NON-NATIVE CHILDREN SPEECH RECOGNITION THROUGH TRANSFER LEARNING

Marco Matassoni, Roberto Gretter, Daniele Falavigna, Diego Giuliani

Center for Information and Communication Technology
Fondazione Bruno Kessler, via Sommarive 18, Trento (Italy)
{matasso,gretter,falavi,giuliani}@fbk.eu

ABSTRACT

This work deals with non-native children's speech and investigates both multi-task and transfer learning approaches to adapt a multi-language Deep Neural Network (DNN) to speakers, specifically children, learning a foreign language. The application scenario is characterized by young students learning English and German and reading sentences in these second-languages, as well as in their mother language. The paper analyzes and discusses techniques for training effective DNN-based acoustic models starting from children's native speech and performing adaptation with limited non-native audio material. A multi-lingual model is adopted as baseline, where a common phonetic lexicon, defined in terms of the units of the International Phonetic Alphabet (IPA), is shared across the three languages at hand (Italian, German and English); DNN adaptation methods based on transfer learning are evaluated on significant non-native evaluation sets. Results show that the resulting non-native models allow a significant improvement with respect to a mono-lingual system adapted to speakers of the target language.

***Index Terms*—** Transfer learning, Multi-task learning, non-native speech recognition, children's speech

1. INTRODUCTION

Nowadays the usage of deep neural networks hidden Markov models (DNN-HMMs) [1, 2] provides effective performance in speech recognition: there are concrete applications ranging from mobile voice search [3], transcriptions of broadcast news, videos [4] or conversations [5] to recognition in noisy environments [6, 7, 8].

The availability of large training corpora for a given application domain allows the training of a DNN with many layers and parameters in order to improve the classification performance. On the contrary, in the absence of sufficient data for training, e.g. in the case of under-resourced languages, the number of DNN parameters that can be reliably estimated greatly reduces and, consequently, classification performance is not always satisfactory. Recognition of children's speech is an application domain often characterized by training data shortage, even for major languages.

As alternative to complete training of DNN parameters, adaptation of an existing DNN by using the available small data set is a viable approach. This has been investigated in [9, 10], where an initial DNN trained on adult speakers is then adapted using a limited set of children's data. Another approach, to address the lack of training data is represented by *multi-task learning*. This approach has been demonstrated effective for multi-lingual speech recognition, especially if the size of training data for each language is small [11, 12, 13, 14, 15]. The reason of this is due to the fact that the shared hidden layers of the DNN used to estimate the emission probabilities of HMM states in a hybrid Automatic Speech Recognition (ASR) system [1], are language independent if the DNN itself is trained on multi-lingual data. This DNN can then be used to initialize a new one which can be trained only with data of the target language. When the size of training data is small only a subset of the connection weights, usually those of the output layer, are re-estimated. This training procedure is often called *transfer learning*, to indicate the fact that an initial set of learned parameters is transferred to the final acoustic model used by the ASR system.

In this work we address the problem of automatic speech recognition for children speaking a non-native language, specifically: (*a*) Italian students, speaking both English and German, and (*b*) German students speaking English.

It is known that non-native speakers articulate sounds very differently from native ones, because they try to use the phonology of their mother language, giving rise to pronunciation errors [16], mainly: when speakers try to apply spelling-to-sound rules of their native language to the new language, resulting in an inappropriate choice of target phoneme string or when the influence of their native phonological system interferes with the production of sound in the target system [17]. In the past several approaches have been proposed to take into account the pronunciation errors of non-native speakers [18], spanning from the usage of non-native pronunciation lexicon [19, 20, 21, 22, 23] to acoustic model adaptation using either native data and non native data [24, 25, 26, 27].

As previously mentioned, we use transfer learning to adapt the multi-lingual DNN trained on native data from Italian, German and English children. Basically, only the weights of the output layer of the network are updated, through back propagation, using data from non-native speakers of a given language while the weights of the lower layers are frozen and remain unchanged during adaptation. In addition, we propose to use multi-lingual data even in the adaptation phase, that is to update the weights of the output layer of the original DNN with all available non-native data.

The novelties of this work are the application of both multi-task learning and transfer learning to the recognition of speech from children speaking in a foreign language and the usage of (non-native) multi-lingual data for updating the weights of the original transferred DNN.

Experimental results reported in the paper show that: *(a)* the multi-lingual DNN performs similarly, on native data, to mono-lingual networks, i.e. trained only with data from a single language; *(b)* the multi-lingual DNN performs significantly better on non-native data than mono-lingual DNNs; *(c)* the usage of multi-lingual data in the adaptation process further increases the performance on non-native data.

This paper is organized as follows: Section 2 describes the experimental data used for testing the approaches proposed in this work, Section 3 gives details of the acoustic models, language models and IPA based lexicon used in the multi-lingual ASR system employed in this work; section 4 reports experiments and related results. Finally, Section 5 concludes the paper, presenting directions for future work.

2. SPEECH DATA

In this work we exploited speech data collected within the European funded project PF-Star (2002-2004). During the PF-Star project, noticeable amount of speech data were collected from English, German, Italian and Swedish children [28](for one of the Italian corpora see also [29]). For the purposes of this work, children's speech pronounced by English, German and Italian students in the three languages was considered, as shown in Table 1. As already mentioned, data from native speakers were used to train Acoustic Models (AMs), while test was carried out on both native and non-native data sets. In addition, a non-native adaptation set was used in transfer learning and, finally, performance, measured in terms of Word Error Rate (WER), was computed on both native and non-native evaluation data sets. The overall age range of the children is 7-13 and there is no age mismatch between training and test children.

speakers\lang	Italian	German	English
Italian	train + eval	ada + eval	ada + eval
German	–	train + eval	ada + eval
English	–	–	train + eval

Table 1. Native and non-native children's speech corpora used in the paper with *train*, *eval* and *ada* denoting training, evaluation and adaptation data sets.

number of speakers		language spoken	total duration	running words	lexicon size
mono-lingual native training corpora					
115	Italian	Italian	07:15:53	49233	9519
168	German	German	07:45:31	49326	7451
70	English	English	06:04:49	26873	1267
mono-lingual native eval corpora					
42	Italian	Italian	02:37:07	17936	5042
11	German	German	01:21:18	7859	1948
30	English	English	01:40:38	9224	1036

Table 2. Details for mono-lingual, native, training and eval data sets.

Table 2 reports details about the mono-lingual training and eval data—in all cases, native children's speech—in terms of number of speakers, duration, number of running words and lexicon size. Table 3 presents some statistics about the non-native speech data. In particular, Italian children read both English and German sentences, while German children read English sentences. In all cases, the speech data was split into *ada* (used to perform transfer learning) and *eval* data (for evaluation purposes only). Overlapping among training, ada and eval speakers never occurs.

number of speakers		language spoken	total duration	running words	lexicon size
non-native ada corpora					
9	Italian	German	00:29:54	2575	438
21	Italian	English	00:58:24	2753	390
42	German	English	00:30:19	3081	597
non-native eval corpora					
13	Italian	German	00:46:14	3769	474
27	Italian	English	01:16:29	3632	444
52	German	English	00:43:14	4440	630

Table 3. Details for non-native ada and eval data sets.

3. ASR SYSTEM

As mentioned in the Introduction, a multi-lingual DNN was first trained on the data of native speakers.

3.1. Multi-lingual DNN

The ASR system is based on the KALDI open source software toolkit [30]. The baseline acoustic model is build following the Karel's DNN recipe [31]: the preliminary HMM is trained on the usual 13 mel-frequency cepstral coefficients (MFCCs), which are mean and variance normalized; fMLLR-transformed coefficients are then estimated and used as input features for the DNN. The learning procedure features layer-wise pre-training based on Restricted Boltzmann Machines, per-frame cross-entropy training and sequence-discriminative training (lattice framework and State Minimum Bayes Risk criterion). Besides the mono-lingual DNNs trained on native German, Italian, English speech, a multi-lingual model is derived from a shared lexicon (see Section 3.3 for details): Table 4 reports the number of phonetic units used by the mono-lingual lexica as well as by the multi-lingual lexicon; it also reports the size of the output layer of the DNNs trained for mono-lingual and multi-lingual speech recognition.

	units	dnn output
Italian	28	1679
German	45	1592
English	43	1526
multi	67	1632

Table 4. Number of phonetic units and size of the output layer of the DNNs trained for mono- and multi-lingual DNNs.

In the context of the work the term ”transfer learning” is intended as re-estimation of the weights of the DNN output layer.

3.2. Language Models

Texts used to train the LMs are the transcriptions of the audio training data. Since the focus of this paper is on acoustic modeling, we did not cope with Out-Of-Vocabulary (OOV) word issues, which would have complicated the analysis of results; as such, the words of the test sets have been added to the training text data as unigrams. Also, in order to avoid to use different LMs for native and non-native corpora, we decided to build only a single LM for each language. For this reason, the lexicon of each language has to contain at least all of the words included in the native eval, non-native ada, and eval data sets. We decided to use a bigram LM, with Witten-Bell smoothing, that assures a reasonable perplexity over the ada + eval data sets, as reported in Table 5.

language	lexicon size	running words	2-grams	PP	OOV
Italian	5042	17936	13854	38.2	0.0%
German	2194	14203	6959	20.6	0.0%
English	1289	23130	4983	25.0	0.0%

Table 5. Text data used to build the three LMs and related perplexities (PP).

3.3. Lexicon

Concerning the lexicon, we have at our disposal grapheme to phoneme converters for the three languages, that were used in the past to build mono-lingual ASR systems. For this work, we decided to convert all the mono-lingual phones in IPA format, shown here as ASCII sequences. Of course, some of the choices we made (for instance, we replaced geminate consonants with simple ones for the Italian lexicon) are questionable and some of them could be revised in the future. Table 6 contains the list of all the 67 phones resulting from the merging of the three lexica. Of these, 18 phones are common for all three languages, 13 are common to only 2 languages (9 de+en, 2 it+de, 2 it+en), and 36 are present in one language only (16 de, 14 en, 6 it).

A”		de		OW			en	j	it	de	en
AA			en	OY		de	en	k	it	de	en
AE			en	O		de	en	l	it	de	en
AH			en	R		de		m	it	de	en
AI		de	en	S	it	de	en	n	it	de	en
AU		de	en	TH			en	o:		de	
AX			en	U”		de		o	it		
C		de		UA			en	p	it	de	en
DH			en	U		de	en	pf		de	
E@		de		Z			en	r	it	de	en
EA			en	a:		de		s	it	de	en
ER6		de		a	it	de		tS	it	de	en
ER			en	b	it	de	en	t	it	de	en
EY			en	dZ	it	de	en	ts	it	de	
E		de	en	d	it	de	en	u”		de	
IA			en	dz	it			u:		de	
I		de	en	e:		de		u	it		en
J	it			e	it			v	it	de	en
L	it			f	it	de	en	w	it		en
NG		de	en	g	it	de	en	x		de	
O”2		de		h		de	en	z	it	de	en
O”9		de		i:		de					
OH			en	i	it						

Table 6. List of IPA-like (expressed in ASCII characters) phones used for the three languages.

4. EXPERIMENTS AND RESULTS

Table 7 reports the WER results obtained using the four reference acoustic models: the three mono-lingual models, trained on clean data from native children speaking Italian, German and English, perform slightly better than the multi-lingual model trained on the three data sets using the multi-lingual lexicon. Nevertheless, the multi-lingual system shows more robustness against non-native speech: the off-diagonal WERs, that represent the performance of young students speaking a foreign language, are significantly lower when the target language is English (39.8% against 53.2 and 31.7 against 40.4) while it slightly worse in case of German (i.e. 18.5% against 17.4%). Note that, although the perplexity is higher, the WER of the Italian baseline is the lowest one, due to better recording quality.

The forthcoming experiment shows the effectiveness of transfer learning in this applicative context: the baseline DNNs are adapted to non-native speech using limited data from the target domain. The adaptation sets comprise data of Italian students reading German and English sentences and German students reading English text. The adaptation is im-

speakers \ language	Italian	German	English
mono-lingual AMs			
Italian	2.1	17.4	53.2
German	-	7.3	40.4
English	-	-	8.0
multi-lingual AM			
Italian	2.2	18.5	39.8
German	-	7.9	31.7
English	-	-	10.4

Table 7. WERs using mono and multi-lingual acoustic models. For mono-lingual non-native experiments (the off-diagonals numbers), the system is trained on the native speech of the target language.

plemented by keeping fixed all the layers of the DNNs except the last one. A few additional learning iterations (5) are then performed using the adaptation material. We evaluated three modalities related to transfer learning: *(m1)* the mono-lingual model is adapted to non-native speakers using adaptation data of the single target language; *(m2)* similarly, the multi-lingual model is adapted to a single target language; *(m3)* the multi-lingual model is adapted to multi-lingual non-native speech using together the three adaptation sets (i.e. English and German data coming from Italian speakers and English data coming from German speakers).

The results presented in Table 8 demonstrate that the multi-lingual system is better capable of coping with the acoustic variability of non-native speech. The multi-lingual system exhibits better behavior, producing more effective models for non-native speech: the WERs related to Italian students speaking in both German and English decrease from 11.1% to 9.6% and from 16.0% to 15.4%, respectively; a larger gain is observed in case of Germans speaking in English (WER from 18.3% to 15.2%). Similar improvement can be observed adapting the multi-lingual system with non-native speech of the three target languages.

speakers \ language	German	English
adapted mono-lingual AMs (*m1*)		
Italian	11.1	16.0
German	-	18.3
adapted multi-lingual AMs (*m2*)		
Italian	**9.6**	15.4
German	-	15.2
adapted multi-lingual AM (*m3*)		
Italian	10.4	15.0
German	-	15.1

Table 8. WERs obtained with mono- and multi-lingual acoustic models adapted to non-native speech using the three modalities *m1*, *m2*, and *m3*.

The next experiment (see Table 9) investigates the case where, starting from the multi-lingual model, the adaptation data is aggregated with respect to the native language of speakers and to the target language. In the first case, we use the adaptation sentences of Italian speakers reading in German or English. In the second case, the adaptation set is defined in terms of the language spoken by the students, regardless of their original mother language, that is English sentences uttered by Italian and German students. Also in this case, the gain with respect to the mono-lingual case is noticeable. Moreover, this combination can lead to the best results (highlighted in bold) for two cases; of course, as expected, the pairs with no adaptation data (German-English pair and Italian-German pair, respectively) produce worst results.

speakers \ language	German	English
adapted multi-lingual AM with Italian speakers		
Italian	10.3	**14.2**
German	-	19.8
adapted multi-lingual AM with English utterances		
Italian	16.7	14.8
German	-	**15.0**

Table 9. WERs related to experiments exploring modalities in which non-native adaptation data sets are merged according to source or target language (i.e. Italian native speakers and students speaking in English, respectively).

As a consequence, we conclude that transfer-learning in the context of children non-native speech, where usually a limited amount of data is available for training purposes, can be successfully applied to mitigate the acoustic mismatch. Moreover, it seems evident that the hidden layers of the multi-lingual DNN are able to build a more general representation of the phonetic space and this turns out to be suitable for adaptation to non-native speakers.

5. CONCLUSIONS

In this work we have investigated the application of transfer learning for adapting a multi-lingual DNN, trained on native speech from three languages (Italian, German, English), to non-native data. The approach is implemented updating the output layer of the DNN with small adaptation sets; the experimental results confirm the validity of this technique and show the positive effect of a multi-language model to compensate for the pronunciation differences of a non-native speaker.

For future work we plan to further address non-native acoustic modeling, experimenting on other types of acoustic features and exploiting some a-priori knowledge about phonetics of the first and foreign languages. In particular, it seems promising to investigate alternative lexicons that take into account the possible pronunciation variations introduced by non-native students.

6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, and Y. Wang, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 9, no. 3, pp. 82–97, 2012.

[2] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. on Audio Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[3] K. Yao, D. Yu, F. Seide, H. Su, and Y. Gong, "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," in *Proc. of ICASSP*, Kyoto (Japan), March, 25-30 2012, pp. 366–369.

[4] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Trans. on Audio Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[5] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcriptions Using Context-Dependent Deep Neural Networks," in *Proc. of Interspeech*, Florence (Italy), 2011, pp. 437–440.

[6] M. Seltzer, D. Yu, and Y. Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP*, Vancouver, Canada, May 2013, pp. 7398–7402.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proc. of IEEE ASRU Workshop*, Scottsdale, Arizona (USA), 2015.

[8] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proc. of Hands-free Speech Communication and Microphone Arrays (HSCMA) Wokshop*, Villers-les-Nancy, 2014, pp. 172–176.

[9] M. Matassoni, D. Falavigna, and D. Giuliani, "DNN adaptation for recognition of children speech through automatic utterance selection," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 644–651.

[10] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. FirstView, pp. 1–26, 7 2016.

[11] Y. Miao and F. Metze, "Improving low-resource cd-dnnhmm using dropout and multilingual dnn training," in *Interspeech*, 2013, pp. 2237–2241.

[12] N. T. Vu, D. Imseng, P. Mot. Schultz D. Povey, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. of ICASSP*, Forence, Italy, May, 4-9 2014, pp. 7639–764.

[13] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of ICASSP*, 2013, pp. 7304–7308.

[14] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. of ICASSP*, 2013, pp. 7319–7323.

[15] Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for speech and speaker recognition," *arXiv preprint arXiv:1603.09643*, 2016.

[16] Silke M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 1–8.

[17] M. Russell, "Analysis of Italian children's English pronunciation," *http://archive.is/http://www.eee.bham.ac.uk/russellm*, 2007.

[18] G. Bouselmi, D. Fohr, I. Illina, and J. P. Haton, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *Proc. of ICSLP*, 2006, pp. 109–112.

[19] Z. Wang and T. Schultz, "Non-native spontaneous speech recognition through polyphone decision tree specialization," in *Proc. of Eurospeech*, 2003, pp. 1449–1452.

[20] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. of ICASSP*, 2003, pp. 540–543.

[21] Y. R. Oh, J. S. Yoon, and H. K. Kim, "Adaptation based on pronunciation variability analysis for non native speech recognition," in *Proc. of ICASSP*, 2006, pp. 137–140.

[22] H. Strik, K.P. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[23] S. Steidl, G. Stemmer, C. Hacker, and E. Nöth, "Adaptation in the pronunciation space for non-native speech recognition," in *Proc. of ICSLP*, 2004, pp. 2901–2904.

[24] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Articulatory modeling for pronunciation error detection without non-native training data based on dnn transfer learning," *IEICE Transactions on Information and Systems*, vol. E100D, no. 9, pp. 2174–2182, 2017.

[25] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *Proc. of ICASSP*, 2016, vol. 2016-May, pp. 6135–6139.

[26] A. Lee and J. Glass, "Mispronunciation detection without non-native training data," in *Proc. of Interspeech*, 2015, vol. 2015-January, pp. 643–647.

[27] A. Das and M. Hasegawa-Johnson, "Cross-lingual transfer learning during supervised training in low resource scenarios," *Proc. of Interspeech*, vol. 2015-January, pp. 3531–3535, 2015.

[28] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF-STAR children's speech corpus," in *Proc. of Eurospeech*, 01 2005, pp. 2761–2764.

[29] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10, pp. 847 – 860, 2007, Intrinsic Speech Variations.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselỳ, "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE ASRU Workshop*, Hawaii (US), December 2011.

[31] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. of Interspeech*, Florence, Italy, August 2011, pp. 2345–2349.