Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus

Korin Richmond[1], *Phil Hoole*[2] *and Simon King*[1]

[1]The Centre for Speech Technology Research, Informatics Forum, Edinburgh University, UK
[2]Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, Munich, Germany
korin@cstr.ed.ac.uk, hoole@phonetik.uni-muenchen.de, Simon.King@ed.ac.uk

Abstract

This paper serves as an initial announcement of the availability of a corpus of articulatory data called mngu0. This corpus will ultimately consist of a collection of multiple sources of articulatory data acquired from a single speaker: electromagnetic articulography (EMA), audio, video, volumetric MRI scans, and 3D scans of dental impressions. This data will be provided free for research use. In this first stage of the release, we are making available one subset of EMA data, consisting of more than 1,300 phonetically diverse utterances recorded with a Carstens AG500 electromagnetic articulograph. Distribution of mngu0 will be managed by a dedicated "forum-style" web site. This paper both outlines the general goals motivating the distribution of the data and the creation of the mngu0 web forum, and also provides a description of the EMA data contained in this initial release.

Index Terms: articulography, corpus, EMA

1. Introduction

Speech can be viewed as having two parallel, interrelated representations: the acoustic domain, in which the speech signal is transmitted between speaker and listener, and the articulatory domain in which the speech signal is formed. Although the majority of speech research has focused on the speech signal in the acoustic domain, a significant sub-field of speech research has investigated ways to exploit articulatory representations to improve both speech technology and our understanding of speech. For example, [1] and [2] review many attempts to incorporate articulatory information in automatic speech recognition (ASR) systems, while [3] describes an articulatorily controllable HMM-based speech synthesis system. The articulatory representation of speech employed in such research has taken multiple forms, ranging from symbolic features derived from phone labels through to direct measurements of human articulators. Several articulography methods have been employed to acquire the latter, such as electropalatography (EPG), electromagnetic articulography (EMA), X-ray cinematography, ultrasound and MRI.

Two well-known freely available corpora of articulography data are the Wisconsin X-ray microbeam (XRMB) corpus [4] and the MOCHA EMA corpus [5]. Both these data sets have proved invaluable and have been used extensively in a broad range of research. The purpose of this paper is to announce the availability of a new corpus of articulatory data, called mngu0, which we hope will likewise be useful to other researchers and will come to be equally widely-used.

We begin in Section 2 with a closer look at the benefits brought by freely available articulatory data, citing several examples of research that has been supported by MOCHA in particular. In Section 3, we discuss why a release of new articulatory data is desirable, paying particular attention to two shortcomings of the Wisconsin XRMB and MOCHA data sets. Finally, in Section 4 we outline the mngu0 corpus that will in time all be made publicly available, as well providing details of the subset released at this stage.

Research area	Example work
speech recognition	Richardson et al. (2003) [6]
speech synthesis	Toda et al. (2008) [7] Kello & Plaut (2004) [8]
inversion mapping	Richmond (2007) [9]
signal processing	Shiga (2005) [10]
voice transformation	Toth (2008) [11]
speech segmentation	Akdemir & Eiloglu (2008) [12]
tongue modelling	Qin & Carreira-Perpiñán (2010) [13]
speech representation	Gutkin (2005) [14]
speech production	Jackson & Singampalli (2008) [15]

Table 1: Examples of work facilitated by MOCHA EMA data.

2. Why public articulatory corpora?

To appreciate the benefit brought by the free availability of articulatory corpora one only has to review the range of research that has made use of MOCHA. This corpus consists of 2 speakers reading 460 British TIMIT utterances: msak0 (male) and fsew0 (female), which has been particularly widely-used.

Table 1 presents a summary of research that has used MOCHA. This summary captures only a fraction of this work (e.g. Google Scholar returns over a hundred papers, with more than a dozen on the inversion mapping alone), but it does at least give an idea of the scope of the work and with some examples. For ASR, [6] introduced the "Hidden-Articulator Markov model" (HAMM) and used MOCHA to evaluate the articulatory movements predicted by their model (which resulted in decreased word error rates, especially in noise). For speech synthesis, researchers have attempted to model the mapping from articulatory to acoustic synthesis parameters, e.g. using Gaussian mixture models (GMM) [7] or a neural net [8]. Going in the opposite direction, i.e. the acoustic-to-articulatory inversion mapping, [9] used MOCHA to train various nonlinear regression mappings. In a technique they term "multiframe analysis", [10] used MOCHA articulatory data to cluster corresponding acoustic frames and perform spectral envelope estimation for multiple frames simultaneously, with the aim of improving accuracy. [11] attempted, amongst other things, to use articulatory data to improve the performance of voice-conversion between the two speakers of MOCHA. For automatic segmentation of speech (e.g. phone labelling), [12] reported a reduction of 18% in average absolute boundary error with respect to manual labels when features derived from the articulatory data were included.

In [13], MOCHA data for 3 tongue points were used to drive a statistical model to predict entire tongue contours for whole utterances. [14] experimented with the Evolving Transformation System formalism to induce a formal articulatory representation of speech from MOCHA articulatory data. Finally, [15] used statistical techniques to analyse MOCHA data with the aim of identifying the varying roles of articulators during speech as either 'critical', 'dependent' or 'redundant'.

Though articulatory data is clearly useful, it is unfortunately not easy to acquire, requiring specialist equipment and expertise. Although it may be possible to use the articulography resources of another site[1], it is obviously beneficial for researchers who need articulatory data to avoid the trouble and expense of recording their own. Not only does this minimise effort and facilitate novel research, but using common data sets in theory allows comparison between different methods. This is discussed further in the next section.

3. The case for a fresh corpus

Publicly available articulatory corpora such as MOCHA clearly already provide an invaluable resource, so why is there need for another corpus of articulatory data? The most straightforward answer is that for many empirical investigations and machine-learning modelling techniques it is simply better to have more data. That aside, there are further reasons why another corpus of freely available data is desirable. Here, we consider two example drawbacks of the currently available corpora: difficulty in comparing results, and data inconsistency.

3.1. Comparison of results

Although multiple researchers working on the same problem may use the same corpus in their experiments, it does not follow that it is straightforward, or even possible, to directly compare results, and hence to directly compare methods. This is because researchers typically perform their own preprocessing. Even when the processing steps are reported, for myriad reasons they may not always be repeated exactly. This confounds comparison of methods and impedes progress. This situation arises where data has been released solely in a "raw" and "static" form. Releasing preprocessed versions together with the raw data might help somewhat, although it is impossible to foresee future developments. Ideally, data should be distributed via an infrastructure designed to keep pace with developments in its use. For example, if an effective feature extraction method were devised, it would be ideal to add that parameterisation for distribution with the raw data. Unfortunately, it is likely to be too late for such an effort in the case of MOCHA and the Wisconsin XRMB corpus. It is unlikely researchers who have conducted experiments using these data would go back and rerun them using a standardised parameterisation of the data. A release of fresh data, however, offers the opportunity to put such infrastructure in place at the outset. This is discussed further in Section 4.3.

3.2. Evidence for inconsistency

Recording human articulator movements is not straightforward. Although the currently available articulatory data sets have provided a valuable resource, they cannot be assumed to be perfect. As an example, in this section we briefly consider the presence of inconsistency we have previously identified [17] in MOCHA.

One clear source of inconsistency in EMA data is introduced when a coil becomes detached during recording. This is problematic not only because it disrupts the recording session, but because it is unfortunately not possible to re-attach the coil in exactly the same place. For example, when recording `fsew0` the velum coil was re-attached at file recording index 125, while the middle tongue coil was re-attached at index 284. These inconvenient breaks must be taken into account when using this data, or they will undoubtedly influence results.

[1]The Edinburgh Speech Production Facility [16] offers a studio with two Carstens AG500 machines for hire for example

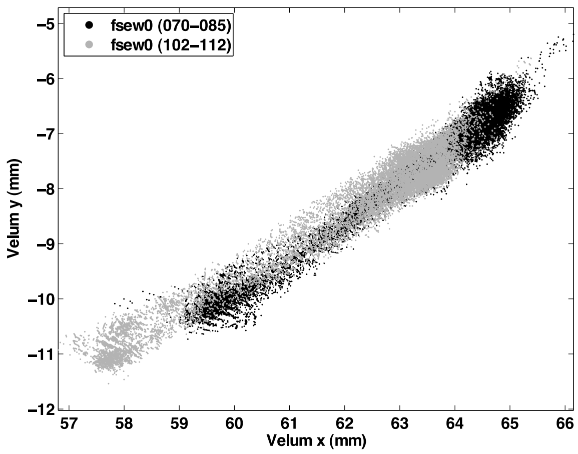


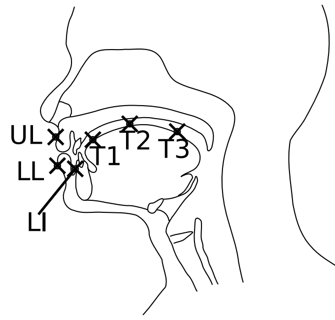
Figure 1: A scatter plot of velum position in multiple files of MOCHA speaker `fsew0`. Two sets of contiguously recorded files are shown: one group (black) shows files $070 - 085$, the other (grey) shows files $102 - 112$. (reproduced from [17]).

Alas, there is also evidence to suggest an additional, less clear-cut source of inconsistency is present in both MOCHA speakers. Fig. 1 illustrates this with a scatter plot of velum position throughout multiple utterances. These utterances comprise two groups of contiguously-recorded files. The first group of points, shown in black, are taken from files $070 - 085$. The second group, shown in grey, contains points from files $102 - 112$. In both these groups, we observe that the velum moves in a regular way, in a slight arc with well defined limits. However, this pattern of velum movement appears to be *shifted* in one group relative to the other. The cause of this variation is not known (though potential causes are discussed further in [17]). However, it is clear that, while this inconsistency is easiest to spot in the constrained movement of the velum, it is very likely to have affected the other EMA channels too. This too is bound to influence results in experiments using MOCHA data.

4. A new articulatory corpus - `mngu0`

In the course of our recent research into various ways of incorporating articulatory information into speech technology, we have compiled a set of multiple forms of articulatory data acquired from a single native British English speaker. So far, this data set includes EMA data for read speech with accompanying audio and video of the lower face, Volumetric MRI scans for all sounds in the speaker's inventory and 3D scans of dental impressions of the speaker's upper and lower jaw. In time, this will all be released and hosted together in one place, and will hopefully provide a useful resource of multi-modal articulatory data for the research community. This paper announces the first part of this data that will be released at this initial stage, consisting of a large part of the EMA recordings.

The EMA part of the `mngu0` corpus was recorded at Ludwig-Maximilians-Universität München, using a Cartsens



label	location	label	location
UL	Upper lip	T1	Tongue tip
LL	Lower lip	T2	Tongue body
LI	Lower incisor	T3	Tongue dorsum

Figure 2: Sensor coil locations for the "day 1" EMA subset. All 6 coils were placed in the midsagittal plane. Additional coils (not shown) were used for head-movement correction.

AG500 electromagnetic articulograph. This data set consists of over 2,000 utterances recorded over two consecutive days. On the first day, over 1,300 utterances were recorded with EMA sensors attached as indicated in Fig. 2. On the second day, this configuration was changed slightly by placing a coil on the velum and using only two coils on the tongue instead of three. Around 800 utterances were recorded with this configuration. The data to be released first are all the utterances recorded on day 1. EMA data from day 2, as with the rest of the `mngu0` data, will be released at a later date.

4.1. Day 1 EMA subset - raw data

The Day 1 subset contains 1,354 utterances, giving a total of approximately 67 mins of speech data, discounting initial and final silences. This is a large amount of speech data from a single speaker relative to the Wisconsin XRMB corpus or MOCHA (where the amount of speech for each speaker is only 15–20 minutes for example). So, a primary advantage of the `mngu0` day 1 EMA data is simply how much of it there is.

This data was originally collected with unit selection speech synthesis in mind, and therefore we tried to maximise variety in several respects. The sentences were selected from newspaper text using an algorithm designed for building `Multisyn` [18] voices for `Festival`. This algorithm aimed to maximise coverage of context-specific diphones in as few sentences as possible. Sentence lengths range from 1 to 48 words, and comprise a variety of types, such as questions, statements, exclamations etc. Overall, this set contains approx. 1,715 and 12,322 unique diphone and triphone types respectively (with 1,562 and 6,737 appearing at least twice respectively).

Raw EMA data: In addition to providing a large and phonetically rich source of data alone, the `mngu0` EMA subset offers other significant advantages. Having used the AG500 confers several benefits. Unlike the preceding 2D AG200 or AG100 systems, the AG500 tracks sensor coils in 3D space with two angles of rotation, resulting in 5 measurements per sensor coil. This provides richer information about each articulatory point. It also means a speaker's head is free to move, which increases comfort, which in turn allows longer recording times. Finally, it also avoids the problem found with the AG200, for example, of inaccuracy being introduced when a sensor moves off the midline plane of the transmitter coils (the AG200 makes the assumption that coils lie only in this plane)[2].

It should be noted that none of the the sensor coils became detached during recording, which is a significant advantage as it immediately increases the chances for a consistent data set. This contrasts with the Wisconsin XRMB data, in which pellets may "disappear" in some sections due to mis-tracking, and MOCHA, in which sensor coils became detached as mentioned in Section 3.2. From preliminary experiments `mngu0` indeed seems relatively consistent; in [17] our best inversion mapping system achieved average root mean square error of 0.99mm using `mngu0`, compared to 1.54mm using MOCHA `fsew0`.

The EMA data is distributed in Edinburgh Speech Tools format files, containing 84 channels (5 coordinates plus 2 reliability indicators for each of the total 12 coils used in the AG500) plus sample times. Files include a header noting the identity of each channel, amongst other information. The `ch_track` utility in the Edinburgh Speech Tools distribution can manipulate these files, and modules are distributed on the web forum which make it possible to read the data into Matlab and Python scripts.

Raw audio data: Audio was recorded concurrently using 2 microphones: a Sennheiser MKH50 hypercardioid, at 50-60 cm from lips; and a PHON-OR noise-cancelling optical microphone. Each of these has pros and cons. The standard microphone is susceptible to the electromagnetic transmitter coils of the AG500, which show up as dark bands across the spectrogram (lowest at approx. 7.5kHz), though these may be attenuated using a notch filter. The optical microphone is not affected by this, but its frequency response is not as wide which is especially noticeable at low frequencies. Both sets of acoustic waveforms are distributed in RIFF wavefiles. Finally, since this data was recorded with speech synthesis in mind, special care was taken to ensure good audio quality. A professional actor and voice-talent was employed, and unlike MOCHA EPG was not used, so the impact on pronunciation is negligible.

4.2. Day 1 EMA subset - processed data

In addition to distributing the raw EMA and audio data, we are also distributing the processed versions of this data used in previous experiments [17, 3, 19]. This should allow others to conduct their own experiments using *exactly* the same data. This makes it more convenient to use the data, but more importantly also should allow more direct comparison between results, which will hopefully prove illuminating. Specific details of the processing carried out are included in the distribution of this portion of `mngu0`, but are also briefly summarised here.

Processed EMA data: Although the raw EMA data provides 5 measurements for each of the 6 sensor coils indicated in Fig. 2, we have so far only used x- and y-coordinates in the midsagittal plane. Hence, our processed version comprises 12 channels of EMA, sampled at 200Hz, with initial and final silences omitted. These have been z-score normalised by subtracting the respective global mean from each channel, and then dividing by 4 times each channel's respective global standard deviation.

Processed audio data: The corresponding audio data has been converted to frequency-warped line spectral frequencies (LSF) of order 40 plus a gain value. We have used the waveforms recorded using the standard microphone, lowpass filtered and downsampled to 16kHz (so omitting transmitter coil traces). STRAIGHT was used to estimate the spectral envelope for these, with a frame shift of 5msec to match the EMA sample rate. The LSF feature vectors have also been z-score normalised in the same way as the EMA data.

[2]Though AG500 coil tracking becomes a non-linear optimisation problem itself!

Subsets: Three subsets have been used in previous inversion mapping work using mngu0 data: a validation and test set each containing 63 utterances, and a training set containing the rest. The lists of which utterances have been used in which set are also available for download, making it straightforward to recreate reported experiment conditions.

Labelling: Phone labelling is provided in addition to the processed audio and EMA data. This was produced automatically using the combined forced alignment tools of Multisyn [18] and the Combilex lexicon [20], with labels given in the Combilex phone set. We also provide Festival *Utterance* structures for each utterance, generated by the front-end text analysis modules of the Festival text-to-speech synthesis system, but using the forced-alignment phone sequence.

4.3. Distribution and the mngu0 web forum

All data in the mngu0 corpus will be made available via the dedicated web site: http://www.mngu0.org. Public access to this site will be enabled shortly before the Interspeech 2011 conference begins. Data will be freely available for research use, although prospective users will be required to accept the licence agreement and provide contact details in order to register an account prior to downloading. The main benefits of requiring this are two-fold. First, it will make it easier to keep track of who is using mngu0. This will make it possible to contact users and notify them of updates or new releases. It could also provide useful statistics to support future grant applications in which funds are requested to acquire further articulatory data.

Second, we aim to encourage all those who download mngu0 data to look upon the web site as a hub for research activity related to mngu0. For example, as mentioned in Section 4.2, in addition to distributing the raw EMA data, we intend to post various processed versions derived from it that we have used in our experiments. It would be ideal if other mngu0 users were subsequently willing to post their own processed versions of the data too where that may prove useful (e.g. any hand-labelling, or special features extracted from the raw data), so that other researchers would in turn be able to conduct experiments using *exactly* the same data. As another example, the web forum holds a repository of papers that use mngu0 data, and we would encourage all users to upload their own papers to this collection. We hope this too will provide additional benefit to the research community.

Finally, to support the aims of the mngu0 web forum, we would ask all users not to pass on any data directly to other prospective users if asked, but instead encourage them to register themselves as a user at www.mngu0.org and to obtain the data directly from there.

5. Summary

We have highlighted the usefulness of current freely available articulatory data, and hereby announce our own new contribution to this, which is to be distributed via a dedicated web forum. We hope this will be useful, and will in time become well-used.

6. Acknowledgements

Thanks to Susanne Waltl for help with the EMA recordings. K. Richmond is supported by EPSRC grant EP/I027696/1. EMA recording was supported by EPSRC grant GR/R94688/01.

7. References

[1] E. Mcdermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE Transactions on Information and Systems*, vol. 89, no. 3, pp. 1006–1014, 2006.

[2] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[3] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1171–1185, 2009.

[4] J. R. Westbury, "X-ray microbeam speech production database user's handbook," University of Wisconsin, Madison, Tech. Rep. Version 1.0, 1994.

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

[6] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Speech Communication*, vol. 41, no. 2-3, pp. 511 – 529, 2003.

[7] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[8] C. Kello and D. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *The Journal of the Acoustical Society of America*, vol. 116, p. 2354, 2004.

[9] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007*, ser. Lecture Notes in Computer Science, M. Chetouani, A. Hussain, B. Gas, M. Milgram, and J.-L. Zarader, Eds., vol. 4885. Springer-Verlag, Berlin Heidelberg, 2007, pp. 263–272.

[10] Y. Shiga, "Precise estimation of vocal tract and voice source characteristics," Ph.D. dissertation, The Centre for Speech Technology Research, Edinburgh University, Edinburgh, UK, 2005.

[11] A. Toth, "Using articulatory position data to improve voice transformation," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, USA, 2008.

[12] E. Akdemir and T. Çiloglu, "The use of articulator motion information in automatic speech segmentation," *Speech Communication*, vol. 50, no. 7, pp. 594 – 604, 2008.

[13] C. Qin and M. Á. Carreira-Perpiñán, "Reconstructing the full tongue contour from EMA/X-ray microbeam," in *Proc. ICASSP*, 2010, pp. 4190–4193.

[14] A. Gutkin, "Towards Formal Structural Representation of Spoken Language: An Evolving Transformation System (ETS) Approach," Ph.D. dissertation, School of Informatics, University of Edinburgh, UK, 2005.

[15] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[16] A. Turk, J. M. Scobbie, C. Geng, C. Macmartin, E. G. Bard, B. Campbell, B. Diab, C. Dickie, E. Dubourg, B. Hardcastle, P. Hoole, E. Kainada, S. King, R. Lickley, S. Nakai, M. Pouplier, S. Renals, K. Richmond, S. Schaeffler, R. Wiegand, K. White, and A. Wrench, "An Edinburgh speech production facility," http://www.lel.ed.ac.uk/projects/ema/home.html, 2010.

[17] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proc. Interspeech*, 2009, pp. 2835–2838.

[18] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[19] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834 – 846, 2010.

[20] S. Fitt, K. Richmond, and R. Clark, "The Combilex lexicon," http://www.cstr.ed.ac.uk/research/projects/combilex.