR, systems are generally combined using confusion networks or using late fusion at transcription level. The solution chosen here is to either combine the different systems at features level (standard VTLN and the posterior probabilities of the warping factors are combined at the input of the DNN) or at the acoustic model level (acoustic features augmented with the posterior probabilities of the warping factors are used as input to an DNN with age-gender adaptation). A reason to this choice is that this way, the experiments are focus at acoustic model level and remain *independent from any change in the later stage of the decoder or in the language model.*

3 Reviewer 3

Comments III-V

This has been corrected

II. In the under-resourced conditions, the authors should also investigate more adaptation techniques, e.g., regularization, LIN., LON., etc. to see if they perform better. I would like to see more experimental results on more DNN based adaptation approaches.

See comment 26 from reviewer 1 and comment 2 from reviewer.