The ArtiPhon Task at Evalita 2016

Leonardo Badino
Center for Translational Neurophysiology of Speech and Communication
Istituto Italiano di Tecnologia – Italy
leonardo.badino@iit.it

Abstract

English. Despite the impressive results achieved by ASR technology in the last few years, state-of-the-art ASR systems 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 at Evalita 2016 we wanted to evaluate phone recognition systems in mismatched speaking styles. While training data consisted of read speech, most of testing data consisted of single-speaker hypo- and hyper-articulated speech. A second goal of the task was to investigate whether the use of speech production knowledge, in the form of measured articulatory movements, could help in building ASR systems that are more robust to the effects of the mismatch problem. Here I report the result of the only entry of the task and of baseline systems.

Italiano. *Nonostante i notevoli risultati ottenuti recentemente nel riconoscimento automatico del parlato (ASR) le prestazioni dei sistemi ASR peggiorano significativamente in quando le condizioni di testing sono differenti da quelle di training (per esempio quando il tipo di rumore acustico è differente). Un primo gol della ArtiPhon task ad Evalita 2016 è quello di valutare il comportamento di sistemi di riconoscimento fonetico in presenza di un mismatch in termini di registro del parlato. Mentre il parlato di training consiste di frasi lette ad un velocita; di eloquio "standard", il parlato di testing consiste di frasi sia iperche ipo-articolate. Un secondo gol della task è quello di analizzare se e come l'utilizzo di informazione concernente la produzione del parlato migliora l'accuratezza dell'ASR e in particolare nel caso di mismatch a livello di registri del parlato. Qui riporto risultati dell'unico sistema che è stato sottomesso e di una baseline.*

1 Introduction

In the last five years ASR technology has achieved remarkable results, thanks to increased training data, computational resources, and the use of deep neural networks (DNNs, (LeCun et al., 2015)). However, the performance of connectionist ASR degrades when testing conditions are different from training conditions (e.g., acoustic environments are different) (Huang et al., 2014). This is usually referred to as the training-testing mismatch problem. This problem is partly masked by multi-condition training (Seltzer et al., 2013) that consists in using very large training datasets (up to thousands of hours) of transcribed speech to cover as many sources of variability as possible (e.g., speaker's gender, age and accent, different acoustic environments).

One of the two main goals of the ArtiPhon task at Evalita 2016 was to evaluate phone recognition systems in mismatched speaking styles. Between training and testing data the speaking style was the condition that differed. More specifically, while the training dataset 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. Training and testing data are from the same speaker.

The second goal of the ArtiPhon task was to investigate whether the use of speech production knowledge, in the form of measured articulatory movements, could help in building ASR systems that are more robust to the effects of the mismatch problem.

The use of speech production knowledge, i.e., knowledge about how the vocal tract behaves when it produces speech sounds, is motivated by the fact that complex phenomena

observed in speech, for which a simple purely acoustic description has still to be found, can be easily and compactly described in speech production-based representations. For example, in Articulatory Phonology (Browman and Goldstein, 1992) or in the distinctive features framework (Jakobson et al., 1952) coarticulation effects can be compactly modeled as temporal overlaps of few vocal tract gestures. The vocal tract gestures are regarded as invariant, i.e., context- and speaker-independent, production targets that contribute to the realization of a phonetic segment. Obviously the invariance of a vocal tract gesture partly depends on the degree of abstraction of the representation but speech production representations offer compact descriptions of complex phenomena and of phonetic targets that purely acoustic representations are not able to provide yet (Maddieson, 1997).

Recently, my colleagues and I have proposed DNN-based "articulatory" ASR where the DNN that computes phone probabilities is forced, during training, to learn/use motor features. We have proposed strategies that allow motor information to produce an inductive bias on learning. The bias resulted in improvements over strong DNN-based purely auditory baselines, in both speaker-dependent (Badino et al., 2016) and speaker-independent settings (Badino, 2016)

Regarding the Artiphon task, unfortunately only one out of the 6 research groups that expressed an interest in the task actually participated (Piero Cosi from ISTC at CNR, henceforth I will refer to this participant as ISTC) (Cosi, 2016). The ISTC system did not use articulatory data.

In this report I will present results of the ISTC phone recognition systems and of baseline systems that also used articulatory data.

2 Data

The training and testing datasets used for the ArtiPhon task were selected from voice cnz of the Italian MSPKA corpus (http://www.mspkacorpus.it/) (Canevari et al., 2015), which was collected in 2015 at the Istituto Italiano di Tecnologia (IIT).

The training dataset corresponds to the 666-utterance session 1 of MPSKA, where the speaker was required to keep a constant speech rate. The testing dataset was a 40-utterance subset selected from session 2 of MPSKA. Session 2 of MPSKA contains a continuum of ten descending articulation degrees, from hyper-articulated to hypo-articulated speech. Details on the procedure used to elicit this continuum are provided in (Canevari et al., 2015).

Articulatory data consist of trajectories of 7 vocal tract articulators and recorded with the NDI (Northern Digital Instruments, Canada) wave speech electromagnetic articulography system at 400 Hz.

Seven 5-Degree-of-freedom (DOF) sensor coils were attached to upper and lower lips (UL and LL), upper and lower incisors (UI and LI), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD). For head movement correction a 6-DOF sensor coil was fixed on the bridge of a pair of glasses worn by the speakers.

The NDI system tracks sensor coils in 3D space providing 7 measurements per each coil: 3 positions (i.e. x; y; z) and 4 rotations (i.e. Q0;Q1;Q2;Q3) in quaternion format with Q0 = 0 for 5-DOF sensor coils.

Contrarily to other articulographic systems (e.g. Carstens 2D AG200, AG100) speakers head is free to move. That increases comfort and the naturalness of speech.

During recordings speakers were asked to read aloud each sentence that is prompted on a computer screen. In order to minimize disfluencies speakers had time to silently read each sentence before reading out.

The audio files of the MSPKA corpus are partly saturated.

The phone set consists of 60 phonemes, although the participants could collapsed them to 48 phonemes as proposed in (Canevari et al., 2015).

3 Sub-tasks

In Artiphon sub-tasks are phone recognition tasks. The participants were asked to:

- train phone recognition systems on the training dataset and then run them on the test dataset;
- (optional) use articulatory data to build "articulatory" phone recognition systems.

Articulatory data were also provided in the test dataset thus three different scenarios were possible:

- Scenario 1. Articulatory data not available
- Scenario 2. Articulatory data available at both training and testing.

- Scenario 3. Articulatory data available only at training.

Note that only scenarios 1 and 3 are realistic ASR scenarios as during testing articulatory data are very difficult to access.

Participants could build purely acoustic and articulatory phone recognition systems starting from the Matlab toolbox developed at IIT, available at https://github.com/robotology/natural-speech.

4 Phone recognition systems

Baseline systems are hybrid DNN-HMM systems while ISTC systems are either GMM-HMM or DNN-HMM systems with DNN-HMM.

The ISTC systems were trained using the KALDI ASR engine. ISTC systems used either the full phone set (with 60 phone labels) or a reduced phone set (with 29 phones). In the reduced phone set all phones that are not actual phonemes in current Italian were correctly removed. However, important phonemes were also arbitrarily removed, most importantly, geminates and corresponding non-geminate phones were collapsed into a single phone (e.g., /pp/ and /p/ were both represented by label /p/).

ISCT systems used either monophones or triphones.

ISCT systems were built using KALDI (Povey et al., 2011) with TIMIT recipes adapted to the APASCI dataset (Angelini & al., 1994). Two training datasets where used:

- the single-speaker dataset provided within the ArtiPhon task;
- the APASCI dataset.

In all cases only acoustic data were used (scenario 1), so the recognition systems were purely acoustic recognition systems. Henceforth I will refer to ISTC systems trained on the ArtiPhon single-speaker training dataset as speaker-dependent ISTC systems (as the speaker in training and testing data is the same) and to ISTC systems trained on the APASCI dataset as speaker-independent ISTC systems.
Baseline systems were built using the aforementioned Matlab toolbox and only trained on the ArtiPhon training dataset (so they are all speaker-dependent systems).

Baseline systems used a 48 phone set and three-state monophones (Canevari et al., 2015). Baseline systems were trained and tested according to all three aforementioned three scenarios. The articulatory data considered only refer to x-y positions of 6 coils (the coil attached to the upper teeth was exluded).

5 Results

Here I report some of the most relevant results regarding ISTC and baseline systems.

Baseline systems and ISTC systems are not directly comparable as very different assumptions were made, most importantly they use different phone sets.

Additionally, ISCT systems were mainly concerned with exploring the best performing systems (created using well-known KALDI recipes for ASR) and comparing them in the speaker-dependent and in the speaker-independent case.

Baselines systems were created to investigate the utility of articulatory features in mismatched speaking styles.

5.1 ISCT systems

Here I show results on ISCT systems trained and tested on the 29 phone set. Table 1 shows results of the speaker-dependent systems while Table 2 shows results in the speaker-independent case.

The results shown in the two tables 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).

	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 1. Results of ISCT systems on speaker-dependent sub-task with the 29-phone set.
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. MMI: Maximum Mutual Information; BMMI: Boosted MMI; MPE: Minimum Phone Error; sMBR: State-level Minimum Bayes Risk

	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 2. Results of ISCT systems on speaker independent sub-task with 29-phone set.

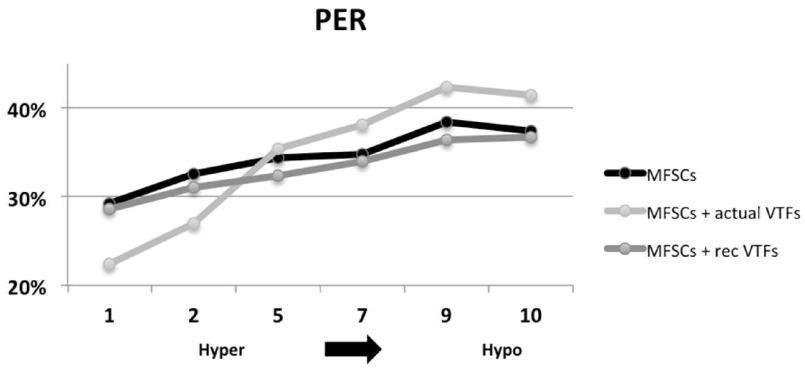

Figure 1. Phone Error Rate (PER) over 10 degrees of articulation when (i) using only MFSCs (black), (ii) using MFSCs appended to actual VTFs (light grey) and (iii) using MFSCs appended to recovered VTFs (dark grey)

The most interesting result is that while DNN-HMM systems outperform GMM-HMM systems in the speaker-independent case (as expected), GMM-HMM and more specifically sub-space GMM-HMM (Povey et al., 2011), outperform the DNN-based systems in the speaker dependent case.

Another interesting result is that sequence based training strategies (Vesely et al., 2013) did not produce any improvement over frame-based training strategies.

5.2 Baseline systems – acoustic vs. articulatory results

The baseline systems addressed the two main questions that motivated the design of the ArtiPhon task: (i) how does articulatory information contribute to phone recognition?; (ii) how does the phone recognition system performance vary along the continuum from hyper-articulated speech to hypo-articulated speech?

Figure 1 shows phone recognition results of 3 different systems over the 10 degrees of articulation from hyper-articulated to hypo-articulated speech.

The three systems, reflecting the three aforementioned training-testing scenarios, are:

- phone recognition system that only uses acoustic feature, specifically mel-filtered spectra coefficients (MFSCs, scenario 1)
- articulatory phone recognition system where actual measured articulatory/vocal tract features (VTFs) are appended to the input acoustic vector during testing (scenario 2)
- articulatory phone recognition system where reconstructed VTFs are appended to the input acoustic vector during testing (scenario 3)

The last system reconstructs the articulatory features using an acoustic-to-articulatory mapping learned during training (see, e.g., (Canevari et al., 2013)for details).

All systems used a 48-phone set as in cc.

One first relevant result is that all systems performed better at high levels of hyper-articulation than at "middle" levels (i.e., levels 5-6) which mostly corresponds to the training condition (Canevari et al., forthcoming). In all systems performance degraded from hyper- to hypo-articulated speech.

Reconstructed VTFs always decrease the phone error rate. Appending recovered VTFs to the acoustic feature vector produces a relative PER reduction that ranges from 4.6% in hyper-articulated speech, to 5.7% and 5.2% in middle- and hypo-articulated speech respectively.

Actual VTFs provide a relative PER reduction up to 23.5% in hyper-articulated speech, whereas, unexpectedly, no improvements are observed when actual VTFs are used in middle- and hypo-articulated speech. That might due to the fact that sessions 1 and 2 of the MSPKA corpus took place in different days so EMA coils could be in slightly different positions.

6 Conclusions

This paper described the ArtiPhon task at Evalita 2016 and showed and discussed results of baseline phone recognition systems and of the submitted systems.

References

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

Badino, L. 2016. Phonetic Context Embeddings for DNN-HMM Phone Recognition. In *Proceedings of Interspeech*. San Francisco, CA.

Badino, L., Canevari, C., Fadiga, L., and Metta, G. 2016. Integrating Articulatory Data in Deep Neural Network-based Acoustic Modeling. *Computer Speech and Language*, 36, 173–195.

Browman, C., and Goldstein, L. 1992. Articulatory phonology: an overview. *Phonetica* 49 (3–4), 155–180.

Canevari, C., Badino, L., and Fadiga, L. 2015. A new Italian dataset of parallel acoustic and articulatory data. *Proceedings of Interspeech*. Dresden.

Canevari, C., Badino, L., D'Ausilio, A., and Fadiga, L. Forthcoming. Analysis of speech production differences between Hypo-and Hyper-articulated speech and implications for Articulatory ASR. Submitted.

Canevari, C., Badino, L., Fadiga, L., and Metta, G. 2013. Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR sytem using EMA data. In *Proceedings of Workshop on Speech Production for Automatic Speech Recognition*. Lyon, France.

Cosi, P. 2016. Phone Recognition Experiments on ArtiPhon with KALDI. *In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Huang, Y., Yu, D., Liu, C., and Gong, Y. 2014. A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models. *Proceedings of Interspeech*. Singapore.

Jakobson, R., Fant, G., and Halle, M. 1952. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge,MA: MIT Press.

LeCun, Y., Bengio, Y., and Hinton, G. E. 2015. Deep Learning. *Nature*, 521, 436-444.

Maddieson, I. 1997. Phonetic universals. In W. Hardcastle, and J. Laver, *The Handbook of Phonetic Sciences*. pp. 619--639. Oxford: Blackwell Publishers.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow , A., Rosei, R. C., Schwarzb, P., and Thomash, S. 2011. The Subspace Gaussian Mixture Model - a Structured Model for Speech Recognition. *ComputerSpeech and Language*. 25(2), 404-439.

Povey, D., Ghoshal, A., & al. 2011. The KALDI Speech Recognition Toolkit . *Proceedings of ASRU2011*.

Rath, S. P., Povey, D., Vesely, K., and Cernocky, J. 2013. Improved feature processing for Deep Neural Networks. *Proceedings of Interspeech*, pp. 109--113. Lyon, France.

Seltzer, M., Yu, D., and Wang, Y. 2013. An Investigation of Deep Neural Networks for Noise Robust Speech Recognition. *Proceedings of ICASSP*. Vancouver, Canada.

Vesely, K., Ghoshal, A., Burget, L., and Povey, D. 2013. Sequence-discriminative training of deep neural networks. *Proceeding of Interspeech*. Lyon, France.