PIERO COSI, GIULIO PACI, GIACOMO SOMMAVILLA, FABIO TESSER [1]

KALDI: Yet Another ASR Toolkit?
Experiments on Italian children speech

Abstract: In this paper, the KALDI ASR engine adapted to Italian is described and the results obtained so far on some children speech ASR experiments are reported. We give a brief overview of KALDI, we describe in detail its DNN implementation, we introduce the acoustic model (AM) training procedure and we end describing some ASR experiments on Italian children speech together with the final test procedures.

1 Introduction

Deep Neural Networks (DNNs) are the latest hot topic in speech recognition. Since around 2010, many papers have been published in this area, and some of the largest companies (e.g. Google, Microsoft) are starting to use DNNs in their production systems. Moreover, many different Automatic Speech Recognition frameworks have been developed for research purposes during the last years and, nowadays, various open-source automatic speech recognition (ASR) toolkits are available to even small research laboratories. Systems such as HTK (Young et al., 2009), SONIC (Pellom, 2001), (Pellom & Hacioglu, 2003), SPHINX (Lee et al., 1990), (Walker et al., 2004), RWTH (Rybach et al., 2009), JULIUS (Lee et al., 2001), KALDI (Povey et al., 2011), the more recent ASR framework SIMON (Simon, WEB), and the relatively new system called BAVIECA (Bolaños, 2012) are a simple and probably not exhaustive list.

Indeed new systems such as KALDI (Povey et al., 2011) demonstrated the effectiveness of easily incorporating “Deep Neural Network” (DNN) techniques (Bengio, 2009) in order to improve the recognition performance in almost all recognition tasks.

[1] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche -
Unità Organizzativa di Supporto di Padova - Italy
[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it.

In this paper, the KALDI ASR engine adapted to Italian is described and the results obtained so far on some children speech ASR experiments are reported. We give a brief overview of KALDI, and in particular of its DNN implementation, we introduce the acoustic model (AM) training procedure and we end describing some experiments on Italian children speech together with the final test procedures.

2 KALDI

As written in his official web site (http://KALDI.sourceforge.net), the KALDI ASR environment should be mainly taken into consideration for the following simple reasons:

- it's "easy to use" (once you learn the basics, and assuming you understand the underlying science);
- it's "easy to extend and modify";
- it's "redistributable": unrestrictive license, community project;
- if your stuff works or is interesting, the KALDI team is open to including it and your example scripts in our central repository: more citation, as others build on it.

Another main reason for choosing KALDI is represented by the fact that complete open source recipes for building speech recognition systems, that work from widely available databases such as those provided by the ELRA or Linguistic Data Consortium (LDC), are easily available for testing the system.

3 DNN (Deep Neural Networks) in KALDI [2]

KALDI currently contains two parallel implementations for DNN training. Both of these recipes are deep neural networks where the last (output) layer is a softmax layer whose output dimension equals the number of context-dependent states in the system (typically several thousand). The neural net is trained to predict the posterior probability of each context-dependent state. During decoding the output probabilities are divided by the prior probability of each state to form a "pseudo-likelihood" that is used in place of the state emission probabilities in the HMM.

The first implementation (Kaldi, WEB-a), (Vesely et al., 2013) supports Restricted Boltzmann Machines (RBM) pre-training (Hinton et al., 2012),

[2] Most of the text used to briefly describe Kaldi in this section is taken from "Deep Neural Networks in Kaldi" (http://kaldi.sourceforge.net/dnn.html) with permission from the Author (Daniel Povey).

(Dahl et al., 2012), (Seide et al., 2011), stochastic gradient descent training using NVidia graphics processing units (GPUs), and discriminative training such as boosted MMI (Povey et al, 2008) and state-level minimum Bayes risk (sMBR) (Gibson & Hain, 2006), (Povey & Kingsbury, 2007).

The second implementation of DNNs in KALDI (Kaldi, WEB-b), (Zhang et al., 2014), (Povey et al., 2015) was originally written to support parallel training on multiple CPUs, although it has now been extended to support parallel GPU-based training and it does not support discriminative training.

One is located in code sub-directories nnet/ and nnetbin/ [3], and is primarily maintained by Karel Vesely. The other is located in code subdirectories nnet2/ and nnet2bin/, and is primarily maintained by Daniel Povey (this code was originally based on an earlier version of Karel's code, but it has been extensively rewritten). Neither codebase is more “official” than the other. Both are still being developed in parallel.

In the example directories (referring to the HKUST Mandarin Telephone Speech, Resource Management, Switchboard, Timit, and Wall Street Journal corpora)[4] neural net example scripts can be found. Before running those scripts, the first stages of “run.sh” in those directories must be run in order to build the systems used for alignment.

Regarding which of the two setups you should use:

- Karel's setup (nnet1) generally gives somewhat better results but it only supports training on a single GPU card, or on a single CPU which is very slow.
- Dan's setup generally gives slightly worse results but is more flexible in how you can train: it supports using multiple GPUs, or multiple CPU's each with multiple threads. Multiple GPU's is the recommended setup. They don't have to all be on the same machine.

The reasons for the performance difference is still unclear, as there are many differences in the recipes used. For example, Karel's setup uses pre-training but Dan's setup does not; Karel's setup uses early stopping using a validation set but Dan's setup uses a fixed number of epochs and averages the parameters over the last few epochs of training. Most other details of the training (nonlinearity types, learning rate schedules, etc.) also differ.

[3] See the code at: https://svn.code.sf.net/p/kaldi/code/trunk/egs

[4] such as egs/hkust/s5b, egs/rm/s5, egs/swbd/s5, egs/timit/s5/, and egs/wsj/s5/. Karel's example scripts can be found in local/run_dnn.sh or local/run_nnet.sh, and Dan's example scripts can be found in local/run_nnet2.sh.

3.1 Karel's DNN training implementation

The implementation of DNNs from Karel Vesely (Kaldi, WEB-a), (Vesely et al., 2013) uses the following techniques:

- layer-wise pre-training based on RBMs (Restricted Boltzmann Machines)
- per-frame cross-entropy training
- sequence-discriminative training, using lattice framework, optimizing sMBR criterion (State Minimum Bayes Risk)

The systems are built on top of MFCC, LDA, MLLT, fMLLR with CMN[5] features - see (Rath, et al., 2013) for all acronyms references - obtained from auxiliary GMM (Gaussian Mixture Model) models. Whole DNN training is run in a single GPU using CUDA (Compute Unified Device Architecture, the parallel computing architecture created by NVidiaTM). However, cudamatrix library is designed to also run on machines without a GPU, but this tends to be more than 10x slower.

The script for standard databases such wsj is located at: "egs/wsj/s5/local/run_dnn.sh" and it is split into several stages. At first the 40-dimensional features (MFCC, LDA, MLLT, fMLLR with CMN) are stored to disk in order to simplify the training scripts.

3.1.1 Pre-training phase

The implementation of layer-wise RBM (Restricted Boltzmann Machine) pre-training is following "www.cs.toronto.edu/~hinton/absps/guideTR.pdf" (Hinton, 2010).

The training algorithm is Contrastive Divergence with 1-step of Markov Chain Monte Carlo sampling (CD-1). The hyper-parameters of the recipe were tuned on the 100 hours Switchboard subset. If smaller databases are used, mainly the number of epochs N needs to be set to 100 hours/set_size. The training is unsupervised, so it is sufficient to provide single data-directory with input features.

When training the RBM with Gaussian-Bernoulli units, there is a high risk of weight-explosion, especially with larger learning rates and thousands of hidden neurons. To avoid weight-explosion, a mechanism has been implemented, which compares the variance of training data with the variance of the

[5] MFCC: Mel-Frequency Cepstral Coefficients; LDA: Linear Discriminant Analysis; MLTT: Maximum Likelihood Linear Transform; fMLLR: feature space Maximum Likelihood Linear Regression; CMN: Cepstral Mean Normalization.

reconstruction data in a minibatch. If the variance of reconstruction is > 2x larger, the weights are shrinked and the learning rate is temporarily reduced.

3.1.2 Frame-level cross-entropy training

In this phase a DNN which classifies frames into triphone-states is trained. This is done by mini-batch Stochastic Gradient Descent[6]. The default is to use Sigmoid hidden units, Softmax output units and fully connected layers AffineTransform. The learning rate by default is 0.008, size of mini-batch 256; no momentum or regularization is used. The optimal learning-rate differs with type of hidden units, the value for sigmoid is 0.008, for tanh 0.00001.

The input_transform and the pre-trained DBN (i.e. Deep Belief Network, stack of RBMs) is passed into to the script using the options '–input-transform' and '–dbn'; only the output layer is initialized randomly. An early stopping criterium is used to prevent over-fitting: for this, the objective function on the cross-validation set (i.e. held-out set) is measured. Therefore two pairs of feature-alignment dirs are needed to perform the supervised training.

3.1.3 Sequence-discriminative training

In this phase, the neural network is trained to classify correctly the whole sentences, which is closer to the general ASR objective than frame-level training. The objective of sequence-discriminative training is to maximize the expected accuracy of state labels derived from reference transcriptions: lattice framework to represent competing hypothesis is used. The training is done by Stochastic Gradient Descent (SGD) with per-utterance updates, low learning rate: 1e-5 kept constant is used and 3-5 epochs are done. Faster convergence when re-generating lattices after 1st epoch are observed.

MMI, BMMI, MPE and sMBR[7] training are all supported – (see Rath et al., 2013 for all acronyms references). All the techniques perform very similarly, even if sMBR is a little bit better. In sMBR optimization silence frames are excluded from accumulating approximate accuracies[8] (Vesely et al., 2013).

[6] A good summary paper on DNN training is “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups” by Geoffrey Hinton’s group (Hinton et al., 2012).

[7] MMI: Maximum Mutual Information; BMMI: Boosted MMI; MPE: Minimum Phone Error; sMBR: State-level Minimum Bayes Risk

[8] More detailed description is at:
http://www.danielpovey.com/files/2013_interspeech_dnn.pdf

3.2 Povey's DNN training implementation

For the full documentation that covers Dan Povey's version of the deep neural network code in KALDI one could refer to the following web link (Kaldi, WEB-b), and to the following papers (Zhang et al., 2014; Povey et al., 2015). In its last implementation stage, as indicated in (Povey et al., 2015), where the Dan's DNN is used on a speech recognition setup called Fisher English, which is English language conversational telephone speech, sampled at 8kHz, for a total amount of training data of 1600 hours, the "… DNN system uses speaker adapted features from a GMM system, so it requires a first pass of GMM decoding and adaptation. …".

"… The GMM system is based on MFCC features, spliced across ±3 frames and processed with LDA+MLLT to 40-dimensional features, then adapted with feature-space MLLR (fMLLR) in both training and test time. See (Povey & Ghoshal, 2011) for an explanation of these terms and the normal system build steps. All these systems used the same phonetic context decision tree with 7880 context-dependent states; the GMM system had 300000 Gaussians in total. ..."

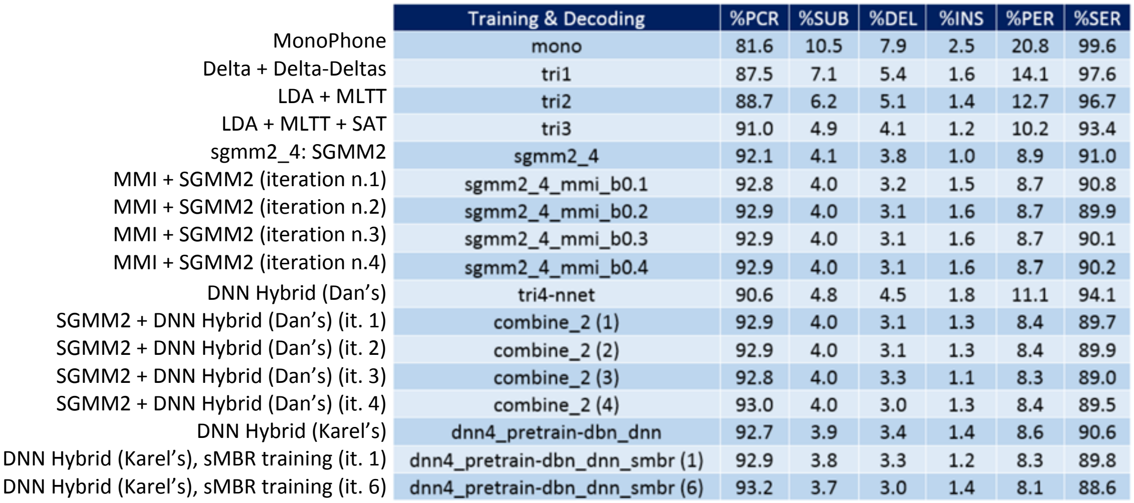
".. The 40-dimensional features from GMM are spliced across ±4 frames of context and used as input to the DNN. DNN is a p-norm DNN (Zhang et al., 2014), with 5 hidden layers and p-norm (input, output) dimensions of (5000, 500) respectively, i.e. the nonlinearity reduces the dimension tenfold. In this framework 15000 "sub-classes" are used - see Section C.3 of (Povey et al., 2015) for explanation- and the number of parameters is 19.3 million. The system is trained for 12 epochs with learning rate varying from 0.08 to 0.008, trained with 8 parallel jobs with online natural gradient SGD (NG-SGD) and, for this DNN system, K=400000 samples per outer iteration for each machine are used for training. …".

As for the TIMIT recipe (s5), Dan's DNN is much simpler and adopts a classic Hybrid Training and Decoding framework using a simple deep network with tanh nonlinearities. Moreover, also system combination using minimum Bayes risk decoding is used, and in this case a lattice combination is used to create a union of lattices normalized by removing the total forward cost from them and using the resulting lattice as input to the last decoding step.

4 KALDI on Italian

In this section, the adaptation of KALDI to Italian is described, starting from the TIMIT recipe, and the results obtained so far on some children speech ASR experiments are reported (see Table 1).

In the experiments described below, the Italian FBK ChildIt Corpus (Gerosa et al., 2007) was taken into consideration. This is a corpus that counts almost 10 hours of speech from 171 children; each child has read about 60 children literature sentences; the audio was sampled at 16 kHz, 16 bit linear, using a Shure SM10A head-worn mic.



	Training & Decoding	%PCR	%SUB	%DEL	%INS	%PER	%SER
MonoPhone	mono	81.6	10.5	7.9	2.5	20.8	99.6
Delta + Delta-Deltas	tri1	87.5	7.1	5.4	1.6	14.1	97.6
LDA + MLTT	tri2	88.7	6.2	5.1	1.4	12.7	96.7
LDA + MLTT + SAT	tri3	91.0	4.9	4.1	1.2	10.2	93.4
sgmm2_4: SGMM2	sgmm2_4	92.1	4.1	3.8	1.0	8.9	91.0
MMI + SGMM2 (iteration n.1)	sgmm2_4_mmi_b0.1	92.8	4.0	3.2	1.5	8.7	90.8
MMI + SGMM2 (iteration n.2)	sgmm2_4_mmi_b0.2	92.9	4.0	3.1	1.6	8.7	89.9
MMI + SGMM2 (iteration n.3)	sgmm2_4_mmi_b0.3	92.9	4.0	3.1	1.6	8.7	90.1
MMI + SGMM2 (iteration n.4)	sgmm2_4_mmi_b0.4	92.9	4.0	3.1	1.6	8.7	90.2
DNN Hybrid (Dan's)	tri4-nnet	90.6	4.8	4.5	1.8	11.1	94.1
SGMM2 + DNN Hybrid (Dan's) (it. 1)	combine_2 (1)	92.9	4.0	3.1	1.3	8.4	89.7
SGMM2 + DNN Hybrid (Dan's) (it. 2)	combine_2 (2)	92.9	4.0	3.1	1.3	8.4	89.9
SGMM2 + DNN Hybrid (Dan's) (it. 3)	combine_2 (3)	92.8	4.0	3.3	1.1	8.3	89.0
SGMM2 + DNN Hybrid (Dan's) (it. 4)	combine_2 (4)	93.0	4.0	3.0	1.3	8.4	89.5
DNN Hybrid (Karel's)	dnn4_pretrain-dbn_dnn	92.7	3.9	3.4	1.4	8.6	90.6
DNN Hybrid (Karel's), sMBR training (it. 1)	dnn4_pretrain-dbn_dnn_smbr (1)	92.9	3.8	3.3	1.2	8.3	89.8
DNN Hybrid (Karel's), sMBR training (it. 6)	dnn4_pretrain-dbn_dnn_smbr (6)	93.2	3.7	3.0	1.4	8.1	88.6

Table 1 – *Preliminary results obtained in the experiments executed on the CHILDIT corpus in various configurations adapting the KALDI's TIMIT-recipe scripts (see text for the definition of all keywords).*

Various experiments have been carried out in many configurations and in all cases, training and test materials have been kept separate. In all the experiments described in this work, the standard MFCC features were considered, setting reasonable defaults, even if other options could be exploited in the future. The results, shown in Figure 1, are the best obtained with KALDI on the CHILDIT corpus and they are the best obtained so far in comparison with those obtained by similar experiments reported in the literature (Gerosa et al., 2007; Giuliani & Gerosa, 2003; Cosi & Pellom, 2005; Cosi, 2008; Cosi, 2009; Cosi & Hosom, 2000; Cosi et al., 2014).

Phone Error Rate (PER) was considered for computing the score of the recognition process. The PER is defined as the sum of the deletion (DEL), substitution (SUB) and insertion (INS) percentage of phonemes in the ASR outcome text with respect to a reference transcription. PCR and SER refers to the Phone Correct Rate and Sentence Error Rate respectively.

Ideally, a hand-labelled reference would have been preferred, because it would have been corrected at the phonetic level to take into account children's speech pronunciation mistakes. Since this was not available for the CHILDIT corpus, the automatic phonetic sequences obtained by a Viterbi alignment of the word-level orthographic transcription have been used. The reference test transcriptions were created with a SONIC-based aligner using a previously trained children speech Italian acoustic model (Cosi & Pellom, 2005). This method was chosen because it allowed for automatically selecting the best pronunciation for each word in the training data among the alternative choices available in the 400,000-word Italian lexicon available.

The results shown in Table 1 refer to the various training and decoding experiments - see (Rath et al., 2013) for all acronyms references:

- MonoPhone (mono);
- Deltas + Delta-Deltas (tri1);
- LDA + MLLT (tri2);
- LDA + MLLT + SAT (tri3);
- SGMM2 (sgmm2_4);
- MMI + SGMM2 (sgmm2_4_mmi_b0.1-4);
- Dan's Hybrid DNN (tri4-nnet),
- system combination, that is Dan's DNN + SGMM (combine_2_1-4);
- Karel's Hybrid DNN (dnn4_pretrain-dbn_dnn);
- system combination that is Karel's DNN + sMBR (dnn4_pretrain-dbn_dnn_1-6).

In the Table, SAT refers to the Speaker Adapted Training (SAT), i.e. train on fMLLR-adapted features. It can be done on top of either LDA+MLLT, or delta and delta-delta features.

If there are no transforms supplied in the alignment directory, it will estimate transforms itself before building the tree (and in any case, it estimates transforms a number of times during training). SGMM2 refers instead to Subspace Gaussian Mixture Model training (Povey, 2009; Povey, Burget, et al. 2011). This training would normally be called on top of fMLLR features obtained from a conventional system, but it also works on top of any type of speaker-independent features (based on deltas+delta-deltas or LDA+MLLT).

On this difficult phonetic ASR task, Karel's DNN looks slightly better than Dan's DNN and it outperforms all other non-DNN configurations.

5 Conclusions

At its early stage, the KALDI toolkit supported modeling of context-dependent phones of arbitrary context lengths, all commonly used techniques that can be estimated using maximum likelihood, and almost all adaptation techniques available in the ASR literature. At present, it also supports the recently proposed Subspace Gaussian Mixture Model (SGMM) training (Povey, 2009;

Povey, Burget, et al., 2011), and the very promising DNN hybrid training and decoding (Kaldi, WEB-a; Veselý et al., 2013; Kaldi, WEB-b; Zhang et al., 2014; Povey et al., 2015). Moreover, developers are working on using large language models in the FST framework, and the development of KALDI is continuing.

The adaptation of KALDI to Italian, and in particular to the ChildIt corpus, was indeed quite straightforward, and results are truly exceptional with respect to those previously obtained on the same data (Gerosa et al., 2007; Giuliani & Gerosa, 2003; Cosi & Pellom, 2005; Cosi, 2008; Cosi, 2009; Cosi & Hosom, 2000; Cosi et al., 2014). Indeed, this is mainly because all the very last ASR techniques, including DNNs, could be easily implemented by adapting to Italian the already available downloadable scripts written by a quite large number of various developers around the world for a similar task in English based on TIMIT.

Acknowledgements

We would like to acknowledge and thank Daniel Povey for his invaluable work on KALDI, indeed a great ASR Toolkit, and for his great help and kindness in letting us use some of the text of the KALDI's web pages describing the software. Moreover, we would like to acknowledge Karel Vesely for his great work on his very effective version of DNN and for his great help and kindness in supporting me on Italian implementation.

References

Bengio, Y., 2009. Learning Deep Architectures for AI, in *Foundations and Trends in Machine Learning*, Vol. 2, No. 1 (2009) 1-127.

Bolaños, D., 2012. The Bavieca Open-Source Speech Recognition Toolkit, in *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, December 2-5, 2012, Miami, FL, USA, 2012.

Cosi, P., 2008. Recent Advances in Sonic Italian Children's Speech Recognition for Interactive Literacy Tutors, in *Proceedings of 1st Workshop On Child, Computer and Interaction (WOCCI-2008)*, Chania, Greece, 2008.

Cosi, P., 2009. On the Development of Matched and Mismatched Italian Children's Speech Recognition Systems, in *Proceedings of INTERSPEECH 2009*, Brighton, UK, 2009, 540-543.

Cosi, P., & Hosom, J. P., "High Performance General Purpose Phonetic Recognition for Italian", in *Proceedings of ICSLP 2000*, Beijing, 527-530, 2000.

Cosi, P., Nicolao, M., Paci, G., Sommavilla, G., & Tesser, F., 2014. Comparing Open Source ASR Toolkits on Italian Children Speech, in *Proceedings of Workshop On Child, Computer and Interaction (WOCCI-2014)*, Satellite Event of INTERSPEECH 2014, Singapore, September 19, 2014.

Cosi, P. & Pellom, B., 2005. Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors, in *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, 2005, 2201-2204.

Dahl, G. E., Dong, Yu, D., Deng, L, & Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, in *Audio, Speech, and Language Processing, IEEE Transactions on*, , 2012, vol. 20, no. 1, 30-42.

Gerosa, M., Giuliani, D. & Brugnara, F., 2007. Acoustic Variability and automatic recognition of children's speech, *Speech Communication*, Vol. 49, 2007, 847–860.

Giuliani, D. & Gerosa, M.. 2003. Investigating Recognition of Children's Speech", in *Proceedings of ICASSP, Hong Kong, 2003*, II 137-140.

Gibson, M. & Hain, T., 2006. Hypothesis Spaces For Minimum Bayes Risk Training In Large Vocabulary Speech Recognition, in *Proceedings of INTERSPEECH, 2006*, 2406-2409.

Hinton, G., 2010. *A Practical Guide to Training Restricted Boltzmann Machines*, UTML TR 2010:003, Momentum, 9:1, 2010.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Abdelrahman, M., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, in *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

Kaldi-WEBa - Karel's DNN implementation: *http://KALDI.sourceforge.net/dnn1.html*

Kaldi-WEBb - Dan's DNN implementation: http://KALDI.sourceforge.net/dnn2.html.

Lee, A., Kawahara, T., & Shikano, K, 2001. JULIUS - an open source real-time large vocabulary recognition engine, in *Proceedings of EUROSPEECH 2001*, Proc. European Conf. on Speech Communication and Technology, Aalborg, Denmark, Sep. 2001, 1691-1694.

Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system, in *IEEE Transactions on Acoustics, Speech and Signal Processing* 38.1, 1990, 35-45.

Pellom, B. 2001. "SONIC: The University of Colorado Continuous Speech Recognizer", *Technical Report TR-CSLR-2001-01*, Center for Spoken Language Research, University of Colorado, USA, 2001.

Pellom, B. & Hacioglu, K. 2003. Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task, in *Proceedings of ICASSP 2003*, I 4-7.

Povey D., 2009. Subspace Gaussian Mixture Models for Speech Recognition, *Tech. Rep. MSR-TR-2009-64, Microsoft Research*, 2009.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A:, Glembek, O., Goel, N.K., Karafiát, M., Rastrow, A., Rose, R.C., Schwarz, P., Thomas, S., 2011. The Subspace Gaussian Mixture Mode - A Structured Model for Speech Recognition, *Computer Speech & Language*, April 2011, vol. 25, no. 2, 404–439.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N.M., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K., 2011. The KALDI Speech Recognition Toolkit, in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE Signal Processing Society, December 2011.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K., 2008. Boosted MMI for Feature and Model Space Discriminative Training, in *Proceedings of ICASSP, 2008*, 3753-3756.

Povey, D. & Kingsbury, B., 2007. Evaluation of proposed modifications to MPE for large scale discriminative training, in *Proceedings of ICASSP, 2007*, IV 321-324.

Povey, D., Zhang, X., & Khudanpur, S., 2015. Parallel Training of DNNs with Natural Gradient and Parameter Averaging, in Proceedings of *ICLR 2015 - 5th International Conference on Learning Representations*, CoRR, vol. abs/1410.7455, 2014. [Online]. Available: http://arxiv.org/abs/1410.7455.

Rath, S. P., Povey, D., Veselý, K., & Cernocky, J., 2013. Improved feature processing for Deep Neural Networks, in *Proceedings of INTERSPEECH 2013*, 109-113.

Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., & Ney, H., 2009. The RWTH Aachen University Open Source Speech

Recognition System, in *Proceedings of INTERSPEECH, 2009*, 2111-2114.

Seide, F., Li, G., & Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks, in *Proceedings of INTERSPEECH, 2011*, 437–440.

Simon-WEB: http://www.simon-listens.com.

Veselý, K., Ghoshal, A., Burget, L., & Povey, D., 2013. Sequence-discriminative training of deep neural networks, in *Proceedings of INTERSPEECH 2013*, 2345-2349.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., & Woelfel, J., 2004. *Sphinx-4: A flexible Open Source Framework for Speech Recognition*, Sun Microsystems Inc., Technical Report SML1 TR2004-0811, 2004.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P., 2009. *The HTK Book* (version 3.4). Cambridge Univ. Eng. Dept., 2009.

Zhang, X, Trmal, J., Povey, D., & Khudanpur, S., 2014. Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks, in *Proceedings of. ICASSP 2014*, 215-219.

Author name, affiliation and email

Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser

Istituto di Scienze e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche -
Unità Organizzativa di Supporto di Padova
Via Martiri della libertà, 2
35137 Padova

[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it