Review of Doctoral Thesis

Submitted by:	Jan Trmal
Title:	Spatio-Temporal Structure of Feature Vectors in Neural Network Adaptation
Reviewer:	František Grézl

The thesis investigate into the adaptation of neural network on very little data. The key points are (1) definition of adaptation problem and proposed solution by linear adaptation, (2) assumption about neural network input which allows reduction of free parameters and (3) experimental validation of proposed technique. The thesis has 104 pages divided into 10 chapters. This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents overall conclusion and recommendation to the committee.

Technical content of the thesis and remarks to chapters

Chapter 1 briefly introduces into speech recognition and points towards the issue of the thesis.

Chapter 2 specifies the goals of the thesis.

These two chapters are very short (about one page each) and could be merged.

Chapter 3 presents speech recognition system. First, several feature extraction techniques are described - standard MFCC feature extraction is followed by TRAP feature extraction including its several modification. Finally, Bottle Neck feature extraction is outlined. Then the acoustic modelling using HMM is introduced followed by definition of GMM and neural network observation probability functions. Next, language modelling is shown including N-gram and neural network based language models. Finally, speech decoding technique is described and discussion about evaluation of results is given.

This chapter can be seen as general introduction into speech recognition. But it goes into too much detail in case of description of concatenation of elementary HMM models into a chain of these models. Also, the description of neural network based language models seems to be redundant as they are not used in the work. A note with citation would be fine. On the other hand, the reader is missing the information about motivation of the author that is sentences like "we will use there technique further".

Chapter 4 is devoted to artificial neural networks starting with biological inspiration, going through definition of perceptron unit to end-up with feed-forward and recurrent neural networks. Great part of this chapter is dedicated to training of multilayer perceptron networks with detailed description of error back-propagation. Different methods for speeding up the training process are described and discussed.

Again, the work goes into direction which will not be explored further in case of recurrent neural networks and a specification of what will be used is missing. This would be helpful especially in case of discussion of techniques for training speed up.

Chapter 5 describes the training of speech recognition system with regard to speaker adaptation. The parameter normalization is discussed first, followed by acoustic model adaptation and finished

with speaker adaptive training. Discussion of presented techniques from the hybrid recognition system and neural network adaptation point of view should be given.

Chapter 6 gives overview of current approaches to adaptation of neural network. Variety of techniques are introduced and their properties are discussed.

Chapter 7 describes proposed method for neural network adaptation. The solution takes into account the neural network input features which allows to reduce the number of free parameters in the adaptation matrix. Another parameter is employed to reduce the number of free parameters further. This allows to train the adaptation matrix on very little data.

The amount of parameters is given as 2^{17} - how much is it? Is the work supposed to be human or computer readable?

Chapter 8 provides information about three databases the experiments will be performed on: the Czech SpeechDat(E), TIMIT and Wall Street Journal phase 0. The phoneme sets are described together with modifications made. Data used for training of phoneme language models are specified and the n-gram coverage given is provided. For Wall Street Journal database, also the word language models were trained.

The reduction of phoneme set was done SpeechDat data set in such a way, that long and short vowels were merged together. This might be fine for standard MFCC features, but for features which capture long temporal context, there is quite difference between long and short vowels. Were some comparison experiments done regarding this phoneme set reduction?

Chapter 9 is dedicated to the experimental evaluation of proposed technique. In the first section, preliminary experiments are done to verify the proper function of the adaptation technique. Further experiments are done on WSJ data set. The reference (MFCC based) system is trained and results are compared to other published ones. Synoptic table would greatly help to orient the reader. Then the baseline neural network based system is trained and results reported. The following sections are dedicated to individual scenarios of speaker adaptive training - supervised, semi-supervised and unsupervised ones.

Section 9.1.1 mentions experiments with different structures and number of parameters in neural network, but no further information is provided, neither the results are given.

When comparison with BUT system is given, the differences in phoneme set are mentioned. A table with used phoneme sets would be helpful. Also, the used phoneme language models are not mentioned. Were LMs used? What order?

The comparison with VTLN technique (used method for VTLN factor estimation is not given) is done on development version of the software. Is it possible to rerun the experiments with the final version to allow for comparison of final results?

The results for WSJ are reported as "accuracy per sentence". I must admit that I have never encounter such measure. The definition of this measure is not given thus it is not clear what the reported results actually are. Is it average accuracy per sentence? Percentage of correctly recognized sentences?

The confidence intervals are given for each experiment, however the number of repetition of the experiments is not given. Neither is described how exactly the data were reordered. The only hint is section 3.5.3 with general description of the methodology.

How are the values of μ in figures 9.2 - 9.4 related to the values in corresponding tables 9.4, 9.8 and 9.9?

Table 9.6. presents the results of rescoring lattices with another LM. What is the meaning of rescoring lattice with lower n-gram LM than it is was generated with?

The result for optimal configuration of MELT and CF factor should be added to table 9.12 to see the difference in performance. The discussion on additional improvement and computation

demand should be given.

The results using standard VTLN technique are missing for TIMIT and WSJCAM0 data sets. This is very important comparison as the proposed MELT technique does "the same" thing. Results and computation demand and other advantages/disadvantages of both techniques should be discussed.

Chapter 10 concludes the work and points the future ways of research.

Summary of technical content of the thesis

The thesis clearly demonstrates the qualities of the candidate - capability to study non-trivial literature, suggest own novel solutions, implement implement and test them. It is also evident that the candidate is skilled in software implementation. The critical points are the following:

- Sections which are not related to the problem solved in the thesis. A note with the pointer to literature would be fine.
- The developed neural network training software is not described at all. As its implementation is one of the goals of the thesis, a chapter/section devoted to it is expected. Also, comparison of the software with other existing and freely available tools (QuickNet, SNet, ...) should be made in terms of achieved classification accuracy and time consumption. Other advantages should be discussed.
- The comparison with standard VTLN technique is done only for one data set and with development version of the software only. However this comparison is essential as both techniques do "the same" thing. Considering other adaptation techniques suitable for the given framework would be beneficial. Comparison of individual techniques in terms of achieved improvements and computational demands would increase the benefits of the thesis.
- It seems that the precise and careful work presented in the theoretical part of the work faded out in the experimental part. I am missing proper discussion of the results. It seems that the end of the work was done in hurry.

Note on goals fulfilment

Comments on the goal of the thesis as stated in sec. 2.1:

1. Fulfilled completely
2. Probably fulfilled. The software is not described but probably exists as the candidate had to obtain the results somehow.
3. Fulfilled completely.
4. Fulfilled completely. One should be careful with the statements like "current techniques do not work..." specially when "current" is not defined. The literature overview mentions a publication solving the same problem by the same approach (e.g. linear adaptation matrix on the input of the NN) with good results.
5. Fulfilled under reservations - the comparison with at least standard VTLN technique should be given.
6. Fulfilled under reservations - the comparison with at least standard VTLN technique should be given.

Note on results and original contribution of the thesis

As main results and contributions of the thesis are, in my opinion, the detailed literature survey and proposed solution of NN adaptation problem for a particular situation (specific NN input). The NN training software seems to be proprietary thus cannot be counted as contribution to speech recognition community. Which in turn also limits the potential impact of the adaptation method.

Note on the candidate's publication

Publication demonstrate the progress of the work.

Comments on the formal aspects

As far as I can judge, the thesis is written in proper English, easy to read and follow, with reasonable number of typos. Sometimes there is too detailed description of an negligible problem or sections devoted to problems not related to the solved problem. Motivation is sometimes missing in the text, especially when several possible solutions is presented. To state which method will be further used would be helpful.

The equations are carefully type-set, however different type of indices (or symbols used in them) in the text and equation is confusing.

The captions of tables should contain the information about the measure WER/PER/WACC... one compact table comparing results over all experiments should be given in conclusion section or the baseline results should be kept over all tables.

Summary and recommendation

I have carefully examined the doctoral thesis of Mr. Jan Trmal. Despite the criticism raised above, in my opinion, it is a solid work that contributes to progress in speech recognition domain, especially in employment of neural networks into recognition systems. However, I would recommend the following:

- Include a chapter/section on the developed NN training software.
- Compare the results to standard VTLN technique.

To conclude, I do recommend the Thesis as a partial requirement for granting Mr. Jan Trmal the doctoral degree at University of West Bohemia

In Brno, 13.6.2012

Ing. František Grézl, Ph.D.
Brno University of Technology, Faculty of Information Technology
Božetěchova 2, 61266 Brno, Czech Republic
tel: +420-54114 1280, fax: +420-5-41141290
email: grezl@fit.vutbr.cz

Západočeská univerzita v Plzni, fakulta aplikovaných věd, katedra kybernetiky

Posudek na disertační práci k získání titulu doktor v oboru Kybernetika
Ing. Jana Trmala

Využití prostoro-časové struktury příznakových vektorů pro adaptaci neuronových sítí

Ing. Jan Trmal se ve své disertační práci zabývá metodikou adaptace neuronových sítí a řečníků. Pro řešení zvolil adaptivní trénování neuronových sítí pro akustické modelování. Práce je příspěvkem k řešení problematiky automatického rozpoznávání řeči.

Téma zvolené pro disertační práci je velmi aktuální. Řeč a její zpracování je otázkou multidisciplinární. Vyplývá to z podstaty řeči. Proto se touto oblastí zabývají intenzivně mnohá výzkumná pracoviště na celém světě. Velký význam předložené disertační práce je nejen v množství experimentů, ale také v tom, že rozvíjí a dále posunuje řešení problému, který spadá do jednoho z témat, kterými se již mnoho let úspěšně zabývají členové katedry kybernetiky na Fakultě aplikovaných věd ZČU v Plzni. Disertační práce byla řešena za finanční podpory grantů MŠMT.

Použité metody jsou moderní, podle dostupných pramenů dosud pro češtinu nepoužívané. Jsou založeny na rozsáhlých znalostech v mnoha oborech, jsou podepřeny velmi dobrým matematickým zázemím. To všechno a ještě velké množství časově náročné a nepopulární monotónní práce při experimentech bylo základem celé disertační práce. Výsledkem je splnění všech vytčených cílů.

Předkládaná práce je psána anglicky, má rozsah 118 stránek, je členěna do 10 kapitol a je doplněna rozsáhlým seznamem prostudované literatury (obsahuje 124 položek) a seznamem vlastních prací nebo prací, v nichž je spoluautorem (16), abstraktem v obou jazycích, obsahem, seznamem obrázků, tabulek a použitých zkratek. V seznamu vlastních prací je 9 publikací, na kterých se podílelo více autorů a ze kterých není vidět spoluautorství disertanta. Uvítala bych v těchto případech konkrétní označení části publikace, u které je skutečně autorem. Chybí také publikace, kde je jediným autorem. Podle mého soudu však Ing. Trmal publikoval dostatečný počet prací na prestižních mezinárodních konferencích, workshopech a v časopisech, což dokazuje nejenom autorovu schopnost vědecky pracovat, ale i schopnost informovat odbornou veřejnost o dosažených výsledcích.

V práci se objevuje určitá nevyváženost ve způsobu a podrobnostech v popisu metod. Mám na mysli především části o umělých neuronových sítích. Na jedné straně v přehledu metod založených na ANN chybí podrobnější popis architektury a použitého algoritmu učení (kapitola 3), na druhé straně popisuje velmi podrobně (5,5 stránky) základní algoritmus učení zpětného šíření chyby (BPG) i přes to, že tento algoritmus je dnes už popsán v mnoha publikacích a disertant tuto základní variantu BPG učení nepoužívá (kapitola 4). Takový popis bych očekávala v diplomové práci, ale ne v práci disertační. Může to být způsobené tím, že víc než polovina uvedených pramenů pochází z 90-tých let, kdy aplikace ANN nebyly ještě tolik rozsirene a algoritmy učení byly méně známé. Od té doby se však metodika neuronových sítí rozvíjela velmi rychle, stejně tak rostl velmi rychle počet publikací v tomto oboru.

Po formální a technické stránce je předkládaná práce na velmi dobré úrovni. Je psána přehledně, autor prokázal schopnost pracovat tvůrčím způsobem. Také logická stavba práce je na velmi dobré úrovni. Vysokou kvalitu poněkud snižuje příliš časté používání zkratek,

z nichž některé nejsou v textu vysvětleny ani se nevyskytují v Seznamu zkratek nebo jsou vysvětleny až v dalších opakováních. Tento fakt působí negativně na čtenáře disertační práce. Podobně to platí u značení proměnných, které není jednotné. Pravděpodobně je ponecháno stejné, jako v původní publikaci, ze které je převzato. Je sice pravda, že v oblasti ANN neexistuje norma pro značení proměnných, pro lepší srozumitelnost při studiu předložené disertační práce by bylo vhodnější, aby autor značení proměnných u různých variant učení sjednotil. Při čtení textu jsem měla místy dojem, že je text zkracován a některé části jsou vynechány.

Při posuzování jazykové úrovně mám připomínku k abstraktu v českém jazyce. Autor občas používá slovosled a slovní tvary, které jsou chybné z pohledu české gramatiky. Je zřejmé, že se jedná o špatný překlad z angličtiny, ve které je disertační práce napsána. K úrovni angličtiny se nebudu vyjadřovat, nejsem rodilá mluvčí.

K práci mám několik připomínek a dotazů. Mezi připomínky patří např. konstatování, že:

- Publikace [5] a [8] uveřejněné v Lecture Notes in Computer Science 6231 (2010) mají uvedeno jednou ISSN, podruhé ISBN a v publikaci [15] je časopis se stejným ISSN uveden jako Lecture Notes in Artificiel Inteligence. Bylo by vhodné způsob citací sjednotit.
- Pro zpracování řeči je vhodnější uvažovat rekurentní vazby mezi výstupní a skrytou vrstvou, mezi výstupní a vstupní vrstvou resp. mezi skrytou a vstupní vrstvou.
- Str. 26, 1.věta odstavce 4.5. – Přenosová funkce u MLP může být lineární, jediná podmínka pro volbu přenosové funkce u tohoto učení je, aby byla diferencovatelná a vstupní data musí být separabilní. Lineární přenosová funkce se často u reálných aplikací používá ve výstupní vrstvě.
- Str. 27 – Píšete, že použitím funkce „tansig" trenink rychleji konverguje. To není tak jednoznačné. Počet matematických operací v algoritmu učení, které se nastavují a ovlivňují jednak rychlost konvergence, jednak úspěšnost natrénování, je u této funkce větší (např. výpočet 4 exponenciál u „tansig" oproti 1 exponenciále u „logsig". Výpočet v jedné epoše tedy trvá déle. Navíc ne vždy je rychlost konvergence nejdůležitějším parametrem, který sledujeme. Běžně používané varianty BPG algoritmu u MLP vedou většinou k nalezení pouze lokálních minim a při velké rychlosti konvergence může být „přeskočeno" nejlepší lokální minimum.
- Str. 29, paragraf 4.6.2. – při obecném popisu trénování je nutné uvažovat i výpočet prahů (ne pouze synaptických vah).
- Str. 55 – v rovnicích (6.10) a (6.11) jsou matice opravdu stejné? Není chyba ve značení?
- V textu není odkaz na obr. 6.2. (str.56).
- Str. 82, paragraf 9.1.1. – Postup, který uvádíte v odstavci „Regularization Influence" je známý z metod klestění ANN a dnes už je , paragraf součástí některých variant modifikovaného algoritmu učení BPG.
- Newrozumím tomu, co plyne z textu na str. 88, paragraf 9.3.1. Zdá se mi opět, že text je vytržen z kontextu nějaké větší práce. Bylo by vhodnější uvádět posloupnost dosažených výsledků v uvedených publikacích chronologicky. Pak by bylo zřejmé, že novější metody vycházejí ve srovnání jako lepší.

Dotazy k práci mám následující:

1) Co je „aktivační hodnota skryté a výstupní vrstvy"? Co je „malý počet neuronů ve 2. skryté vrstvě"? – str. 11

2) Vysvětlete podrobněji, jak funguje proces Parallel Hidden network. Z obr. 6.1, na který se odkazujete, není zřejmý rozdíl od standardní metody (str. 55). K modifikaci parametrů dochází jen v paralelní vrstvě? Nejedná se o částečně rekurentní Elmanovu síť (s rekurencí mezi výstupní a skrytou vrstvou)?
3) Vysvětlete tvorbu požadovaných hodnot (target) – str. 51. Podle Vás původní vstupní vektor projde v dopředném směru na výstup a stane se targetem. Jedná se tedy pouze o jeden průchod sítí, bez modifikace vah pomocí algoritmu BPG? Vstupní vektory pro určení targetu jsou generovány náhodně? Uvažoval jste vliv koartikulace v řeči? Vámi citovaný autor (Anthony Robins) použil tento způsob také pro trénování řečového signálu?
4) Vysvětlete princip Wilcoxonova testu (str.84,85). Co znamená koeficient „α"? Na str. 45 je nazýván „warping coef. a jeho hodnota je v rozmezí <0,88; 1,12>, ale na str. 84 má hodnotu 0,003.
5) Jak vypadá architektura sítí cANN(0) a cANN(k)? Co je inicializace SPK(0) na str. 92?

Na závěr konstatuji, že i přes uvedené nedostatky, které nejsou zásadního rázu, Ing. Jan Trmal projevil schopnost samostatně vědecky pracovat. Oceňuji, že se nespokojil pouze s přebíráním a opakováním vžitých závěrů, ale má odvahu hledat nové směry a metody ve výzkumu. To považuji za velmi významnou vlastnost vědeckého pracovníka. Hledání nových postupů je ale podloženo usilovnou prací a dosáhl velmi dobrých výsledků.

Na základě prostudování práce mohu konstatovat, že **vytčené cíle byly splněny**. Proto **disertační práci doporučuji k obhajobě a po jejím úspěšném obhájení k udělení titulu Ph.D.**

Jana Tučková

V Praze, 24.10.2012

Prof..Ing. Jana Tučková, CSc.
Katedra teorie obvodů
FEL ČVUT v Praze