**Zimbra** **rserizel@enst.fr**

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

**De :** jnle@wlv.ac.uk mar., 30 juin 2015 17:36

**Expéditeur :** onbehalfof+jnle+wlv ac uk <onbehalfof+jnle+wlv.ac.uk@manuscriptcentral.com> 1 pièce jointe

**Objet :** Natural Language Engineering - Decision on NLE-ARTC-15-0019

**À :** romain serizel <romain.serizel@telecom-paristech.fr>

**Cc :** jnle@wlv.ac.uk

30-Jun-2015

Dear Dr. Serizel,

We have now received all reviewers’ reports on your paper ID NLE-ARTC-15-0019 entitled "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children".

As you can see from the reviews enclosed, the reviewers identified a number of issues in your article.
As a consequence, we cannot accept your paper as it stands; however, WE ENCOURAGE YOU TO REVISE YOUR ARTICLE according to reviewers' feedbacks and RESUBMIT IT. (The comments of the reviewer(s) are included at the bottom of this letter).

If you decide to resubmit, please enclose a list of the points that have been addressed, and the criticism which they answer.

Once we have received your resubmission, your re-submission will be checked and will also undergo additional reviews which may require further modification. This is to ensure the paper meets the standards of the journal.

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=ab1b35bfa13e40a0a57188122530fa37

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?

We hope that the reviewers' feedback will be useful for improving your work.

Thank you again for submitting your article to JNLE.

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
This is an interesting paper about the application of HMM-DNN based speech recognition to child, adult male and adult female speech. Specifically the paper is concerned with the interaction between DNN training, DNN adaptation to a particular speaker group (children, adult females, adult males) and explicit vocal tract length normalisation (VTLN). The paper compares the performances of a baseline system (a HMM-DNN system trained on data from all three groups), an "adapted" system, in which the baseline is subject to further training using only data from one of the groups, VTLN pre-processing of the acoustic data prior to DNN training and recognition, and a system in which VTLN warping factor posterior probabilities are estimated using a DNN, and the result is appended to the acoustic feature vector input to the main DNN. The Gillick-Cox test is used to assess the statistical significance of the results. Finally, combinations of these different techniques are tried to expoit their complememary nature.
I think this is an interesting paper, and that the most interesting outcome is the extent to which DNN adaptation, especially to children's speech, works almost as well as all of the techniques that

include some form of VTLN. Reading the paper, though, my feeling was that the authors are trying very hard to get the best performance out of one of the approaches that uses explicit VTLN!

Here are some more detailed comments:

1. Page 2, first line: "one of the major sources" or "a major source"
2. Page 3, line 1, "... reported in the literature ..."
3. Page 3, line 17 "Assuming that there is enough training data..."
4. Page 7. 3rd equation. How is the quantity p(X) calculated?
5. Page 8, line 1. " limitation has been partly overcome"
6. Page 8, line 6. "... better starting point than..."
7. Page 10. ""Similarly to the method proposed in Seide et al." is better than "Similarly to as proposed in Seide et al."
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.
9. Page 11. line 2 "..., acoustic feature vectors..." (repeated in next sentence). Also state explicitly that "normalised" means "VTLN normalised".
10. Page 12. The first sentence is not grammatical and needs attention. 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.
11. Page 12, line 8. "...and the DNN-HMM are concatenated..." and later in the same sentence "fined tuned" rather than "fine tune".
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. $.1.1 and 4.1.2 say something about the transcriptions of the data that are available - word-level or phone-level?
14. Page 14. Please include a reference for HLDA.
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)
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?
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 tha HMM set used for word recognition use the same set of tied states as the previous HMM set?
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.
21. Page 17, 4.4. In adaptation are all of the other DNN training parameters the same as in 4.2.2
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.
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.
25. Page 18. Where does the learning rate of 0.0002 come from?
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.
27. Page 22. In the earlier overview of the systems, did you include the case of MFCC feature vectors augnmented with VTLN warping factors obtained in the normal way?
28. Page 22. Exactly how are the posterior probabilities averaged to utterance level, and why is this done?
29. Page 23. Final paragraph. This is an example of what I was indicating earler. The first line of the final paragraph is very "pro-VTLN" and ignores the fact that better performance is obtained by simple adaptation.
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 perofrmance is obtained with model selection augmented with VTLN!
31. Page 27, line 1 of first paragraph. Should it be "age-gender" adaptation or "age" adaptation in this case? Also, "for a child speaker" is better than "for children speaker".
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?

Reviewer: 2

Comments to the Author
Please see attached document for my comments. Thanks.

Reviewer: 3

Comments to the Author
In this article, the authors investigate DNN adaption for children/ male and female adults speech recognition, and proposed to use a dedicated DNN to estimate VTLN factors which are further used as auxiliary features to the primary feature (MFCCs). Experimental results show the VTLN DNN-estimated features with or without joint training both help the system and with them there is no need for second-pass decoding. My specific comments are as follows,

I. Why the baseline DNN system is not sequence trained., eg. bMMI, sMBR which is now standard DNN based acoustic modeling techniques and should be used as baselines.

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.

III. In page 7, in the equation, $\sigma(h) = 1/(1+exp(-w y^{-b}))$, $h$ is not a variable, please revise the equation in the right side.

III. In page 7, in the equation $softmax(h)_j$, here it should be written as $softmax(h_j)$

IV. In page 7, in the equation $p(X|S) \propto p(S|X) / p(X)$, here $p(X)$ should be $p(S)$

V. To be consistent with the notation DNN-HMM, you should use GMM-HMM instead of HMM-GMM though the article.

---

**Paper Review.pdf**
47 ko

---