A new Italian dataset of parallel acoustic and articulatory data

Claudia Canevari, Leonardo Badino, Luciano Fadiga

Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia, Genova, Italy

claudia.canevari@gmail.com, leonardo.badino@iit.it, luciano.fadiga@iit.it

Abstract

In this paper we introduce a new Italian dataset consisting of simultaneous recordings of continuous speech and trajectories of important vocal tract articulators (i.e. tongue, lips, incisors) tracked by Electromagnetic Articulography (EMA). It includes more than 500 sentences uttered in citation condition by three speakers, one male (cnz) and two females (lls, olm), for approximately 2 hours of speech material.
Such dataset has been designed to be large enough and phonetically balanced so as to be used in speech applications (e.g. speech recognition systems).
We then test our speaker-dependent articulatory Deep-Neural-Network Hidden-Markov-Model (DNN-HMM) phone recognizer on the set of data recorded from the cnz speaker.
We show that phone recognition results are comparable to the ones that we previously obtained using two well-known British-English datasets with EMA data of equivalent vocal tract articulators. That suggests that the new set of data is a equally useful and coherent resource.
The dataset is the session 1 of a larger Italian corpus, called Multi-SPeaKing-style-Articulatory (MSPKA) corpus, including parallel audio and articulatory data in diverse speaking styles (e.g. read, hyperarticulated and hypoarticulated speech). It is freely available at http://www.mspkacorpus.it for research purposes. In the immediate future the whole corpus will be released.

Index Terms: Electromagnetic Articulography, Articulatory Corpora, Acoustic-to-Articulatory Mapping, Phone Recognition, Deep-Neural-Networks

1. Introduction

Measuring the movements of the actual vocal tract articulators during speech production provides valuable information for research related to speech technologies and helps to understand the human speech production itself.
A lot of approaches were proposed to exploit the articulatory information in speech applications such as Automatic Speech Recognition (ASR) systems (see [12] for a review), silent speech interfaces ([10]), speech synthesizers ([13]), systems for speech therapy ([11]) and head motion synthesizers ([3]).
Although various acquisition techniques (e.g. real-time Magnetic Resonance (rtMR), ultrasound, Electromagnetic Articulography (EMA), X-ray microbeam system, Electropalatography (EPG), Laryngography) can be employed the collection of vocal tract articulatory data remains much more difficult than speech recording. For example, in the case of EMA, sensor coils can come detached and need to be re-glued. That can produce incoherent measurements ([16]). Furthermore due to the invasive nature of some techniques speakers become fatigued soon and often produce many disfluencies.
For such reasons the existing corpora of parallel acoustic and articulatory data are few. Most of them is limited to English language and only contains read speech (i.e. speech in citation condition with fairly uniform acoustic characteristics opposite to those of spontaneous speech).
Three well-known freely available articulatory corpora including continuous speech in citation condition are two British-English datasets, the MOCHA-TIMIT ([18]) and the mngu0 ([17]), and one American-English dataset, the USC-TIMIT ([15]). All of them contain a large and phonetically balanced reading material (i.e. MOCHA-TIMIT and USC-TIMIT contain the same 460 sentences while mngu0 includes over 2000 sentences) and articulatory data acquired in different modalities (e.g. USC-TIMIT includes rtMR images of the vocal tract and EMA data while MOCHA-TIMIT contains speech production data collected by means of EPG, EMA and Laryngography) from a few speakers (i.e. USC-TIMIT and MOCHA-TIMIT consist of the recordings from 10 and 2 speakers respectively while mngu0 contains data from only one speaker).
The unique Italian existing articulatory corpus, the "Lecce corpus", consists of single-word and pseudo-word utterances pronounced by 9 different speakers ([9]). Vocal tract articulators were tracked by using EMA, ultrasound and EPG and each word was pronounced as a declaration and a question. Such corpus is not phonetically rich and large enough to train speech applications although it provides some speech variability and can be useful for studying human speech production.
Recently we collected a new Italian articulatory corpus, called Multi-SPeaKing-style-Articulatory (MSPKA) corpus, consisting of simultaneous recordings of continuous speech and important vocal tract articulators (i.e. lips, tongue, incisors) exclusively tracked by EMA in diverse speaking styles (e.g. read speech, hypoarticulated speech, hyperarticulated speech) over two sessions from three different speakers.
In addition to a large and phonetically balanced set of parallel acoustic and articulatory data for Italian language such corpus provides a much more speech variability than the other existing articulatory corpora. That allows to study the actual vocal tract behaviour in different speaking styles using an extensive set of data. Moreover it is a necessary requirement to better exploit the full potential of speech production information in real speech applications that have to address speech with variabilities (e.g. spontaneous speech).
In the present paper we introduce the session 1 of the MSPKA corpus, only including read speech.
Since obtaining disfluent and unclear speech or inconsistent measurements of the vocal tract articulators ([16]) when using EMA is very likely, we also try to assess the quality of our new articulatory corpus (session 1) through a "task-oriented" evaluation using an articulatory hybrid Deep-Neural-Network Hidden-Markov-Model (DNN-HMM) phone recognition

system in speaker-dependent settings.
In our previous studies we tested such system on different corpora with EMA data of equivalent vocal tract articulators ([1], [2], [7], [8]). These studies provide phone recognition results to use as benchmark.
The articulatory DNN-HMM phone recognizer receives as input measured articulatory features (AFs) appended to speech acoustics. The benefits that the additional AFs can provide are evaluated w.r.t. an equivalent DNN-HMM system that uses speech acoustics only (i.e. the acoustic baseline).
DNNs are both used to estimate the phone posterior probabilities and to carry out the Acoustic-to-Articulatory Mapping (AAM) which recovers the AFs from speech acoustics. Recovering AFs is necessary in realistic scenarios where articulatory data are only available during training.
This paper is organized as follows. Sections 2 and 3 describe the session 1 of the MSPKA corpus and the processing of audio and articulatory data. Sections 4-6 concern the phone recognition experiments. Section 4 is about the experimental setup and sections 5 and 6 deal with the articulatory recovery, phone recognition results and discussion. Section 7 describes the free availability of the MSPKA corpus on a dedicated website and section 8 is about conclusion.

2. Multi-SPeaKing-style-Articulatory (MSPKA) corpus: session 1

The session 1 of the MSPKA corpus consists of the audio and articulatory data collected from three native speakers of Italian, one male (cnz) and two females (lls, olm) in citation condition. It includes more than 500 Italian sentences for approximately 2 hours of speech material (table 1). The sentences were selected from a lot of popular novels and text data available on the web in order to provide a balanced phonetic coverage ([4]) and maximize the speech register variety.
The trajectories of 7 vocal tract articulators and the audio signal were recorded simultaneously using the NDI (Northern Digital Instruments, Canada) wave speech electromagnetic articulography system at 400 Hz and 22050 Hz sampling rate respectively.
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. Q_0, Q_1, Q_2, Q_3) in quaternion format with $Q_0 = 0$ 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 read in their head each sentence before reading out. Mispronounced phrases were prompted until speakers produce them correctly.

dataset	utterances
cnz	666
lls	629
olm	501
msak0	460
mngu0 (day 1)	1354

Table 1: *Number of recorded sentences per speaker in the MSPKA corpus (session 1). For comparison with our previous studies ([1], [2], [7], [8]) we also reported the number of sentences uttered by the msak0 speaker in the MOCHA-TIMIT and the male speaker in the mngu0 corpus (day1).*

3. Data Processing

The acoustic and articulatory data were processed as follows.
We applied an amplitude spectral noise subtraction strategy on the speech signal as described in [5] in order to suppress the background noise produced by the transmitter electromagnets of the articulograph. We computed the noise spectrum in the first 25 ms of each audio recording where there was no speech activity.
Concerning the articulatory data we filtered the vocal tract articulator trajectories (i.e. x, y, z positions of the sensor coils) using an adaptive median filter with a window from 10 ms to 50 ms and a smooth elliptic low-pass filter with 20 Hz cutoff frequency. We did not consider the rotations of sensor coils.
We phonetically transcribed the set of sentences using the Italian letter-to-sound tool for the Festival Speech Synthesis System ([21]). The phone set consists of 49 Italian phones including "silence" to mark the voice activity.
The phone boundaries were estimated using HCompV, HERest and HVite functions of the HTK ([19]) and manually checked using PRAAT speech processing software ([20]).

4. Experimental Setup

We used the set of audio and articulatory data recorded from cnz speaker in the MSPKA corpus (session 1) for the phone recognition experiments presented in this paper.
We extracted 60 mel-filtered spectral coefficients (MFSCs) from speech signal as in [1]. We used MFSCs as acoustic input for the AAM and as observations in the DNN-HMM phone recognition system. In our previous studies using MFSCs ([8]) rather than standard MFCCs ([1], [2], [7]) as acoustic observations turned out to produce slightly better results on the acoustic baseline (i.e. a DNN-HMM phone recognizer that uses acoustic observations only) in the msak0 dataset of the MOCHA-TIMIT. That is in agreement with some studies on speech recognition based on DNNs (see, e.g. [14]).
Concerning the articulatory features we used the midsagittal positions plus their first and second derivatives of LI, UL, LL, TT, TB and TD (for an overall of 36 AFs). We imposed the same time window of the acoustic features.
All the vocal tract articulator trajectories were normalized w.r.t. the UIs that exhibit very small variations.
Training and testing sets were created using a 5-fold cross-validation criterion: the test fold 1 consists of sentences numbered 1, 6, 11, 16 ..., the test fold 2 consists of sentences numbered 2, 7, 12 ... and so on. The corresponding remaining sentences in each test fold are used for training.

4.1. Acoustic-to-Articulatory-Mapping

The Acoustic-to-Articulatory Mapping was performed using the DNN-based-AAM strategy in the simplest configuration ([1], [2]). The DNN is a 3-hidden layer net with 300 nodes per each hidden layer. The input units of the corresponding DBN were Gaussian-distributed while all hidden units were binary. The input consists of 5 acoustic features frames (60 MFSCs x5) and the output is the frame of 36 AFs corresponding to the frame on which the acoustic input is centred on.

4.2. Phone recognizer

We used 3 states per phone. The state boundaries were computed in the training utterances using the HInit, HRest and HERest functions of the HTK ([19]).
The state posteriors were estimated by a 3-hidden layer DNN, with 9 vectors of MFSCs (60 x 9 MFSCs) and the corresponding 9 vectors of AFs (36 x 9 AFs), when AFs were used, as input units. Each hidden layer has 1500 units while the output layer has 144 units (48 Italian phonemes x 3 states).
In order to estimate the phone sequence for each test utterance we first computed the phone unigrams and bigrams and the state bigrams on the training data for each fold through the CMU toolkit ([22]). Then the state posteriors plus phone unigrams and bigrams and state bigrams were fed into the Viterbi decoder.
Theoretically the Viterbi decoder should receive the state emission probabilities (i.e. the state posteriors divided by the state priors). However that resulted into slightly lower phone recognition performance in agreement with our previous studies ([1], [2], [7], [8]).

5. Results

5.1. Articulatory Reconstruction

The AF reconstruction was evaluated in terms of Root Mean Square Error (RMSE) and Pearson Product Moment Correlation (r). Table 2 compares the reconstruction of the 36 AFs in the cnz dataset with the one previously obtained in the msak0 dataset of the MOCHA-TIMIT and the mngu0 corpus (day 1) using the same DNN-AAM strategy ([8]).
The three datasets were compared on an equivalent articulatory feature set (i.e. the midsagittal positions, velocities and accelerations of UL, LL, LI, TT, TB, TD).
UIs were reconstructed as any other articulators from speech acoustics in our previous studies on the msak0 dataset ([1], [2], [7], [8]). Here for an unbiased comparison we did not include their contributions to the overall reconstruction evaluation in the table 2.

Dataset	RMSE	r
cnz	0.617	0.778
msak0	0.650	0.750
mngu0 (day 1)	0.542	0.837

Table 2: *Articulatory reconstruction results in terms of Root Mean Square Error (RMSE) and Pearson product moment correlation (r) averaged on all 5 folds of the cnz dataset in the MSPKA corpus (session 1). We also reported the articulatory reconstruction results previously obtained on the msak0 voice of the MOCHA-TIMIT and the mngu0 corpus (day 1) ([8]). For an unbiased comparison the contributions of the upper incisor positions, velocities and accelerations in the msak0 dataset were excluded.*

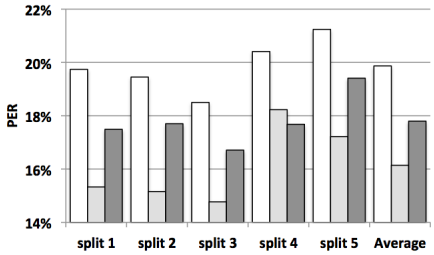

Figure 1: Phone Error Rate (PER) in each fold and on average for the cnz dataset using the following feature sets: MFSCs (white), MFSCs + actual AFs (light grey), MFSCs + recovered AFs (dark grey).

The articulatory reconstruction in the cnz dataset outperforms that in the msak0 dataset and is outperformed by the one in the mngu0 corpus.

5.2. Phone Recognition

Table 3 shows the Frame-wise Classification Accuracy (FwCa) and the Phone Error Rate (PER) for the DNN-HMM phone recognizer using speech acoustics only or appended to actual or recovered articulatory features in the cnz dataset. For comparison it also reports the phone recognition results for the msak0 dataset of the MOCHA-TIMIT and mngu0 corpus (day 1) ([8]).
In the cnz dataset recovered AFs produced a 10.4% relative PER reduction w.r.t. the acoustic baseline. A perfect articulatory reconstruction would lead to a 18.8% relative PER reduction. The additional AFs improved the phone recognition in each fold (figure 1).
In the msak0 and mngu0 datasets actual and recovered AFs leaded up to 25% and 9.8% relative PER reductions respectively ([8]).

Feature set	cnz		msak0		mngu0 (day 1)	
	FwCA %	PER	FwCA %	PER	FwCA %	PER
MFSCs	76.4	19.9	68.0	30.0	83.6	13.4
MFSCs + actual AFs	79.2	16.1	74.9	22.5	87.5	10.7
MFSCs + rec AFs	78.6	17.8	70.8	28.2	85.5	12.1

Table 3: *Frame-wise phone classification accuracy (FwCa) and phone error rate (PER) for the cnz voice of the MSPKA dataset using MFSCs only, MFSCs and actual AFs, MFSCs and reconstructed AFs. Values are averaged over the 5 folds. We also reported the FwCa and the PER previously obtained on the msak0 voice of the MOCHA-TIMIT and the mngu0 corpus (day 1) ([8]).*

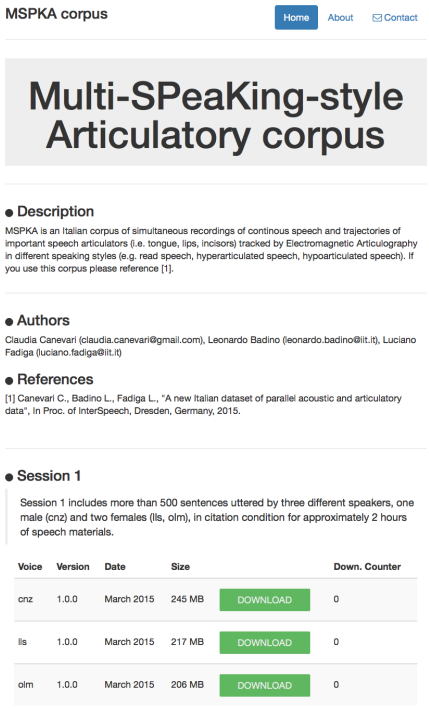

Figure 2: Screenshot of the website dedicated to the free distribution of the MSPKA corpus.

6. Discussion

Phone recognition results for the cnz dataset are comparable to our previous results for the msak0 dataset of the MOCHA-TIMIT and the mngu0 corpus ([8]). That suggests that the new set of data is a equally useful and coherent resource (that was proved using the set of data recorded from the cnz speaker only) .

The highest relative PER reductions are observed in the msak0 and cnz datasets using actual and recovered AFs respectively. It is hard to establish the factors that produce higher benefits of using additional articulatory observations among the three datasets.

The mngu0 corpus is claimed to provide very reliable articulatory measurements ([17]) and was reported to perform very well in the AAM problem ([16]). While in the MOCHA-TIMIT corpus some articulatory inconsistencies are well documented mostly for fsew0 dataset ([16], [17]). Note that here we used the msak0 dataset of the MOCHA-TIMIT.

We know that reconstructed AFs can improve the phone recognition if their recovery is good enough ([6], [8]). However more accurate articulatory reconstruction does not always correspond to higher phone recognition improvements. That was already reported in our previous study where we experimented with different AAM strategies for phone recognition only on the msak0 dataset ([1], [2]). Here we observe that the AFs are reconstructed much better in the mngu0 dataset, but do not produce the highest relative PER reduction.

It is more likely that also factors such as language and phone set can play an important role. Previously we showed that the recognition of some phonetic categories in a given language or dataset more positively benefits from using additional articulatory observations than others in a different language or dataset ([8]). Note that even the two British-English datasets (i.e. mngu0 and msak0 datasets) use slightly different phone sets (different number of allophones for the same phone).

Finally it is worth to point out that supposedly the better the acoustic baseline, the more difficult it is to improve it.

7. MSPKA corpus distribution

The session 1 of the MSPKA corpus is freely available at http://www.mspkacorpus.it for research purposes. In the immediate future the whole corpus will be released.

Figure 2 shows the screenshot of the website dedicated to the distribution of the corpus.

Processed audio and articulatory data together to phonetic labelling are provided.

The articulatory data consist of the x, y, z positions of the seven 5-DOF coil sensors. The rotation values are not included.

It is our plan to continuously enhance the corpus with data acquired from more speakers having different accents in more diverse speaking styles.

8. Conclusion

In this paper we introduced the session 1 of our new Multi-SPeaKing-style Articulatory (MSPKA) Italian corpus and tested its goodness in a phone recognition task using the set of parallel audio and articulatory data from only one speaker.

We then made know that such session is currently freely available on a dedicated website and the whole corpus will be released in the immediate future.

9. Acknowledgements

The authors thank Alessandro Bruchi and Luca Ghigliotti for their support on hosting and maintenance of the website dedicated to the free distribution of the MSPKA corpus.

10. References

[1] Badino, L., Canevari, C., Fadiga, L. and Metta, G., *Deep-level acoustic-to-articulatory mapping for DBN-HMM based phone recognition*, in Proc. IEEE SLT 2012, Miami, Florida, 2012.

[2] Badino, L., Canevari, C., Fadiga, L. and Metta, G., *Deep-Level Acoustic-to-Articulatory Mapping for DBN-HMM Based Phone Recognition - Erratum*. Available at http://www.rbcs.iit.it/online/badino_et_al_slt2012_erratum.pdf

[3] Ben-Youssef A., Shimodaira H., Braude D. A., *Articulatory features for speech-driven head motion synthesis*, in Proc. Interspeech, Lyon, France, 2013.

[4] Berry J., Fadiga L., *Data-driven Design of a Sentence List for an Articulatory Speech Corpus*, in Proc. Interspeech, Lyon, France, 2013.

[5] Boll S.F., *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Transactions on acoustics, speech and signal processing, 27, pp. 113-120, 1979.

[6] Canevari C., Badino L., Fadiga F., Metta G., *Modelling speech imitation and ecological learning of auditory-motor maps*, Frontiers in Psychology, 4(364), pp. 1-12, 2013.

[7] Canevari C., Badino L., Fadiga L., Metta G., *Relevance-weighted reconstruction of articulatory features in Deep Neural Network-based Acoustic-to-Articulatory Mapping*, in Proc. of Interspeech, Lyon, France, 2013.

[8] Canevari C., Badino L., Fadiga L., Metta G., *Cross-corpus and cross-linguistic evaluation of speaker-dependent DNN-HMM ASR system using EMA data*, in Proc. SPASR - Speech Production in Automatic Speech Recognition, Lyon, France, 2013.

[9] Grimaldi M., Gili Fivela B., Sigona F., Tavella M., Fitzpatrick P., Craighero L., Fadiga L., Sandini G., Metta G., *New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph*, in Proc. LangTech., Rome, Italy, 2008.

[10] Hueber T., Benaroya E.L., Chollet G., Denby B., Dreyfus G., Stone M., *Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips*, Speech Communication, 52, pp. 288-200, 2010.

[11] Hueber T., Bailly G., Badin P., Elisei F., *Speaker Adaptation of an Acoustic-Articulatory Inversion Model using Cascaded Gaussian Mixture Regressions*, in Proc. Interspeech, Lyon, France, 2013.

[12] King S., Frankel J., Livescu K., McDermott E., Richmond K., Wester M., *Speech production knowledge in automatic speech recognition*, J. Acoust. Soc. of Am., 121(2), 2007.

[13] Ling Z.H., Richmond K., J. Yamagishi, *Articulatory Control of HMM-Based Parametric Speech Synthesis Using Feature-Space Switched Multiple Regression*, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 21 (1), 2013.

[14] Mohamed, A., Hinton, G. E. and Penn, G. *Understanding how Deep Belief Networks perform acoustic modeling*. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4273-4276, 2012.

[15] Narayanan S., Toutios A., Ramanarayanan V., Lammert A., Kim J, Lee S., Nayak K., Kim Y.C., Zhu Y., Goldstein L., Byrd D., Bresch E., Ghosh P., Katsamanis A., Proctor M., *Real-Time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)*, J. Acoust. Soc. of Am., 136 (3), 2014.

[16] Richmond, K., *Preliminary inversion Mapping Results with a New EMA Corpus*, in Proc. Interspeech, Brighton, UK, 2009.

[17] Richmond, K., Hoole, P., King, S., *Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory corpus*, in Proc. Interspeech, Florence, Italy, 2011.

[18] Wrench A., *The MOCHA-TIMIT articulatory database*, http://www.cstr.ed.ac.uk/artic mocha.html, 1999.

[19] Available at http://htk.eng.cam.ac.uk

[20] Available at http://www.fon.hum.uva.nl/praat

[21] Available at http://www2.pd.istc.cnr.it/FESTIVAL

[22] Available at http://www.speech.cs.cmu.edu/SLM/toolkit.html