A Multitask Learning Perspective on Acoustic-Articulatory Inversion

Korin Richmond[1]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom

korin@cstr.ed.ac.uk

Abstract

This paper proposes the idea that by viewing an inversion mapping MLP from a Multitask Learning perspective, we may be able to relax two constraints which are inherent in using electromagnetic articulography as a source of articulatory information for speech technology purposes. As a first step to evaluating this idea, we perform an inversion mapping experiment in an attempt to ascertain whether the hidden layer of a "multi-task" MLP can act beneficially as a hidden representation that is shared between inversion mapping subtasks for multiple articulatory targets. Our results in the case of the tongue dorsum x-coordinate indicate this is indeed the case and show good promise. Results for the tongue dorsum y-coordinate however are not so clear-cut, and will require further investigation.

Index Terms: Multitask learning, acoustic-articulatory inversion, multilayer perceptron

1. Introduction

Mainstream speech technology focuses closely on the acoustic speech signal. The acoustic domain is where the speech signal exists in transmission between humans, and moreover, we can easily measure an acoustic representation of speech. Nevertheless, it is recognised that the acoustic speech signal is ultimately the result of events in a speaker's articulatory system, and there has long been interest in ways to exploit the underlying articulatory information for speech technology. An articulatory representation of speech has certain attractive properties which may be exploited in modelling. Articulators move relatively slowly and smoothly, and their movements are continuous. The mouth cannot "jump" from one configuration to a completely different one. Using speech production knowledge could improve speech processing methods by providing useful constraints. Previously suggested applications include, for example, automatic speech recognition [1, 2], low bit-rate speech coding [3], speech analysis and synthesis [4], and animating talking heads.

To use articulatory information in speech processing applications, the articulatory representation itself must somehow be obtained. In recent years, electromagnetic articulography (EMA) has become one of the most promising sources for articulatory data. The advantages of EMA include its relatively low cost, negligible impact on articulation and low danger to the subject! In addition, because EMA records the movements of fixed points on the articulators rather than "images" of the articulatory structures, the need for postprocessing to extract articulatory movements from the raw data is greatly reduced. For these four main reasons, we can now record reasonably large amounts of articulatory data using EMA (e.g. MOCHA [5]), which in turn means the use of machine learning algorithms in common use today in speech technology is viable.

Despite the advantages of using EMA to collect articulatory data for speech technology, one or two drawbacks or uncertainties remain. First, there is the possibility that the seven or eight sensor coils that are typically recorded in EMA corpora are not sufficient to provide a full description of the vocal tract, and so important information is missing. There are two reasons for this limitation. The first is that the EMA machines available only provide a limited number of coils. The second, more fundamental problem, is that it is impractical to fix more than a small number of coils to a speaker's articulators at any one time. Attaching coils is a time consuming procedure, so attaching a great number becomes prohibitive. The presence of a great number of receiver coils increases the likelihood of impairing the speaker's articulation. Finally, the coils may interfere with each other; coils in closer proximity to each other than 8mm may affect signal measurement accuracy. In addition, physical collisions between two coils may cause either or both of them to dislodge. For example, care must be exercised when placing coils on the velum and the back of the tongue; if these coils collide during normal articulation, it is likely that one or both of them will become detached.

Even in absence of collisions, EMA coils may spontaneously become detached during the recording of a corpus of significant size. This is a second significant problem, as placing the coils accurately is somewhat difficult, and so the position and orientation at which a coil is re-attached is unfortunately only approximately the same. This introduces the problem of inconsistency of the data throughout the corpus.

We can attempt to normalise the differing positions of coils which have been re-attached, but success may be limited, depending on the extent and nature of the difference. Furthermore, normalisation does not help with the problem of the limited number of coils. In contrast, this paper addresses both these shortcomings jointly by considering the inversion mapping from a *Multitask Learning* perspective.

Caruana [6] provides a compelling review and demonstration of the potential benefits of Multitask Learning. In Multitask Learning, a single empirical learning model (e.g. a multilayer perceptron (MLP)) is trained to perform multiple *related* tasks at the same time. The tasks are made to share the same hidden representation (e.g. the MLP hidden layer), which introduces the possibility that the training signals for each task act as an inductive bias for the other tasks. In other words, what the MLP learns for one task can help the MLP for the other related tasks. This principle is termed inductive transfer.

Intuitively, estimating the movements of multiple articulatory points from the same acoustic signal when performing the inversion mapping is a set of interrelated tasks[1]. The degree of this interrelation depends on which articulators one considers. We might expect the relationship between the inversion mappings for two points on a rigid structure such as the lower jaw to be closely related. Whereas, the relationship between the velum and the jaw is less obvious. Nevertheless, they are all

[1] Although inversion mapping studies have frequently modelled the inversion mapping for each articulator channel separately (e.g.[7]).

in some sense related tasks. Therefore, in an MLP designed to perform the inversion mapping for multiple articulators, it is natural to ask to what extent we can view the hidden layer as a shared internal representation of vocal tract configuration.

If the hidden layer were to function as a shared underlying representation to a significant degree, then we would be able to tackle the two problems with collecting EMA data described above. First, to address the problem of the limited number of sensor coils, it might prove viable to move the EMA coils part way through recording a large corpus and "overlay" the different coil configurations in an inversion mapping MLP. This would yield a fuller final representation of the vocal tract. Similarly, if a coil were to become spontaneously detached, it may subsequently turn out to be possible to exploit a shared hidden representation in order to make full use of the EMA data recorded from the coil in that position. The mechanism for sharing the hidden representation in this way differs slightly from the straightforward method employed in typical applications of MLP-based Multitask Learning, although is still quite elementary. This method is described in Section 2.1.

Addressing the above two issues is the ultimate goal of this line of research. At this preliminary stage, the purpose of this paper is primarily to detail the proposed idea itself, and second to investigate whether there is indeed significant evidence of sharing in the hidden layer of an MLP trained to perform inversion for multiple articulators.

2. Inversion mapping experiment

We will first describe the principle by which multitask learning might be implemented in an MLP, in accordance with the goal of addressing the issues presented in Section 1. We will then describe the dataset used in this investigation. Finally, we will provide the details of the experiment conducted.

2.1. Implementation of a shared representation

In standard applications of Multitask Learning to MLP training, units which are fully connected to the hidden layer may simply be added to the output layer. Target output values in the training set may then accordingly be augmented with the targets for additional related tasks. In this way, all output units for the separate subtasks naturally share the hidden representation for the whole training set. In this approach, the target data is defined for all tasks and for the whole training set.

To address the two problems described in Section 1, however, we cannot assume that each subtask's target data is defined for all training patterns. By way of explanation, consider an example where we have a total training set of 1,000 utterances, and where we may have recorded 500 utterances with coils attached in some configuration "A" and 500 utterance with coils in configuration "B". Ultimately, we want a single MLP which can estimate the positions of coils in configurations A and B at the same time and in response to the same acoustic input pattern. We aim to train a single MLP on these two subtasks, while promoting the development of shared hidden representation.

This condition requires a different approach to implementing Multitask Learning. Fortunately, we can achieve the desired effect of sharing the hidden representation in our case by a slight modification to the backpropagation training in the multitask MLP. Specifically, we do the following:

1. Construct an MLP with output units corresponding to all target articulator positions available in the training set (in our example case, those for configurations A and B).
2. Concatenate all subsections of the training data with coils in configurations A and B.
3. Where a target value for a coil is not known for a particular input acoustic vector, it is set to a value "*undefined*".
4. During MLP error calculation and backpropagation, the error for an *undefined* target is set to 0.0.

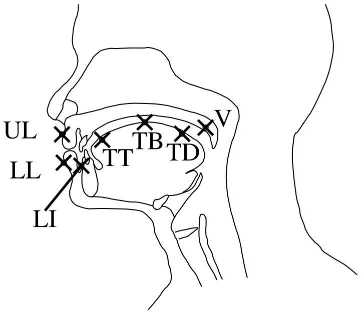

Figure 1: *EMA receiver coil locations for MOCHA speaker* `fsew0`. *Table 1 gives a key to the coil positions.*

label	articulator	label	articulator
UL	Upper lip	TT	Tongue tip
LL	Lower lip	TB	Tongue body
LI	Lower incisor	TD	Tongue dorsum
V	Velum		

Table 1: *Key for coil locations for MOCHA speaker* `fsew0`. *Coil locations may be suffixed with "_x" and "_y" to refer to the x- and y-coordinate in the midsagittal plane respectively.*

For a training pattern recorded with coils in configuration A, the target outputs for coils in configuration B will be *undefined*, and the corresponding error will be 0.0. Therefore, during error backpropagation only the signals from the coils from configuration A will affect the fully connected shared hidden representation. This situation will be reversed for a training pattern which was recorded with coils in configuration B. In this way, both subtasks corresponding to coil configurations A and B will affect and, hopefully, benefit from the shared hidden layer.

2.2. MOCHA articulatory data

The multichannel articulatory (MOCHA) dataset [5] gives the acoustic waveform recorded concurrently with electromagnetic articulograph (2D EMA) data. The sensors in Figure 1 provide x- and y-coordinates in the midsagittal plane at 500Hz sample rate. Speakers were recorded reading a set of 460 phonetically-balanced British-TIMIT sentences. The data of a *single* speaker, female `fsew0`, was used for the experiments here.

2.2.1. Data processing

The acoustic signal was converted to frames of 20 melscale filterbank coefficients using a Hamming window of 20ms with a shift of 10ms. These were z-score normalised and scaled to the range [0.0,1.0]. The EMA trajectories were downsampled to match the 10ms shift rate, then z-score normalised and scaled to the range [0.1,0.9] using the normalisation method described in [8]. Frames of silence at the beginning and end of files were discarded, using the labelling provided with MOCHA.

368 utterances were used for the full training set (this was reduced by varying amounts for the various networks trained;

Num utts	Num frames	Num utts	Num frames
23	4200	184	44159
46	9806	207	48993
69	15351	230	54781
92	20778	253	60768
115	26503	276	67028
138	33105	299	73862
161	38830	322	79367

Table 2: 14 subsets of contiguously recorded utterances from MOCHA speaker `fsew0` used as training data for the tongue dorsum coil.

see Section 2.3), and the validation and test sets contained 46 utterances each. All `fsew0` MOCHA files with filename numbers ending in "2" were selected for the validation set, while all files ending in "6" were selected for the test set. This is the same scheme as used elsewhere [9, 7], and means the three datasets are drawn uniformly from the database taken in recording order.

A context window of 20 consecutive acoustic frames was used as network input, which increased the order of the acoustic vector paired with each articulatory vector to 400.

2.3. Experimental details

The aim of this initial experiment is to verify whether any benefit is observed from sharing the hidden layer in an MLP as described in Section 2.1. To achieve this, we have taken the MOCHA dataset in Section 2.2, and simulated various amounts of missing, or *undefined*, data for one of the coils: the tongue dorsum. Table 2, lists the 14 training sets we have created to simulate this. These training sets differ in the number of utterances for which the tongue dorsum x- and y-coordinates are defined. For example, in the first training set, the tongue dorsum coil is defined for only the first 23 utterances (4200 frames), while in the next training set, it is defined for 46 utterances (9806 frames) and so on. We have purposely divided up the training set into blocks of *consecutively* recorded utterances, as this matches what we are likely to encounter when coils become detached and are re-attached during EMA recording. It also matches the condition where we might choose to move coils part way through recording in order to get a richer representation of inferred vocal tract shape via an inversion mapping.

Next, using each of these 14 training sets, we trained the following network configurations:

1. An MLP with 14 outputs trained on data for all coils. We will refer to this as the **multitask (MT)** MLP.
2. An MLP with 2 outputs trained on only the tongue dorsum data which is *defined* within the various training sets in Table 2. We will call this the **single task (ST)** MLP.

Thus, a total of 28 MLPs were trained, all with a single hidden layer of 80 units. All output units used a linear activation function. The scaled conjugate gradients non-linear optimisation algorithm was run for a maximum of 4000 epochs, and the separate validation set was used to identify the point at which an optimum appeared to have been reached. The unseen test set was then used to compare the performance of the inversion mapping for the tongue dorsum coil by the MT and ST MLPs.

3. Results

Figure 2 shows the correlation of network output with target trajectories for the tongue dorsum x-coordinate on the test set for all 28 networks trained. Figure 3 gives RMS error, expressed in millimetres, between the network output and the test set target trajectories for the tongue dorsum x-coordinate. These plots indicate the multitask MLP does indeed derive benefit from the shared hidden representation, with increased performance over the single task MLP when the training data for the tongue dorsum coil is reduced by varying amounts.

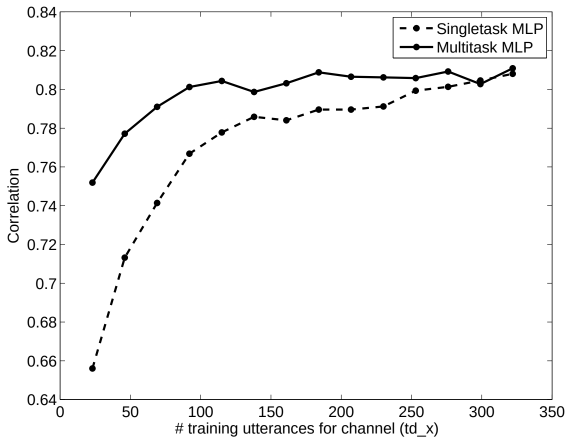

Figure 2: *Correlation for the tongue dorsum x-coordinate as a function of number of training utterances. Results for the multitask (MT) and single task (ST) MLPs are shown.*

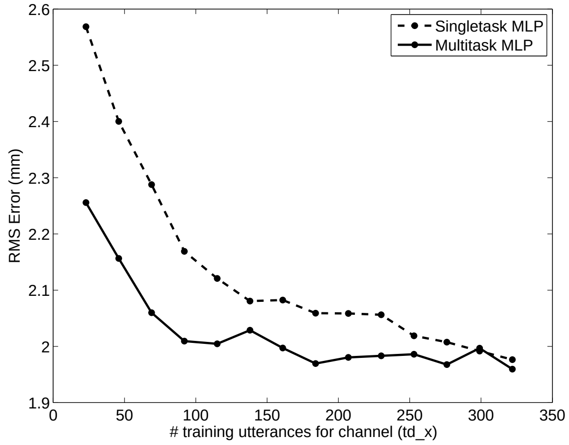

Figure 3: *RMSE(mm) for the tongue dorsum x-coordinate as a function of number of training utterances. Results for the multitask (MT) and single task (ST) MLPs are shown.*

This is a very encouraging result and suggests, for example, that by exploiting the shared hidden representation, the multitask MLP is able in this case to perform the inversion mapping equally well when trained with only around 184 utterances. This potentially implies that during recording of a future EMA dataset, we could move the coil from this location after collecting around this number of utterances (or seconds) and collect additional data from a different articulator location. Alternatively, it implies that if this coil had become detached during recording after around only 92 utterances, we would still be able to use this data to learn the inversion mapping for this articulatory position to a level only fractionally worse than with the coil firmly attached at this location for the whole corpus.

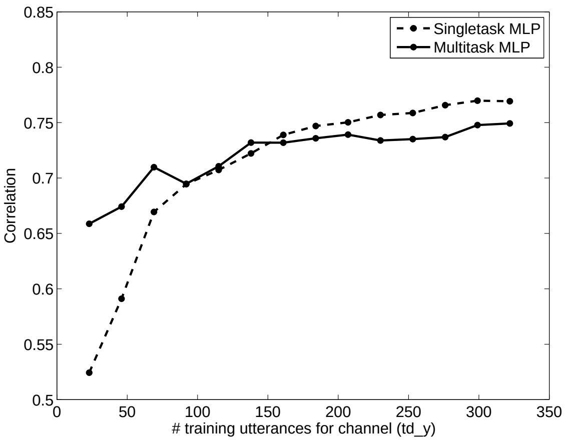

Figure 4: *Correlation for the tongue dorsum x-coordinate as a function of number of training utterances.*

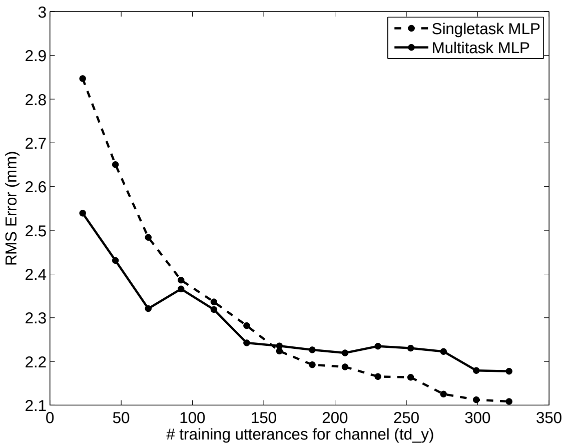

Figure 5: *RMSE(mm) for the tongue dorsum y-coordinate as a function of number of training utterances.*

Figures 4 and 5 present the equivalent results for the tongue dorsum y-coordinate. With a reduced training set of up to around 69 utterances, we observe the same benefit of the shared hidden representation in the multitask MLP; correlation is increased and RMS error is reduced compared with the single task MLP trained on the same amount of tongue dorsum data alone. With a training set of data for this channel larger than around 161 utterances, however, we observe that the multitask MLP performance is consistently *worse* than that of the single task MLP. This partly contradicts the previous results, and is not what we would expect according to the theory of MLP learning.

4. Discussion

One possible explanation for the results observed is that a shared hidden representation really *is* detrimental in the case of the inversion mapping for the y-coordinate of the tongue dorsum. Before accepting this conclusion though, further investigation is needed to rule out other causes. No straightforward explanation has yet been isolated, but potential causes range from the possibility that the multitask MLP does not contain sufficient hidden units to the possibility that the data for other coils is somehow corrupt or inconsistent (for example, during recording, the velum coil was reattached at utterance 125, and the tongue body coil was reattached at utterance 284).

In addition to explaining the results observed here, we will in future investigate how the sharing effect varies among articulator locations. We also have the opportunity to use a larger EMA dataset of 1,263 utterances to establish the behaviour of the effect with larger overall amounts of training data. For example, we would like to evaluate the effect of overall database size on the minimum data required for each articulatory location; with a total dataset of 1,263 utterances, it might be the case we need less data from any individual coil to give the same performance overall.

5. Conclusions

This paper has put forward the idea that by viewing an inversion mapping MLP from a Multitask Learning perspective, we might address at least two difficulties of using EMA to provide articulatory data for speech technology purposes.

As a first step, we have investigated whether there is any evidence that the hidden layer of an inversion mapping MLP can act beneficially as a shared representation. Our results in the case of the inversion mapping for the tongue dorsum x-coordinate are promising. They demonstrate there is indeed evidence to suggest this is the case. However, our results for the tongue dorsum y-coordinate indicate there may be complications which prohibit a straightforward approach to exploiting a shared hidden representation. More work will be required before clearer conclusions may be drawn.

6. References

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

[2] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. ICSLP 2000*, Beijing, China, 2000.

[3] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker Inc, 1992, ch. 8, pp. 231–268.

[4] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proc. 5th ISCA Workshop on Speech Synthesis*, 2004.

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

[6] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: citeseer.ist.psu.edu/caruana97multitask.html

[7] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. Interspeech*, Pittsburgh, USA, September 2006.

[8] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, The Centre for Speech Technology Research, Edinburgh University, 2002.

[9] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.