Phone Recognition Experiments on ArtiPhon with KALDI

Piero Cosi
Istituto di Scienze e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche
Sede Secondaria di Padova – Italy
piero.cosi@pd.istc.cnr.it

Abstract

English. In this work we present the results obtained so far in different recognition experiments working on the audio only part of the ArtiPhon corpus used for the EVALITA 2016 speech-mismatch ArtiPhon task.

Italiano. *In questo lavoro si presentano i risultati ottenuti sinora in diversi esperimenti di riconoscimento fonetico utilizzanti esclusivamente la sola parte audio del corpus ArtiPhon utilizzato per il task ArtiPhon di EVALITA 2016.*

1 Introduction

In the last few years, the automatic speech recognition (ASR) technology has achieved remarkable results, mainly thanks to increased training data and computational resources. However, ASR trained on thousand hours of annotated speech can still perform poorly when training and testing conditions are different (e.g., different acoustic environments). This is usually referred to as the mismatch problem.

In the ArtiPhon task participants will have to build a speaker-dependent phone recognition system that will be evaluated on mismatched speech rates. While training data consists of read speech where the speaker was required to keep a constant speech rate, testing data range from slow and hyper-articulated speech to fast and hypo-articulated speech.

The training dataset contains simultaneous recordings of audio and vocal tract (i.e., articulatory) movements recorded with an electromagnetic articulograph (Canevari et al., 2015).

Participants were encouraged to use the training articulatory data to increase the generalization performance of their recognition system. However, we decided not to use them, mainly for the sake of time, but also because we wanted to compare the results with those obtained in the past on different adult and children speech audio-only corpora (Cosi & Hosom, 2000; Cosi & Pellom, 2005; Cosi, 2008; Cosi, 2009; Cosi et al., 2014; Cosi et al., 2015).

2 Data

We received the ArtiPhon (Canevari et al., 2015) training data by the Istituto Italiano di Tecnologia - Center for Translational Neurophysiology of Speech and Communication (CTNSC) late in July 2016, while the test material was released at the end of September 2016. The ArtiPhon dataset contains the audio and articulatory data recorded from three different speakers in citation condition. In particular for the EVALITA 2016 ArtiPhon - Articulatory Phone Recognition task only one speaker (cnz - 666 utterances) was considered.

The audio was sampled at 22050 Hz while articulatory data were extracted by the use of the NDI (Northen Digital Instruments, Canada) wave speech electromagnetic articulograph at 400 Hz sampling rate.

Four subdirectories are available:

- wav_1.0.0: each file contains an audio recording
- lab_1.0.0: each file contains phonetic labels automatically computed using HTK
- ema_1.0.0: each file contains 21 channels: coordinate in 3D space (xul yul zul xll yll zll xui yui zui xli yli zli xtb ytb ztb xtm ytm ztm xtt ytt ztt)

Head movement correction was automatically performed. First an adaptive median filter with a window from 10 ms to 50 ms and secondly a smooth elliptic low-pass filter with 20 Hz cutoff frequency were applied to each channel.

Unfortunately, we discovered that the audio data was completely saturated both in the training and the test set, thus forcing us to develop various experiments both using the full set of phonemes but also a smaller reduced set in order to make more effective and reliable the various phone recognition experiments.

3 ASR

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).

Deep Neural Networks (DNNs) are indeed the latest hot topic in speech recognition and 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.

DNNs has been already applied on different adults and children Italian speech corpora, obtaining quite promising results (Cosi, 2015; Serizel & Giuliani, 2014; Serizel & Giuliani, 2016).

In this work, the KALDI ASR engine adapted to Italian was adopted as the target ASR system to be evaluated on the ArtiPhon data set.

At the end we decided not to use the articulatory data available in the ArtiPhon data set, because we wanted to compare the final results of this task with those obtained in the past on different audio-only corpora which were not characterized by the above cited speech mismatch problem.

4 The EVALITA 2016 - ArtiPhon Task

A speaker dependent experiment characterized by training and test speech type mismatch was prepared by using the ArtiPhon task training and test material. A second speaker independent experiment was also set by testing the ArtiPhon test data using a previously trained ASR acoustic model on APASCI (Angelini et al., 1994), thus having in this case both speech type and speaker mismatch.

For both experiments, we used the KALDI ASR engine, and we started from the TIMIT recipe, which was adapted to the ArtiPhon Italian data set.

Deciding when a phone should be considered incorrectly recognized was another evaluation issue. In this work, as illustrated in Table 1, two set of phones, with 29 and 60 phones respectively, have been selected for the experiments, even if the second set is far from being realistic given the degraded quality of the audio signal.

60	29	60	29	60	29
a	a	j	j	pp	p
a1	a	J	J	r	r
b	b	JJ	J	rr	r
bb	b	k	k	s	s
d	d	kk	k	ss	s
dd	d	l	l	S	S
dz	dz	ll	l	SS	S
ddz	dz	L	L	t	t
dZ	dZ	LL	L	tt	t
ddZ	dZ	m	m	ts	ts
e	e	mm	m	tts	ts
e1	e	ng	n	tS	tS
E	e	nf	n	ttS	tS
E1	e	n	n	u	u
f	f	nn	n	u1	u
ff	f	o	o	v	v
g	g	o1	o	vv	v
gg	g	O	o	w	w
i	i	O1	o	z	z
i1	i	p	p	sil	sil

Table 1: 60 and 29 phones set (SAMPA).

Considering that, in unstressed position, the oppositions /e/ - /E/ and /o/ - /O/ are often neutralized in the Italian language, it was decided to merge these couples of phonemes. Since the occurrences of /E/ and /O/ phonemes were so rare in the test set, this simplification have had no influence in the test results.

Then, the acoustic differences between stressed (a1, e1, E1, i1, o1, O1, u1) and unstressed vowels (a, e, E, i, o, O, u) in Italian are subtle and mostly related to their duration. Furthermore, most of the Italian people pronounce vowels according to their regional influences instead of "correct-standard" pronunciation, if any, and this sort of inaccuracies is quite common. For these reasons, recognition outputs have been evaluated using the full 60-phones ArtiPhon set as well as a more realistic reduced 29-phones set, which do not count the mistakes between stressed and unstressed vowels, geminates vs

single phones and /ng/ and /nf/ allphones vs the /n/ phoneme.

In Table 2, the results of the EVALITA 2016 ArtiPhon speaker dependent experiment with the 60-phones and 29-phones are summarized in Table 2a and 2b respectively, for all the KALDI ASR engines, as in the TIMIT recipe.

Training & Decoding	%PCR	%SUB	%DEL	%INS	%PER
MonoPhone	mono	61.9	29.9	8.2	3.3	41.3
Delta + Delta-Deltas	tri1	66.4	25.9	7.6	2.4	35.9
LDA + MLTT	tri2	66.1	25.8	8.1	2.5	36.4
LDA + MLTT + SAT (SI)	tri3.si	67.1	25.4	7.5	2.6	35.5
LDA + MLTT + SAT	tri3	67.9	25.5	6.6	1.9	34.0
sgmm2_4: SGMM2	sgmm2_4	68.7	24.5	6.8	1.7	33.1
MMI + SGMM2 (iteration n.1)	sgmm2_4_mmi_b0.1	68.8	24.6	6.6	1.7	32.9
MMI + SGMM2 (iteration n.2)	sgmm2_4_mmi_b0.2	68.7	24.6	6.7	1.6	32.9
MMI + SGMM2 (iteration n.3)	sgmm2_4_mmi_b0.3	68.7	24.5	6.8	1.6	32.9
MMI + SGMM2 (iteration n.4)	sgmm2_4_mmi_b0.4	68.8	24.5	6.8	1.6	32.8
DNN Hybrid (Dan's)	tri4-nnet	64.6	27.7	7.7	3.0	38.3
SGMM + DNN Hybrid (Dan's) (it. 1)	combine_2 (1)	67.8	25.4	6.8	2.1	34.3
SGMM + DNN Hybrid (Dan's) (it. 2)	combine_2 (2)	68.3	25.4	6.3	2.6	34.3
SGMM + DNN Hybrid (Dan's) (it. 3)	combine_2 (3)	68.4	25.4	6.3	2.5	34.1
SGMM + DNN Hybrid (Dan's) (it. 4)	combine_2 (4)	68.1	25.3	6.6	2.3	34.2
DNN Hybrid (Karel's)	dnn4_pretrain-dbn_dnn	67.2	24.8	8.0	1.7	34.5
DNN Hybrid (Karel's), sMBR training (it. 1)	dnn4_pretrain-dbn_dnn_smbr (1)	67.1	24.8	8.1	1.8	34.7
DNN Hybrid (Karel's), sMBR training (it. 6)	dnn4_pretrain-dbn_dnn_smbr (6)	67.6	24.9	7.5	2.1	34.5

Table 2a: results for the EVALITA 2016 ArtiPhon speaker dependent task in the 60-phones case.

Training & Decoding	%PCR	%SUB	%DEL	%INS	%PER
MonoPhone	mono	80.1	11.0	8.9	2.6	22.4
Delta + Delta-Deltas	tri1	85.4	7.7	6.9	2.6	17.2
LDA + MLTT	tri2	85.8	7.3	6.9	2.4	16.6
LDA + MLTT + SAT (SI)	tri3.si	85.2	7.5	7.3	2.7	17.6
LDA + MLTT + SAT	tri3	86.7	6.5	6.8	2.1	15.3
sgmm2_4: SGMM2	sgmm2_4	87.2	6.5	6.3	2.3	15.1
MMI + SGMM2 (iteration n.1)	sgmm2_4_mmi_b0.1	87.2	6.5	6.3	2.3	15.1
MMI + SGMM2 (iteration n.2)	sgmm2_4_mmi_b0.2	86.8	6.3	6.9	1.9	15.0
MMI + SGMM2 (iteration n.3)	sgmm2_4_mmi_b0.3	87.4	6.3	6.3	2.3	14.9
MMI + SGMM2 (iteration n.4)	sgmm2_4_mmi_b0.4	87.4	6.3	6.3	2.3	14.9
DNN Hybrid (Dan's)	tri4-nnet	82.8	8.5	8.8	2.4	19.7
SGMM + DNN Hybrid (Dan's) (it. 1)	combine_2 (1)	87.4	6.1	6.5	2.5	15.1
SGMM + DNN Hybrid (Dan's) (it. 2)	combine_2 (2)	87.4	6.1	6.5	2.5	15.1
SGMM + DNN Hybrid (Dan's) (it. 3)	combine_2 (3)	87.3	6.1	6.6	2.5	15.2
SGMM + DNN Hybrid (Dan's) (it. 4)	combine_2 (4)	87.3	6.1	6.6	2.5	15.2
DNN Hybrid (Karel's)	dnn4_pretrain-dbn_dnn	86.1	6.8	7.1	2.3	16.2
DNN Hybrid (Karel's), sMBR training (it. 1)	dnn4_pretrain-dbn_dnn_smbr (1)	86.1	6.8	7.2	2.3	16.2
DNN Hybrid (Karel's), sMBR training (it. 6)	dnn4_pretrain-dbn_dnn_smbr (6)	86.0	6.6	7.4	2.2	16.2

Table 2b: results for the EVALITA 2016 ArtiPhon speaker dependent task in the 29-phones case.

	Training & Decoding	%PCR	%SUB	%DEL	%INS	%PER
MonoPhone	mono	61,3	23.8	14.9	2.3	41.0
Delta + Delta-Deltas	tri1	66.8	21.3	11.9	3.7	36.9
LDA + MLTT	tri2	70.0	19.5	10.4	4.6	34.5
LDA + MLTT + SAT (SI)	tri3.si	70.2	18.3	11.5	2.2	32.0
LDA + MLTT + SAT	tri3	74.5	16.8	8.7	3.0	28.4
sgmm2_4: SGMM2	sgmm2_4	75.7	15.3	9.0	4.4	28.7
MMI + SGMM2 (iteration n.1)	sgmm2_4_mmi_b0.1	75.7	15.2	9.1	4.1	28.4
MMI + SGMM2 (iteration n.2)	sgmm2_4_mmi_b0.2	76.3	15.6	8.1	4.7	28.4
MMI + SGMM2 (iteration n.3)	sgmm2_4_mmi_b0.3	76.3	15.5	8.2	4.5	28.2
MMI + SGMM2 (iteration n.4)	sgmm2_4_mmi_b0.4	76.3	15.4	8.3	4.6	28.2
DNN Hybrid (Dan's)	tri4-nnet	70.7	17.3	12.0	3.7	31.8
SGMM + DNN Hybrid (Dan's) (it. 1)	combine_2 (1)	76.1	15.2	8.7	3.7	27.5
SGMM + DNN Hybrid (Dan's) (it. 2)	combine_2 (2)	76.2	15.0	8.8	3.6	27.4
SGMM + DNN Hybrid (Dan's) (it. 3)	combine_2 (3)	76.1	15.1	8.8	3.5	27.4
SGMM + DNN Hybrid (Dan's) (it. 4)	combine_2 (4)	76.2	15.2	8.6	3.4	27.1
DNN Hybrid (Karel's)	dnn4_pretrain-dbn_dnn	75.6	14.6	9.8	2.4	26.9
DNN Hybrid (Karel's), sMBR training (it. 1)	dnn4_pretrain-dbn_dnn_smbr (1)	75.3	14.6	10.2	2.3	27.1
DNN Hybrid (Karel's), sMBR training (it. 6)	dnn4_pretrain-dbn_dnn_smbr (6)	75.6	14.7	9.7	2.5	27.0

Table 3: results for the EVALITA 2016 ArtiPhon speaker independent task in the 29-phones case.

The results of the EVALITA 2016 ArtiPhon speaker independent experiment using the acoustic models trained on APASCI with the 29-phones are summarized in Table 3.

All the systems are built on top of MFCC, LDA, MLLT, fMLLR with CMN features[1] - see (Rath, et al., 2013) for all acronyms references - obtained from auxiliary GMM (Gaussian Mixture Model) models. At first, these 40-dimensional features are all stored to disk in order to simplify the training scripts.

Moreover MMI, BMMI, MPE and sMBR[2] training are all supported - see (Rath et al., 2013) for all acronyms references.

KALDI currently contains also two parallel implementations for DNN (Deep Neural Networks) training: "DNN Hybrid (Dan's)" (Kaldi, WEB-b), (Zhang et al., 2014), (Povey et al., 2015) and "DNN Hybrid (Karel's)" (Kaldi, WEB-a), (Vesely et al., 2013) in Table 3. Both of them are DNNs 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(see Cosi et al. 2015, for a more detailed description).

The Phone Error Rate (PER) was considered for computing the score of the recognition process. The PER, which 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 was computed by the use of the NIST software SCLITE (sctk-WEB).

The results shown in Table 3 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),

[1] 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.

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

- 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 Models Training (Povey, 2009; Povey, 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).

5 Conclusions

As expected, due to the degraded clipped quality of the training and test audio signal, the 60-phones set is far from being realistic for obtaining optimum recognition performance even in the speaker dependent case (ArtiPhon training and test material).

On the contrary, if the reduced 29-phones set is used, the phone recognition performance is quite good and more than sufficient to build an effective ASR system if a language model could be incorporated.

Moreover, also in the speaker independent case (APASCI training material and ArtiPhon test material) the performance are not too bad even in these speech type and speaker mismatch conditions, thus confirming the effectiveness and the good quality of the system trained on APASCI material.

In these experiments, the DNNs results do not overcome those of the classic systems and we can hypothesize that this is due partially to the low quality of the signal, and also to the size of the corpus which is probably not sufficient to make the system learn all the variables characterizing the network. Moreover, the DNN architecture was not specifically tuned to the ArtiPhon data but instead the default KALDI architecture used in previous more complex speaker independent adult and children speech ASR experiments was simply chosen.

References

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

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

Bourlard H:A. & Morgan N., 1994. Connectionist Speech Recognition: a Hybrid Approach, volume 247. Springer.

Canevari, C., Badino, L., Fadiga, L., 2015. A new Italian dataset of parallel acoustic and articulatory data, Proceedings of INTERSPEECH, Dresden, Germany, 2015, 2152-2156.

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

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

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, 540-543, 2009.

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., Paci G., Sommavilla G., & Tesser F., 2015. KALDI: Yet another ASR Toolkit? Experiments on Italian Children Speech. In *Il farsi e il disfarsi del linguaggio. L'emergere, il mutamento e la patologia della struttura sonora del linguaggio,* 2015.

Dahl, G.E., Yu, D., Deng,L. & Acero, A., 2012. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, Jan. 2012, 20(1):30–42.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. & Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, Nov. 2012, 29(6):82-97.

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

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

Mohamed, A., Dahl, G.E. & Hinton, G., 2012. Acoustic Modeling Using Deep Belief Networks. IEEE Transactions on Audio, Speech, and Language Processing, Jan. 2012, 20(1):14-22.

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, vol. 25, no. 2, pp. 404–439, April 2011.

Povey, D., Ghoshal, A. et al., 2001. The KALDI Speech Recognition Toolkit, in Proceedings of ASRU, 2011 (IEEE Catalog No.: CFP11SRW-USB).

Povey, D., Zhang, X., & Khudanpur, S., 2014. Parallel Training of DNNs with Natural Gradient and Parameter Averaging, in Proceedings of ICLR 2015, International Conference on Learning Representations (arXiv:1410.7455).

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

sctk-WEB - Speech Recognition Scoring Toolkit https://www.nist.gov/itl/iad/mig/tools.

Serizel R., Giuliani D. (2014). Deep neural network adaptation for children's and adults' speech recognition. In Proceedings of ClicIt 2014, 1st Italian Conference on Computational Linguistics, Pisa, Italy, 2014.

Serizel R., Giuliani D. (2016). Deep-neural network approaches to speech recognition in heterogeneous groups of speakers including children, in Natural Language Engineering, April 2016.

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

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.