## Natural Language Engineering - Decision on NLE-ARTC-15-0019.R1

**De :** jnle@wlv.ac.uk ven., 04 déc. 2015 18:42
**Expéditeur :** onbehalfof+jnle+wlv ac uk <onbehalfof+jnle+wlv.ac.uk@manuscriptcentral.com>
**Objet :** Natural Language Engineering - Decision on NLE-ARTC-15-0019.R1
**À :** romain serizel <romain.serizel@telecom-paristech.fr>
**Cc :** jnle@wlv.ac.uk

04-Dec-2015

Dear Dr. Serizel,

We have now received all reviewers' reports on your revised paper ID NLE-ARTC-15-0019.R1 entitled "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children". We are pleased to inform you that YOUR PAPER WILL BE ACCEPTED for publication in Natural Language Engineering, PROVIDED THAT you take into account the following remarks and incorporate the very small corrections deemed necessary by the reviewers ( The comments of the reviewer(s) are included at the bottom of this letter).

Some of the criticisms go beyond simple presentation matters, so I would like you to provide us with a list of all the changes that you have made to the paper in response to each criticism. Your paper will then be subject to another short review, to ensure that all the required changes have been completed.

I would be grateful if you could let me know for administrative purposes whether you are prepared to make the suggested amendments, and, if so, when I can expect to receive the revised version.

To revise your manuscript, log into https://mc.manuscriptcentral.com/nle and enter your Author Center, where you will find your manuscript under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number will be appended to denote a revision. You may also click this link to start your revision:

https://mc.manuscriptcentral.com/nle?URL_MASK=41cf000284944c24a296953b1c1a2b93

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. Please use this space to document any changes you make to the original manuscript.

Could you please acknowledge receipt of this e-mail? Thank you very

much in advance.

Kind regards,

Dr. Sara Moze
Research Associate in Lexicography
Editorial Assistant for the Cambridge Journal Natural Language Engineering
Research Group in Computational Linguistics
Research Institute of Information and Language Processing
University of Wolverhampton
MC133
Stafford Street
WOLVERHAMPTON
WV1 1LY
Tel: +441902 322 409

_____

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

Overall, I can not find satisfied responses from the authors w.r.t my previous comments, specifically,

"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.

"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.

Reviewer: 2

Comments to the Author

I have read the revised version of the manuscript and I confirm that the authors have addressed all of the issues that I raised in my first review. I think this is a very interesting paper and also that

it is very timely, in that it deals with the application of DNN-HMM systems for ASR of children's speech. While reading the revised paper I noted a number of instances where grammar could be improved. I have listed these below:
P1: l -5: "... in a single pass when THE VTLN approach ..."
P2: l -8: "As A consequence..." or "Consequently..."
P3: l3 I think "contrast" is the wrong word. Maybe use "counter" or "compensate for"
P4: l6: ".. and generalisation capabilities, DNNs have been..." (delete THE and make "DNN" "DNNs")
P5: last line: "MFCCs" or "MFCC vectors" rather than "MFCC"
P6: l-5: "... Section 2 briefly introduces DNNs" (add "s") and l -4: " ... in ASR and presentS..." (*add "s")
P7: l 5: "DNNs are called deep..." (add "s") and l 7: "... may require a huge of ..." should be " ... may require a huge NUMBER of..." or " ... may require a huge SET of..." and l -4: ".. as the outputS.." (add "s")
P8: l 2: "Following Bayes' rule:" (delete "the" and add an apostrophe after "Bayes") also first line of 2.1 "Training a DNN is a difficult task mainly ..." ("task" not "tasks")
P11: final line: " ... to form an augmented acoustic feature vector" (delete "s")
P13: 4.1: l 2: " ... consisting of children's speech ..." (I believe this is correct. I know that we talk about "adult speech" but I think that the equivalent for children is "child speech" I think that "children speech" is wrong. It certainly sounds wrong! Two lines later "while the ChildIt and the APASCI corpora provide A similar amount..." or "while ChildIt and APASCI provide A similar amount..."
P16: l 5: " ... tied state triphone HMMs based ... "
P18: l 8: maybe (ChildIt + IBN) should be (IBN + ChildIt) for consistency with the previous phrase. Also 4.3.3 l 2: "... domain consisting OF about..." Also 3 lines from bottom " .... a group specific corpUS.." or ... group specific corpora..."
P19 (and elsewhere) l 11: "features vector" is wrong. Use "feature vector" for singular or "feature vectors" for plural. This occurs several times in the manuscript.
P20: l5: " ... in aN attempt..." Also, 2 lines from bottom "... with A limited amount of training data ... " or "... with limited training data ..." Also, final line "... to VTLN for A/THE DNN-HMM framework ... "
P21: 5.1.1: "In this experiment, DNNs..." (add "s") Also, two lines later, " ... and performanceS are compared ... " or "... performance is compared ..."
P22: l-4: "by is"? Also, final line, "From THE results in ..."
P23: l 1: "... adapted to children's voices performS much better ..."
P24: l -9: "This experiment allowS..." l -7: "The impact of having A hard or soft..."
P25: l -13 "... each group of speakerS" Also, l -7 "children's speech"
P26: L 2: "..with the hard decision THAT is the standard..." l 5 " ... constraints are not compatible..." (delete "s") l 8: "... introduced until here..." maybe replace with "... introduced up until this point..." l -10 " ... and THE combination ..."

P28: l 8: " .. the combination of several approaches improves... " (delete "allows to")  l 11: sme thing again.
P29: l 2: "... with the performance obtained with A system..." and on the next line " ... age adapted systemS for child speakers and ..."  l -7: " ... than for THE experimentS on ... "This allows affective training to be achieved on the adult groups" is better than "This allows to achieve an effective training on the adult groups"
P30 l 1: "... VTLN based approaches:  VTLN applied..." (delete "the").  l 9 "features vector" (see previous comment) l -2 "(11.57% WER and 11.58% WER ..." or "(11.57% and 11.58% WER ...".
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.  l 6: "THE performance differenceS ..."   l -4 " ... adult corpora .. " or " ... adults' corpora ..."
P32: l 2: "speakers' group" or "speaker group"   l -5: " ... operateS in A single pass"    l -2: "Two different kinds of approach have ..." (delete "s")
P33: The second paragraph reads awkwardly.  There are too many instances of  "approach".  l -11: "The trend observed on PER continues/persists..."?

Reviewer: 3

Comments to the Author

In general, this is an interesting article, which shows 1) VTLN warping factors posterior vector learned by a DNN can serve as an effective form of age/gender code when combined with raw MFCC features, which performs as well as VTLN-normalized MFCC as the input to a DNN acoustic model. 2) one existing DNN model adaptation approach seems to be able to compensate age/gender group variations well without VTLN, and even better with VTLN information. Both the VTLN-based and model adaptation approaches seem to be mainly useful in the limited resource scenario, whereas DNN appears to learn speaker variations directly from all of the data if they are sufficient.

Extensive experiments are conducted, which are well designed for the most part. However, the model selection presented for adapted systems is not automatic. Also it would be interesting to see how the proposed VTLN posterior feature is compared against standard speaker ID models/features like i-vectors, fMLLR.

The text is clearly written and well organized. But it needs to be proof read more carefully for grammatical corrections.

Below are some detailed comments.

1. Page 2, last paragraph, line 1. Define "ASR" - first occurrence.

2. Page 4, line 1. Gaussian mixture modelisation -> Gaussian mixture model.

3. Page 5, paragraph 3. not be able reach -> not be able to reach

4. Page 7, line 7. model any function they may require -> model any function, they may require

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.

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.

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.

8. Page 24. The impact of having of hard or soft decision -> The impact of having hard or soft decision

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.

10. Page 26, paragraph 1, last line. compatibles -> compatible. Also, I don't understand why the two constraints are not compatible, as argued in comment 9.

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.

12. Page 28, paragraph 1. Twice: the combination of several approaches allow to improves ... and improves ... -> different combinations of several approaches allow to improve ... and improve ...

13. Page 28, 5.2, 1st paragraph. in terms WER -> in terms of WER

14. Page 31, Table 8. Twice: Warp-post - VTLN (...) -> Warp-post + VTLN (...)

15. Page 31, Table 8. Did the authors have the result for just warp-post + VTLN, i.e. not joint optimisation or model adaptation?

16. Page 31, line 2. on children corpus -> on the children corpus

17. Page 31, paragraph 2. During these experiment, the corpus was unbalanced ... -> During these experiments, the corpora were unbalanced ...

18. Page 32, paragraph 2, last line. confirm -> confirms

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.

---