Deep neural network adaptation for children's and adults' speech recognition

Romain Serizel and Diego Giuliani
HLT research unit
Fondazione Bruno Kessler (FBK)
Trento, Italy
(serizel,giuliani)@fbk.eu

Abstract

English. This paper introduces a novel application of the hybrid deep neural network (DNN) - hidden Markov model (HMM) approach for automatic speech recognition (ASR) to target groups of speakers of a specific age/gender. We target three speaker groups consisting of children, adult males and adult females, respectively. The group-specific training of DNN is investigated and shown to be not always effective when the amount of training data is limited. To overcome this problem, the recent approach that consists in adapting a general DNN to domain/language specific data is extended to target age/gender groups in the context of hybrid DNN-HMM systems, reducing consistently the phone error rate by 15-20% relative for the three different speaker groups.

Italiano. *Questo articolo propone l'applicazione del modello ibrido "rete neurale artificiale multistrato - modelli di Markov nascosti" al riconoscimento automatico del parlato per gruppi di parlanti di una specifica fascia di età o genere che in questo caso sono costituiti da: bambini, maschi adulti e femmine adulte. L'addestramente della rete neurale multistrato si è dimostrato non sempre efficace quando i dati di addestramento erano disponibili solo in piccola quantità per uno specifico gruppo di parlanti. Per migliorare le prestazioni, un recente approccio proposto per adattare una rete neurale multistrato pre-addestrata ad un nuovo domino, o ad una nuova lingua, è stato esteso al caso di gruppi di parlanti di diverse età e genere. L'adozione di una rete multistrato adattata per ciascun gruppo di parlanti ha consentito di ottenere una riduzione dell'errore nel riconoscimento di fonemi del 15-20% relativo per ciascuno dei tre gruppi di parlanti considerati.*

1 Introduction

Speaker-related acoustic variability is one of the major source of errors in automatic speech recognition. In this paper we cope with age group differences, by considering the relevant case of children versus adults, as well as with male/female differences. Here DNN is used to deal with the acoustic variability induced by age and gender differences.

When an ASR system trained on adults' speech is employed to recognise children's speech, performance decreases drastically, especially for younger children (Wilpon and Jacobsen, 1996; Das et al., 1998; Claes et al., 1998; Potamianos and Narayanan, 2003; Giuliani and Gerosa, 2003; Gerosa et al., 2007). A number of attempts have been reported in literature to contrast this effect. Most of them try to compensate for spectral differences caused by differences in vocal tract length and shape by warping the frequency axis of the speech power spectrum of each test speaker or transforming acoustic models (Potamianos and Narayanan, 2003; Das et al., 1998; Claes et al., 1998). However, to ensure good recognition performance, age-specific acoustic models trained on speech collected from children of the target age, or group of ages, is usually employed (Wilpon and Jacobsen, 1996; Hagen et al., 2003; Nisimura et al., 2004; Gerosa et al., 2007). Typically, much less training data are available for children than for adults. The use of adults' speech for reinforcing the training data in the case of a lack of children's speech was investigated in the past (Wilpon and Jacobsen, 1996; Steidl et al., 2003). However,

This work was partially funded by the European project EU-BRIDGE, under the contract FP7-287658.

in order to achieve a recognition performance improvement when training with a mixture of children's and adults' speech, speaker normalisation and speaker adaptive training techniques are usually needed (Gerosa et al., 2009).

During the past years, DNN has proven to be an effective alternative to HMM - Gaussian mixture modelisation (GMM) based ASR (HMM-GMM) (Bourlard and Morgan, 1994; Hinton et al., 2012) obtaining good performance with context dependent hybrid DNN-HMM (Mohamed et al., 2012; Dahl et al., 2012).

Capitalising on their good classification and generalisation skills the DNN have been used widely in multi-domain and multi-languages tasks (Sivadas and Hermansky, 2004; Stolcke et al., 2006). The main idea is usually to first exploit a task independent (multi-lingual/multi-domain) corpus and then to use a task specific corpus. One approach consists in using the different corpora at different stages of the DNN training. The task independent corpus is used only for the pretraining (Swietojanski et al., 2012) or for a general first training (Le et al., 2010; Thomas et al., 2013) and the task specific corpus is used for the final training/adaptation of the DNN.

This paper introduces the use of the DNN-HMM approach for phone recognition in age and gender dependent groups, extending the idea introduced in (Yochai and Morgan, 1992) to the DNN context. Three target groups of speakers are considered here, that is children, adult males and adult females. There is only a limited amount of labeled data for such groups. To overcome this problem, a DNN trained on speech data from all the three groups of speakers is adapted to the age/gender group specific corpora. First it is shown that training a DNN only from a group specific corpus is not effective when only limited labeled data is available. Then the method proposed in (Thomas et al., 2013) is adapted to the age/gender specific problem.

The rest of this paper is organized as follows. Section 2 introduces the general training and adaptation methods. Experimental setup is described in Section 3 and results are presented in Section 4. Finally, conclusions are provided in Section 5.

2 DNN training and adaptation

In ASR what is called DNN is a feedforward network with at least one hidden layer (generally more than three). When applied in a hybrid context, the DNN is used to classify the acoustic features into HMM states. The output of the DNN is then used to estimate the HMM's state emission likelihoods. Recent experiments exhibit that DNN-HMM provides better performance for ASR than shallow networks (Dahl et al., 2012).

2.1 Age/gender independent training

The general training procedure described above can be applied, by using all training data available, in an attempt to achieve a system with strong generalisation capabilities. Estimating the DNN parameters on speech from all groups of speakers, that is children, adult males and adult females, may however, have some limitation due to the inhomogeneity of the speech data that may negatively impact on the classification accuracy compared to group-specific DNN.

2.2 Age/gender adaptation

ASR systems provide their best recognition performances when the operating (or testing) conditions are consistent with the training conditions. To be effective, the general training procedure described above requires that a sufficient amount of labeled data is available. Therefore, when considering training for under-resourced population groups (such as children or males/females in particular domains of applications) it might be more effective to train first a DNN on a large amount of data (including the target group specific corpora) and then to adapt this DNN to the group specific corpora. A similar approach has been proposed in (Thomas et al., 2013) for the case of multilingual training. Here the language does not change and the targets of the DNN remain the same when going from age/gender independent training to group specific adaptation. The DNN trained on speech data from all groups of speakers can then be used directly as initialisation to the adaptation procedure where the DNN is trained to convergence with back-propagation only on group specific corpora.

3 Experimental setup

3.1 Speech corpora

For this study we relied on two Italian speech copora: the ChildIt corpus consisting of children speech and the APASCI corpus consisting of adults' speech. Both corpora were used for evaluation purposes, while the ChildIt and the APASCI provided similar amount of training data for children and adults, respectively.

3.1.1 ChildIt

The ChildIt corpus (Giuliani and Gerosa, 2003; Gerosa et al., 2007) is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. The overall duration of audio recordings in the corpus is 10h:48m hours. Speech was collected from 171 children. The corpus was partitioned into: a training set consisting of data from 115 speakers for a total duration of 7h:15m; a development set consisting of data from 14 speakers, for a total durations of 0h:49m; a test set consisting of data from 42 speakers balanced with respect to age and gender for a total duration of of 2h:20m.

3.1.2 APASCI

The APASCI speech corpus (Angelini et al., 1994) is a task-independent, high quality, acoustic-phonetic Italian corpus. APASCI was developed at ITC-irst and consists of speech data collected from 194 adult speakers for a total durations of 7h:05m. The corpus was partitioned into: a training set consisting of data from 134 speakers for a total duration of 5h:19m; a development set consisting of data from 30 speakers balanced per gender, for a total durations of 0h:39m; a test set consisting of data from 30 speakers balanced per gender, for a total duration of 0h:40m.

3.2 ASR systems

3.2.1 DNN-HMM

The DNN use 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. The context spans on a 31 frames window on which Hamming windowing is applied. This 403 dimensional feature vector is then projected on a 208 dimensional feature vector by applying Discrete Cosine Transform (DCT) and normalised before being used as input to the DNN. The targets of the DNN are the 3039 tied-states obtained from a HMM-GMM system trained on the mixture of adults' and children's speech (ChildIt + APASCI). The DNN have 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 3039.

The DNN are trained with the TNet software package (Vesely̌ et al., 2010). The DNN weights are initialised randomly and pre-trained with Restricted Boltzmann Machines (RBM) (Hinton et al., 2006; Erhan et al., 2010) with mini-batch size of 250. For the back propagation training the starting learning rate is 0.02 and the mini-batch size is 512. In both pre-training and training, a first-order momentum of 0.5 is applied.

The DNN are trained either on all speech data available (ChildIt + APASCI) or on group specific corpora (ChildIt, adult female speech in APASCI, adult male speech in APASCI).

3.2.2 Age/gender adapted DNN for DNN-HMM

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). The DNN is then trained with back propagation on a group specific corpus (ChildIt, adult female speech in APASCI and adult male speech in APASCI). The learning rate follows the same rule as above.

4 Experiment results

The experiments presented here are designed to verify the validity of the following statements:

- The age/gender group specific training of the DNN does not necessarily lead to improved performance, specially when a small amount of data is available
- The age/gender group adaptation of a general DNN can help to design group specific systems, even when only a small amount of data is available.

During the experiments the language model weight is tuned on the development set and used to decode the test set. Results were obtained with a phone loop language model and the PER was computed based on 28 phone labels. Variations in recognition performance were validated using the matched-pair sentence test (Gillick and Cox, 1989) to ascertain whether the observed results were inconsistent with the null hypothesis that the output of two systems were statistically identical. Considered significance levels were .05, .01 and .001.

4.1 Age/gender specific training for DNN-HMM

In this experiment, DNN are trained on group specific corpora (children's speech in ChildIt,

Training Set	Evaluation Set					
	ChildIt		APASCI (f)		APASCI (m)	
	Dev	Test	Dev	Test	Dev	Test
Mixture	13.98%	15.56%	**10.12%**	**10.91%**	**10.70%**	**8.62%**
ChildIt	**12.08%**	**12.76%**	24.46%	29.59%	50.93%	46.16%
APASCI (f)	32.23%	34.23%	10.92%	12.75%	36.01%	31.21%
APASCI (m)	53.85%	56.11%	29.73 %	30.81%	11.36%	9.83%

Table 1: Phone error rate achieved with the DNN-HMM trained age/gender groups specific data.

Adaptation Set	Evaluation Set					
	ChildIt		APASCI (f)		APASCI (m)	
	Dev	Test	Dev	Test	Dev	Test
No adaptation	13.98%	15.56%	10.12%	10.91%	10.70%	8.62%
ChildIt	**11.68%**	**12.43%**	13.82 %	16.93%	28.89 %	24.96%
APASCI (f)	19.77%	21.91%	**8.30%**	**9.65%**	20.40%	17.01%
APASCI (m)	30.04 %	32.33%	16.78 %	16.99%	**9.33%**	**7.61%**

Table 2: Phone error rate achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to speficic age/gender groups.

adult female speech in APASCI and adult male speech in APASCI) and performance are compared with the DNN-HMM baseline introduced above. Recognition results are reported in Table 1, which includes results achieved with the DNN-HMM baseline in the row *Mixture*. In ChildIt there is about 7h of training data which is apparently sufficient to train an effective DNN and we can observe an improvement of 2.8% PER ($p < .001$), from 15.56% to 12.76%. However, in adult data there is only about 2h:40m of data for each gender. This is apparently not sufficient to train a DNN. In fact, the DNN-HMM system based on a DNN that is trained on gender specific data consistently degrades the PER. The degradation is 1.84% PER on female speakers in APASCI ($p < .001$) and 1.21% PER on male speakers in APASCI ($p < .001$).

4.2 Age/gender adapted DNN-HMM

In this experiment the DNN trained on all available corpora is adapted to each group specific corpus and recognition performance is compared with that obtained by the DNN-HMM baseline (where the DNN is trained on all available corpora). PER performance is presented in Table 2 which also reports the results achieved by the DNN-HMM baseline (in row *No adaptation*). The group adapted DNN-HMM consistently improve the PER compared to the DNN-HMM baseline. On children's speech the PER improvement is of 3.13% ($p < .001$), from 15.56% to 12.43%, for adult female speakers in APASCI the PER improvement is 1.26% ($p < .001$), from 10.91% to 9.65% and for adult male speakers in APASCI the PER improvement is of 1.01% ($p < .05$), from 8.62% to 7.61%.

5 Conclusions

In this paper we have investigated the use of the DNN-HMM approach in a phone recognition task targeting three groups of speakers, that is children, adult males and adult females. It has been shown that, in under-resourced condition, group specific training does not necessarily lead to PER improvements. To overcome this problem a recent approach, which consists in adapting a task independent DNN for tandem ASR to domain/language specific data, has been extended to age/gender specific DNN adaptation for DNN-HMM. The DNN-HMM adapted on a low amount of group specific data have been shown to improve the PER by 15-20% relative with respect to the DNN-HMM baseline system trained on speech data from all the three groups of speakers.

In this work we have proven the effectiveness of the hybrid DNN-HMM approach when training with limited amount of data and targeting speaker populations of different age/gender. Future work will be devoted to embed the results presented here in a large vocabulary speech recogniser especially targeting under-resourced groups of speakers such as children.

References

B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. 1994. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proc. of ICSLP*, pages 1391–1394, Yokohama, Japan, Sept.

Herve A Bourlard and Nelson Morgan. 1994. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer.

T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle. 1998. A Novel Feature Transformation for Vocal Tract Length Normalisation in Automat ic Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549–557, Nov.

G.E. Dahl, Dong Yu, Li Deng, and A. Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, Jan.

S. Das, D. Nix, and M. Picheny. 1998. Improvements in Children's Speech Recognition Performance. In *Proc. of IEEE ICASSP*, Seattle,WA, May.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660.

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2007. Acoustic variability and automatic recognition of childrens speech. *Speech Communication*, 49(1011):847 – 860.

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2009. Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499 – 509.

L. Gillick and S. Cox. 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. of IEEE ICASSP*, pages I–532–535, Glasgow, Scotland, May.

D. Giuliani and M. Gerosa. 2003. Investigating Recognition of Children Speech. In *Proc. of IEEE ICASSP*, volume 2, pages 137–140, Hong Kong, Apr.

A. Hagen, B. Pellom, and R. Cole. 2003. Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. of IEEE ASRU Workshop*, St. Thomas Irsee, US Virgin Islands, Dec.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov.

Viet-Bac Le, L. Lamel, and J. Gauvain. 2010. Multistyle ML features for BN transcription. In *Proc. of IEEE ICASSP*, pages 4866–4869, March.

A. Mohamed, G.E. Dahl, and G. Hinton. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, Jan.

R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. 2004. Public Speech-Oriented Guidance System with Adult and Child Discrimination Capabi lity. In *Proc. of IEEE ICASSP*, Montreal, Canada, May.

A. Potamianos and S. Narayanan. 2003. Robust Recognition of Children's Speech. *IEEE Trans. on Speech and Audio Processing*, 11(6):603–615, Nov.

S. Sivadas and H. Hermansky. 2004. On use of task independent training data in tandem feature extraction. In *Proc. of IEEE ICASSP*, volume 1, pages I–541–4, May.

S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann. 2003. Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In *Pattern Recognition, 25th DAGM Symposium*, pages 600–607, Sep.

A. Stolcke, F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *Proc. of IEEE ICASSP*, volume 1, pages 321–334, May.

P. Swietojanski, A. Ghoshal, and S. Renals. 2012. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. of IEEE SLT Workshop*, pages 246–251, Dec.

S. Thomas, M.L. Seltzer, K. Church, and H. Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. of IEEE ICASSP*, pages 6704–6708, May.

Karel Veselỳ, Lukáš Burget, and František Grézl. 2010. Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, pages 439–446. Springer.

J. G. Wilpon and C. N. Jacobsen. 1996. A Study of Speech Recognition for Children and Elderly. In *Proc. of IEEE ICASSP*, pages 349–352, Atlanta, GA, May.

Konig Yochai and Nelson Morgan. 1992. GDNN: a gender-dependent neural network for continuous speech recognition. In *Proc. of Iternational Joint Conference on Neural Networks*, volume 2, pages 332–337, Jun.