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

January 5, 2016

1 Reviewer 1

"Sequence trained systems have indeed proved really efficient during the past years." I didn't mean it is efficient. It should serve as your baseline system. If you can not experimentally prove that your proposed approach has significant gains over well-defined baselines, you can't really strengthen your proposal.

Systems trained without sequence training are still widely used as baseline systems (see ICASSP 2015 and Interspeech 2015 for example). Therefore, even if we agree that using sequence training as a baseline could be relevant we do not think this is a mandatory thing. We believe that in our particular case, using a DNN without sequence training can help to keep the message simple and focus on the proposed approaches.

"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." After I went though comment 26 from review 1 and comment 2 from review2, I didn't find the satisfied responses from the authors.

LIN and LON (as I-vector and speaker codes) are speaker dependent transform which are out of focus here as we decided to focus on group of speakers and not on individuals speakers. As for regularization, we could of course use dropout for regularization, early stopping, l1, l2 regularization. All these are relevant and would probably improve performance but this is not the point of the paper. Indeed, for clarity sake, we decided to focus in a first stage on the age and gender groups problem which requires only annotations about the age class (adult or children) and the gender class (male or female).

2 Reviewer 2

Every comment have been addressed.

Maybe (ChildIt + IBN) should be (IBN + ChildIt) for consistency with the previous phrase.

We left this unchanged to keep the format (ChildIt + adult corpus)

P31: l 2: The sentence that begins "This tends to confirm...." needs to be re-written. It's not clear what point is being made.

The sentence has been modified into:

The experiments on children corpus tend to confirm this improvement. Indeed, the systems Warp-post + MFCC *and* VTLN-normalisation *improve the baseline performance by 6% WER relative (from 12.83% to 12.11% with* $p < .001$*) and 5% WER relative (from 12.83% to 12.21% with* $p < .001$*), respectively.*

P33: The second paragraph reads awkwardly. There are too many instances of "approach".

The paragraph has been modified into:

The different approaches presented here have been tested extensively in terms of PER on small corpora first. Systems based on VTLN have been shown to provide a significant improvement compared to the baseline (up to 19% relative) but were still outperformed by the DNN adaptation (23% relative improvement compared to the baseline). The combination of several techniques on the other hand effectively takes advantage of the complementarity of the different approaches introduced in this paper and improves the baseline performance by up to 35% relative PER. Besides, the combination of several techniques is shown to consistently outperform each approach used separately.

3 Reviewer 3

Comments 1-4

This has been corrected.

5. Page 17, 4.2.3. If I understand correctly, "uniform transition probabilities are associated to looped transitions" means there isn't any phone language model in use. Why were bigram or trigram phone LMs not used at all? They should be standard for the phone recognition task.

The purpose of the paper is to compare acoustic models and not to achieve the highest possible phone recognition performance. To this end, we think that the adoption of a phone loop LM, not making use of n-gram probabilities, is adequate. In general, when the purpose is to measure variations in phone recognition induced by changes in the acoustic model some authors prefer a

simple phone-loop LM and others instead n-gram LMs. One reason for using a phone-loop LM is that of avoiding to depend on the quality of the LM.

6. Page 23, 5.1.2. The model selection assumes there exists a perfect age/gender classifier, making it less practical to apply the model adaptation approach in reality. A simple automatic selection approach based on the highest recognition scores among each individual adapted system, as suggested by reviewer 1, could be applied. This should be able to do easily since the authors already decoded each utterance with all 3 models anyway, according to Table 4. Even though the classification might not be perfect, the results should be informative and more directly comparable to VTLN-based approaches which don't assume the age/gender group is known during test time. I think it fits well with the scope of this article and will give a more complete picture of the problem being studied.

We have updated Table 4 with the requested results and modified the related text accordingly:
For comparison purposes, last row of Table 4 Model selection *reports results obtained with an automatic approach for acoustic model selection. In this case each utterance is decoded three times by using each individual group adapted acoustic model and, as final recognition result, the recognition hypothesis resulting in the highest likelihood is retained. Comparing recognition results in the last two rows of Table 4 it is possible to note that the automatic model selection approach results in an overall decrease of performance: from 11.59% to 12.26% PER. This decrease of performance is consistent across the three groups of speakers. It would probably be possible to obtain better model selection for exemple by training a DNN to perform the selection but the is out of the scope of this paper. Therefore, in the rest of the paper,* Model selection *approach is assumed to be the* Model selection (oracle) *approach and recognition experiments are always conducted with matching adapted acoustic models.*

7. Page 24, Table 5. The performance improvement of the DNN-warp based systems over the baseline could also be partially due to more parameters. The authors should also train another baseline system with an 8-layer vanilla DNN-HMM systems and verify if it is worse than the proposed 4-layer DNN-warp + 4-layer DNN-HMM system.

As explained in the previous revision in answer to comment 17 from reviewer 1. The hyper-parameters have been chosen according to the task and the corpus size. These parameters include the number of hidden layers. We have found that on a small corpus as those we are using increasing the number of hidden layers will not improve the performance and that increasing it too much (from 6 layers and more) will actually result in a performance drop. This is a well known effect and the authors believe that such study is not really relevant for the papers and we decided to keep the experiment report really concise and

focused in order to deliver a clear message.

Training a 8-layers vanilla DNN-HMM will therefore results in lower performance than the baseline. We agree with the reviewer's comment. Yet we believe that as the above mentioned effect is well-known in the community it is probably not relevant to add the performance report for a 8-layers vanilla DNN-HMM system.

The choice of the number of hidden layers is now explained in the text:
The values of the hyper-parameters (network topology and learning parameters) are standard values, in the range of the values commonly used for these parameters in the literature. Considering the relatively small size of the corpora, the number of hidden layers is set to 4. Increasing the number of layers with the amount of data available has been observed to provide no significant performance improvement. Besides, training a system with more than 6 hidden layers will result in lower performance than with 4 hidden layers.

Comment 8

This has been corrected.

9. Page 25, line 8. I don't understand why "This latter system however, does not allow for joint optimisation as the averaging operation take place between the DNN-warp and the DNN-HMM." The averaging operation should be differentiable so that the errors can be back-propagated from the DNN-HMM to the DNN-warp for calculating the gradients. In particular, since warp-post(utt) performs better than warp-post according to Table 5, it seems to make more sense to use warp-post(utt) for joint optimisation. But of course, the downside is the two-pass process for decoding. Please provide the justification of not doing that.

What you mention is true with fixed length utterance. The datasets used here have utterances of variable length and back-propagating the gradient is then not trivial in practice. This is the reason why we did not use warp-post(utt) for joint optimisation. This is now clarified:
In this latter system however, the averaging operation over utterances of variable length take place between the DNN-warp and the DNN-HMM. Back-propagating the gradient through the variable length averaging in not trivial to implement in practice. Therefore the system Warp-post(utt) is not used for joint optimisation.

10

See above.

11. Page 26, 5.1.4. "combine ... at features level" seems fine to me since different features are combined. But "combine ... at the acoustic model level" seems a bit confusing since there are no two different acoustic models to be combined - it only means the adapted models use the DNN-warp features. Consider rephrasing it.

This has been rephrased as follows: *It was chosen to either combine the different approaches at features level (standard VTLN normalised features and the posterior probabilities of the warping factors are combined at the input of the DNN) or to use acoustic features augmented with the posterior probabilities of the warping factors are as inputs to a DNN with age-gender adaptation.*

12 - 14
This has been corrected.

15. Page 31, Table 8. Did the authors have the result for just warp-post + VTLN, i.e. not joint optimisation or model adaptation?
We did extensive experiments on PER (including *warp-post + VTLN*) in order to select the systems to be tested on WER as these experiments take much longer. *Warp-post + VTLN* was shown not to improve significantly the performance over *model selection* therefore this combination was not tested on WER. The PER for *warp-post + VTLN* are now included in Table 6 and the corresponding paragraph has been modified as follows:

On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the combination of approaches outperform all the individual approaches presented until here. The combination Warp-post + MFCC (model selection) *improves the baseline by 30% relative (from 14.32% PER to 10.98% PER with* $p < .001$*).* Warp-post + VTLN *improves the baseline by 20% relative (from 14.32% PER to 11.90% PER with* $p < .001$*) and* VTLN (model selection) *improves the baseline by 35% relative (from 14.32% PER to 10.61% PER with* $p < .001$*). The combination of the three approaches presented in this paper (*Warp-post + VTLN (model selection)*) improves the baseline by 34% relative (from 14.32% PER to 10.68% PER with* $p < .001$*). The difference between* VTLN (model selection) *and* Warp-post + VTLN (model selection) *is not statistically significant. When compared to the best system until now (*Model selection*), the combination of different approaches improves from 5% relative (*Warp-post + MFCC (model selection) *with* $p < .001$*) to 9% relative (*VTLN (model selection) *with* $p < .001$*). The combination* Warp-post + VTLN *on the other hand does not significantly improve the performance compared to* Model selection. *Therefore, this approach will not be considered for further experiments.*

16 - 18
This has been corrected.

19. Page 32, paragraph 3, last line. Model adaptation also operates in a two-pass process if automatic model selection is considered the first pass, depending on whether the age/gender group is known during test. This is now mentioned:
The drawback of this approach, however, is that it requires a two-pass decoding

whereas ChildIt + general model *operates in a single-pass granted that the age or gender group group is known during decoding.*