Using deep neural networks to estimate tongue movements from speech face motion

Christian Kroos [1], *Rikke L. Bundgaard-Nielsen* [2,3], *Catherine T. Best* [3,4], *Mark D. Plumbley* [1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]La Trobe University, Australia
[3]MARCS Institute, Western Sydney University, Australia
[4]Haskins Laboratories, USA

c.kroos@surrey.ac.uk, rikkelou@gmail.com, c.best@uws.edu.au, m.plumbley@surrey.ac.uk

Abstract

This study concludes a tripartite investigation into the indirect visibility of the moving tongue in human speech as reflected in co-occurring changes of the facial surface. We were in particular interested in how the shared information is distributed over the range of contributing frequencies. In the current study we examine the degree to which tongue movements during speech can be reliably estimated from face motion using artificial neural networks. We simultaneously acquired data for both movement types; tongue movements were measured with Electromagnetic Articulography (EMA), face motion with a passive marker-based motion capture system. A multiresolution analysis using wavelets provided the desired decomposition into frequency subbands. In the two earlier studies of the project we established linear and non-linear relations between lingual and facial speech motions, as predicted and compatible with previous research in auditory-visual speech. The results of the current study using a Deep Neural Network (DNN) for prediction show that a substantive amount of variance can be recovered (between 13.9 and 33.2% dependent on the speaker and tongue sensor location). Importantly, however, the recovered variance values and the root mean squared error values of the Euclidean distances between the measured and the predicted tongue trajectories are in the range of the linear estimations of our earlier study.

Index Terms: Face motion, tongue movements, deep neural networks, speech articulation, multiresolution analysis, wavelets, electromagnetic articulography

1. Introduction

This study is the third and concluding part of an investigation into how much unobservable articulatory information can be recovered from observable face motion during speech and how much the different frequency bands contribute. Our starting point was the proposition that Articulatory Phonology [1] does not need to be modified or extended to be applied to visible speech processing in humans [2]. If the hypothesis holds, comparing visual speech with both auditory-only and auditory-visual speech should allow gaining insights into the mechanisms underpinning how articulatory gestures are perceived and processed by the perceivers and might reveal how the inverse problem of acoustic-to-articulatory conversion is solved in human speech perception. If the same articulatory gestures are perceived – whether auditorily or visually – and if this occurs in the same manner, examining the perception differences resulting from the modality-specific differences in the amount of available information should shed some light on the minimal requirements for identifying articulatory gestures. Since the primary modality of speech is acoustic, it is reasonable to assume that human speech is most strongly adapted to this modality and information transfer is close to optimal here. All acoustics-based accounts of speech perception have this as their central tenet.

On the other hand, the increase in intelligibility by watching the speaker's face during acoustic speech production in noisy conditions is well documented (starting on the experimental side with Sumby & Pollack (1954) [3]) as is visual speech reading of silent speech (e.g, [4]) and interference when incongruent auditory and visual speech is presented (e.g., in the McGurk-Effect [5]), sometimes interpreted as indicating auditory-visual integration. There is, however, a marked difference in the degree and detail of gestural information that can be recovered by human perceivers via the visual versus the auditory modality [6]. Some speech articulators are only partially or not at all directly observable. This applies for instance to the velum, the inner workings of the larynx and most crucially the tongue. Some or all of the motion of these articulators, however, might be still reflected in changes of the facial surface, either as direct mechanical consequences of muscles activity and the connectivity of the entire speech apparatus via ligaments and tissue connections, or through an indirect functional link, that is, a gesture being made more visible in the production process to aid visual speech perception. A prerequisite for experimentation into the effects of the factors discussed above upon perception is to determine how much information about the unobservable articulators is available on the facial surface.

A couple of studies examined the association between movements of the speech articulators and face motion (e.g., [7, 8]), though their number is far less than the number of studies investigating associations between tongue movements and acoustics (central to articulatory phonetics) and associations between face motion and acoustics (e.g., for multi-modal speech recognition). The relatively small number of studies into comotion of the face and vocal tract articulators in speech might at least partially be caused by the difficulty of capturing face motion simultaneously with the measurements of the 'hidden' articulators. Yehia et al. (1998) [9] found high correlations between their articulatory data acquired with EMA and their face motion data acquired with Optotrak (Northern Digital Inc.). Between 72% to 91% of the variance was accounted for via cross-modality linear prediction. However, those findings were limited by several factors. The researchers were not able to record articulatory data and face motion data simultaneously and used dynamic time warping to align the data from separate experiment runs. Their articulatory data were two-dimensional and limited to the mid-sagittal plane and only a small data set was available for each of the two speakers (American English and Japanese).

High face-articulator correlations were also found in a study investigating American English with a slightly larger stimulus set, as reported in [10]. For the two male and two female speakers of American English average correlation values in the range of 0.74 to 0.83 were registered. All data were recorded simultaneously; however, the articulatory data were still constrained to the mid-sagittal plane. Two-dimensional articulatory data cannot capture any lateral variations of tongue shape, e.g., the difference in tongue shape between /t/ and /l/. As a consequence the relationship between face motion and the movement of the articulators might be overestimated dependent on the phoneme under investigation. In addition the cross-modality estimation was based on CV syllables only.

Substantially lower association strengths were detected by Bailly & Badin (2002) [11]. In the study several linear modelling steps were used to combine tongue traces from cineradiographic data with face motion data. The data sets were not recorded simultaneously but linked via estimated vocal tract target configurations. Moreover, they did not consist of natural speech, but were mostly 'hyperarticulated sustained articulations' [11]. The recovered variance of four parameters capturing tongue motion ranged from 37% to 71%.

The experiments reported in [12] and [13] also resulted in relatively low correlation values and high root mean square errors (RMSE). For methodological reasons, however, they are not comparable with the current study and the other studies cited above. On the one hand, the face motion data were limited to a single 2D-sensor each at the upper and lower lip (mid-sagittal location). As the lips are active articulators and involved in coarticulation, it is highly unlikely that much information about tongue location can be recovered from only lip hight and protrusion information. More fine-grained differences in lip shape might indirectly reveal tongue position differences, but they were not available in these studies. On the other hand, a jaw sensor was added to the face motion data. The mandible itself is not directly visible and its position can only be inferred – be it by human observers or camera-based machine vision system. The sensor, however, was attached to the gums of the lower teeth and registered the movements of the mandible directly. It therefore adds ground truth data to the predicand variable set, which facilitates estimating tongue tip and dorsum position considerably whenever the tongue moves in unison with the mandible or is passively moved by it.

As we pointed out previously in [14] the prior high correlation results seem to be at odds with human performance in silent speech reading. For instance, Auer (2009) [4] tested 20 participants with severe-to-profound hearing loss who relied primarily on vision for speech communication in a word recognition task. Recognition rates for low lexical frequency words dropped below 40% even for words that were assumed to have no visually similar competitor, and below 20% and 10% for words from visually medium dense and very dense neighbourhoods, respectively. This was the case despite that these tasks included phonemes that constitute highly visible visemes such as the bilabial stop consonants /b/ and /p/ or the labiodental fricative /f/.

Our own findings using Partial Least Squares (PLS) as the linear estimation method of choice [14] resulted in relatively low correlation values, comparable to the performance of human speech readers. Extending the research to non-linear relationships in [15], using Mutual Information, a wealth of shared information was asserted. Mutual Information (MI) does not allow the prediction of one modality from the other and returns only relative values if the overall information contents of the signals cannot be determined. Comparing computed MI values to to the ones obtained from a single jaw EMA sensor and 5 lower chin Vicon markers made a rough assessment possible, characterising most of the tongue-face association as very weakly linked. However, as suggested in [15] non-linear estimations method might still profit from the shared information that goes beyond what is captured by linear methods (the linear relationships are of course also registered by MI).

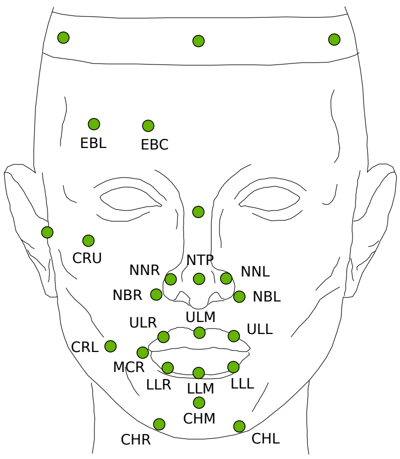

Figure 1: *OPT marker target locations with the marker code displayed next to the indicator dots.*

In the current study we therefore employed artificial neural networks on the non-linear regression problem. Because of their superior performance in many tasks, we choose Deep Neural Networks (DNNs [16, 17], see [18] for a comprehensive overview). The history of Deep Neural Networks extends back in time for at least two decades, but only in recent years has their success in a variety of classification, detection and synthesis tasks made them the most popular techniques in machine learning. Shallow artificial neural networks with only one hidden layer have been shown to be able to theoretically approximate any continuous function [19]. However, given complex functions, the number of nodes in the hidden layer might become exceedingly large. It has also been shown that Deep Neural Networks, that is, networks with many hidden layers, reduce the number of required nodes substantially [20]. Initial procedural problems of how to train DNNs have now been overcome, the necessary computational resources are available and very large data sets are also by now widespread.

Given the considerations outlined above, we hypothesised that a DNN approach would yield a modest, but noticeable improvement over the previously-applied techniques in terms of the root mean squared Euclidean distance error between measured and predicted tongue trajectories and in terms of the recovered variance. The overall values for the latter, however, were expected to remain in the typical range for human speech reading performance. As in our previous studies, we were particularly interested in how articulatory information is distributed over the contributing frequency subbands. Frequency-dependent findings have the potential to aid computer vision systems for visual speech reading by providing clues at where to focus. This can be used, for instance, in improving acoustic automatic speech recognition in very noisy environments such as many industrial manufacturing sites. Frequency-dependent findings also offer opportunities to better understand human speech reading as they might enable to determine on which level (e.g., syllable or phoneme) human speech reading can potentially work most effectively.

2. Method

2.1. Motion data acquisition

Three female speakers of American English (aged 22-28 years) were recorded reading a slightly modified version of the traditional children's story 'Chicken Little'. The story was divided into seven passages comprising about 6-9 sentences each and was read in a very lively manner by the participants.

Flesh point measurements of the tongue (3 sensors) and jaw (1 sensor) were obtained using three-dimensional Electromagnetic Articulography (Carstens AG500) – abbreviated as EMA hereafter. Head motions were tracked with three sensors: one each at the left and right mastoid process of the ears and one at the maxilla. The tongue sensors were attached mid-sagitally with the orientation of the sensor axis aligned with the sagittal axis. This sensor arrangement makes the current study comparable to previous studies reviewed in the Introduction, but with the benefit/difficulty of adding the third spatial dimension (lateral movements) to the estimation targets.

Face motion was captured using the optical Vicon (Vicon Motion Systems Ltd) motion capture system (abbreviated as OPT hereafter). Eight MX40 cameras were placed at two different height levels in front of the EMA cube (in the direction the speakers faced) and two each at each side (right and left sides of the speaker's face). We used 21 half-spherical 3-mm markers attached at key locations on the facial surface of the participant, primarily on the right side, since the wires of the EMA sensors were brought out of the left corner of the mouth and were attached with micropore tape to the participant's left cheek (all speakers included in this study were right-handed).

Three face markers were attached on the chin, 7 around the vermilion border of the lips, 4 on the nose (wings, tip and bridge), 5 on the cheek, and 2 at the right eye brow. In order to track head motion also with the OPT system, three 9-mm spherical markers were sewn to a head band the speaker wore. Figure 1 shows a schematic with the target locations of the markers.

Since the EMA cube is not fixed relative to the OPT system and can be moved e.g., by the speaker involuntarily touching the cube, we tracked potential EMA cube movements with three 14-mm markers fixed to the front of the cube with plastic screws. However, only very little movement was detected and computationally compensated. Both systems operated with a sample rate of 200 Hz. To enable temporal synchronisation in the post-processing the trigger signal from the EMA's AG500 Sybox was recorded with the Vicon analogue signal recording unit MX Control.

2.2. Data post-processing

Face motion and tongue movement data were temporally aligned using the synchronisation signal mentioned in the previous section. The two data types were then spatially aligned by determining the global offsets between the two coordinate systems: Immediately after the recording of the speaker four EMA sensors were wrapped with reflective tape, turning them into simultaneous OPT markers, and positioned at random locations in the measurement field of both systems used in the speech recording. Several trials provided the coordinates of these four points in both the EMA and the OPT coordinate reference frame (method details in [21]). The offsets were then computed using conventional pose estimation via the General Procrustes Method [22].

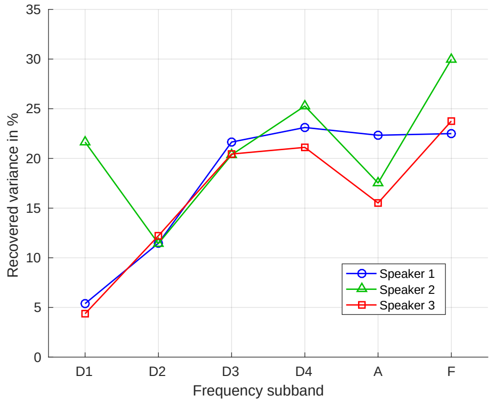

Figure 2: *Recovered variances for each speaker in percent averaged over all three sensors except for speaker 2, for whom only tongue tip and tongue back sensor data were available. The multiresolution subbands (from higher to lower frequencies: D1, D2, D3, D4, A) and the full signal (F) are shown along the x-axis.*

The speakers head was computationally stabilised using the methods proposed in [13], removing the impact of head movements from the motion measurements. Residual and smoothness analyses indicated that the OPT tracking using the head band markers yielded the most reliable results and it was employed. The face motion tracking data were cleaned manually frame by frame: spurious 'ghost' markers due to mistracking were removed and short-lived passages of tracking loss were interpolated using Vicon's Woltring quintic spline filter [23]. Finally, the motion signals were downsampled to 50 Hz after a appropriate low-pass filtering. The tongue dorsum sensor of speaker 2 exhibited repeated and prolonged periods of tracking failure and had to be discarded in its entirety.

2.3. Frequency decomposition

We applied a multiresolution analysis [24] using the discrete wavelet transformation (DWT) [25]. Wavelet transformations represent functions in terms of base functions at different scales and positions [26]:

$$f(t) = \sum_{s=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} c_{s,l}\, 2^{-\frac{s}{2}}\, \psi_{s,l}\left(2^{-s}t - l\right) \tag{1}$$

where $c_{s,l}$ are the wavelet coefficients and $\psi_{s,l}(t)$ the wavelet function.

Wavelet transformations use 'small waves', wavelets, that have their energy concentrated around a point in time, i.e. the energy of the wavelet function is finite. This is in contrast with the Fourier transformation which expands signals (or functions) in terms of sines and cosines (or equivalently in terms of complex exponentials) that are infinite. As a consequence Wavelet transforms are localised in time and frequency, the degree of localisation being dependent on the frequency range: at lower frequencies they trade off a relatively poor localisation in time for a relatively good frequency resolution, but the trend is gradually reversed when moving towards higher frequencies. The

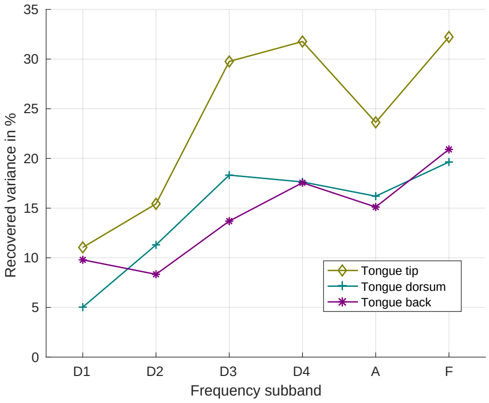

Figure 3: *Recovered variances for each sensor in percent averaged over all speakers. Note that due to problems with the tongue dorsum sensor of speaker 2 only the two other speakers contribute to the average for this sensor. The multiresolution subbands (from higher to lower frequencies: D1, D2, D3, D4, A) and the full signal (F) are shown along the x-axis.*

discrete wavelet transformation can be implemented with a set of cascading digital halfband filters [27]. Starting with the raw signal, the input signal is decomposed at each level of the multiresolution analysis into a high frequency and a low frequency part, using a pair of highpass and lowpass filters, which are orthogonal to each other. The lowpass data are then used as the input signal for the subsequent level.

We used filters corresponding to a biorthogonal scheme with cubic spline wavelets (see [28] for algorithm details) as implemented in the Uvi_Wave wavelet toolbox for Matlab (The Mathworks, Inc.). A multiresolution analysis entails as its last step the application of a corresponding set of synthesis filters to transform the wavelet coefficients back into the original signal domain. The result is a set of signals at different scales band-limited to one octave (given the usual dyadic DWT). The first one contains frequencies between the Nyquist frequency f_n and its half $f_n/2$, the second one between $f_n/2$ and $f_n/4$, and so on. In this study, we considered four wavelet levels sufficient, thus we obtained four 'details' and an 'approximation' as follows: **D1**: $12.5 - 25$ Hz; **D2**: $6.25 - 12.5$ Hz; **D3**: $3.13 - 6.25$ Hz; **D4**: $1.56 - 3.13$ Hz; **A**: $0 - 1.56$ Hz.

2.4. DNN-based estimation

In the current study, the Cartesian coordinates of the face markers constitute the predictor variable and the Cartesian coordinates of the tongue sensors the predicand variable.

There were 59 sentences for speaker 1, 64 for speaker 2 and 65 for speaker 3. The differences are due to the exclusion of single sentences because of mistracking by one or both of the capture systems. This resulted in 9660, 11196 and 10690 samples, respectively, adding up to 31546 available samples in total.

The data were split sentence-wise into a training set (80% of the data) and a final evaluation set (20% of the data). The training set was in turn arranged for a four-fold validation: Each fold consisted of a new random (without replacement) split of the files into training data (75%) and test data (25%). The final evaluation set was only used to test the trained network on previously unseen data after all hyper-parameters were determined and all network parameters learned on the full training set.

We used the Matlab Neural Networks toolbox (The Mathworks, Inc.) for training and evaluating the deep neural networks. The toolbox offers a class of networks intended for non-linear regression, a fully-connected feed-forward architecture with hyperbolic tangent sigmoid activation functions on the hidden layers and a linear activation on the output layer.

As the loss function, the mean squared error between the target data and the output estimation of the network was chosen. Using the four-fold cross-validation we evaluated several different network topologies, ranging from a single hidden layer with 513 nodes to deep narrow architectures (e.g., number of nodes in of the hidden layers: 60-60-60-60-60-60-60) to relatively broad deep topologies (e.g., 513-228-171-171). Based on the four-fold evaluation results we fixed the layout eventually as a 117-117-117-117-117 network.

The Matlab toolbox appears to have implemented neither dropout nor mini-batches for the 'fitnet' class of networks and relies mostly on early stopping to prevent overfitting. In this study, we considered stronger regularisation highly necessary given the high number of parameters in the network and the fact that we noticed substantially lower training errors than test errors initially. As dropout could have only be added in the form of a work-around, we implemented mini-batch training. Three mini-batches containing a randomly selected 33.3% of the training data were employed. To learn the weights, the scaled conjugate gradient algorithm was used in backpropagation.

The following additional parameters were used: Maximum number of training epochs: 8000; error goal: 0; maximum number of validation tests without improvement before stopping: 400; minimum gradient for proceeding: 10^{-7}; learning rate at start: 0.001; learning rate decrease (automatic adaptation): 0.1; learning rate increase (automatic adaptation): 10; maximum learning rate: 10^{9}.

To be able to compare results across frequency subbands and with the linear prediction from [14], the mean was subtracted and the data normalised to a standard deviation of 1.

3. Results

The recovered variance (*R*-squared) of the tongue motion data by predicting them from face motion data is shown in Figures 2 and 3. The approximation and all details from the wavelet transformation are compared with each other and the full signal. Figure 2 depicts the averages across sensors for each individual speaker, while Figure displays 3 the averages across speakers for each sensor.

As can be seen in Figure 2, tongue movements are best predicted when the full frequency range is available in both the predictor and the predicand, with the exception of speaker 1, where a plateau was formed starting with D3 ($3.13 - 6.25$ Hz). There are pronounced speaker differences affecting primarily D1 ($12.5 - 25$ Hz) and A ($0 - 1.56$ Hz).

In line with expectations, the coordinates of the tongue tip sensor are estimated more accurately than the coordinates of the tongue dorsum and tongue back sensors. The difference is larger in the lower frequencies and the full signal. The top values of recovered variance remain just below the one third mark. Note that the tongue tip moves frequently in conjunction with the jaw, which is tracked relatively well with chin markers on the face surface, the only disruption coming from independent movements of the skin at the chin, e.g., for lip rounding.

The overall root mean squared error of the Euclidean dis-

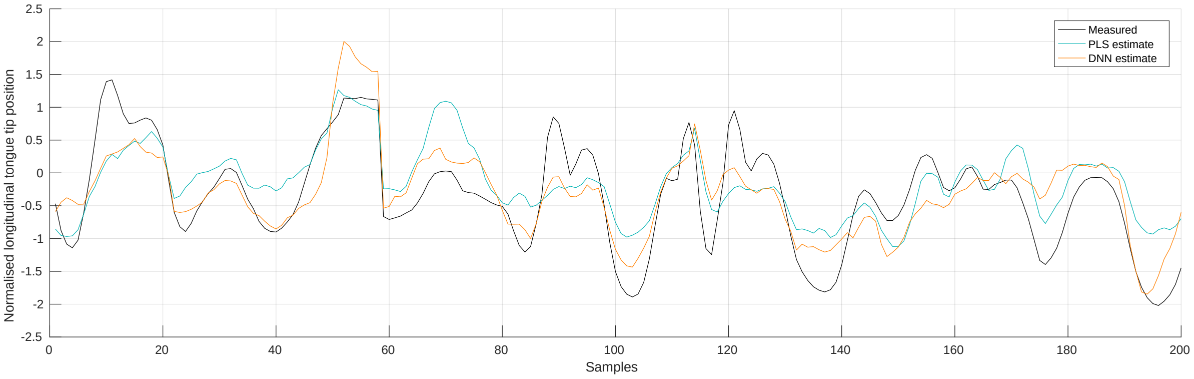

Figure 4: *200-sample point comparison of the PLS-based estimation from [14] and the DNN-based estimation of the current study regarding tongue tip location along the longitudinal axis for speaker 2.*

tances between the measured and the estimated tongue trajectories over the three spatial dimensions and three speakers is displayed in Table 1. The values are given separately for each tongue sensor.

4. Discussion

The results concerning the amount of recovered variance using a DNN-based estimation differ in some details from those of the best linear models reported in [14], but are overall in the same range. This was confirmed by the RMSE of the Euclidean distances between the measured and estimated signals. For speaker 1 the linear estimations result in an even lower error, but because of the small magnitude of the differences, chance variation cannot be ruled out. Figure 4 gives an example: The PLS-based estimation of the longitudinal location (i.e. vertical location if the head is upright) of the tongue tip of speaker 2 is compared with the DNN-based estimation and the original signal. Note that all of them are normalised to have unit standard deviation for the reasons described above. The start and end points were randomly selected.

Two possible explanations for the surprisingly low performance of the DNN-based estimation can be put forward:

1. The relationship between face motion and tongue movements is essentially linear. This would be likely if there is a mostly mechanical connection between the two movement types. If, for instance, muscles that move the tongue also deform the facial surface, their impact on the face is likely to move partial areas of the face surface in the same manner as the tongue albeit not necessarily in the same direction. The magnitude of this secondary movement, however, can be assumed to be linearly related to the magnitude of the tongue movement. It appears, though, that the amount of shared information found in our previous study [15] contradicts this explanation, since the shared information was much more spread over the facial surface and covered relations that contributed little or nothing in the linear estimation.

2. The DNN was not able to learn any associations beyond linear relationships. The shared information exceeding linear relations detected in [15] might have been too small and largely insignificant and thus not been able to help with the prediction given the relatively low number of samples (for a DNN-based estimation) in the current study. Learning might have been additionally hampered by the fact that the available corpus – consisting only of a number of natural speech sentences – is not very well balanced, requiring an even higher number of samples for the DNN to be able to generalise.

Currently both explanations appear equally likely, and more research is needed to favour one of them over the other. In fact, we cannot even rule out that both are partially appropriate in that the linear portion is rather robust (since based on mechanical coupling), but the non-linear portion is heavily affected by the interference of other face motions (since being the result of a malleable functional coupling, that is, speech gestures made visible for the purpose of improving verbal communication in noisy environments). Our participants read the story in a very lively way, almost as if reading them to a small child, a positive side effect of having a continuous story among the stimulus material. It provided naturalness for the speech production in an otherwise highly constrained lab situation and helped the participants to ignore, for instance, the EMA sensor wires leading out of their mouth and being taped to the left cheek. The affective component might have interfered with the more subtle non-linear relational components of the facial and articulator motions and made them less accessible and more complex for humans and machines alike.

According to both, PLS-based and DNN-based estimation, most of the face motion that conveys information about tongue motion can be found in the wavelet subbands D3 and D4, spanning together the range from 1.56 to 6.25 Hz. This covers events that show a single oscillatory change within a time window in the range from 160 to 641 ms. This corresponds approximately to durations of phonemes on the faster side and syllables in the mid-range.

Overall the results are in line with human speech reading capabilities. In terms of their direct or indirect visibility, speech gestures realised by the tongue take up a position in the middle. They are not as directly visible as lip gestures are, but they are also not as difficult to detect as velum activations due to the tongue being situated on the floor of the mouth and passively moved by the visible motions of the jaw as well as being moved by muscles some of which connect to the facial muscle system and tissue. Given the results from this study, it can be expected that slightly less than a third of tongue gestures are recognised correctly from facial motion data. To reach these levels, it might or might not be a condition that the speech reader acquires some familiarity with the speech production of the specific speaker: The PLS and DDN models were trained speaker-dependently.

Table 1: *Normalised RMS prediction error (Euclidean distance) for tongue tip (TT), tongue dorsum (TD), and tongue back (TB) shown for the five multiresolution subbands (from higher to lower frequencies: D1, D2, D3, D4, A) and the entire signal (F).*

	Subband	TT	TD	TB
Speaker 1	D1	1.722	1.713	1.770
	D2	1.590	1.650	1.692
	D3	1.390	1.569	1.677
	D4	1.369	1.592	1.618
	A	1.441	1.577	1.676
	F	**1.452**	**1.576**	**1.647**
Speaker 2	D1	1.695	-	1.550
	D2	1.658	-	1.650
	D3	1.502	-	1.614
	D4	1.524	-	1.540
	A	1.644	-	1.654
	F	**1.451**	**-**	**1.527**
Speaker 3	D1	1.778	2.087	2.242
	D2	1.598	1.628	1.690
	D3	1.479	1.576	1.589
	D4	1.469	1.591	1.613
	A	1.562	1.665	1.627
	F	**1.446**	**1.580**	**1.547**

We have not yet conducted cross-speaker evaluations but intend to do so in future research.

5. Conclusion

We estimated tongue trajectories of natural speech from simultaneously recorded face motion measurements using a Deep Neural Network. The results were comparable to the best linear model employed in an earlier study using the same data set [14]. We did not detect a substantial improvement suggested by the presence of additional non-linear relationships, determined using Mutual Information in another previous study employing the same data set [15]. Signal components in the frequency range between 1.56 to 6.25 Hz emerged as the most salient contributors to the observed face-tongue movement relationship.

6. Acknowledgements

The research leading to this submission was funded by UK EPRSC grant EP/N014111/1. The data were collected at the MARCS Institute with support by the Western Sydney University (Australia).

7. References

[1] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[2] C. Kroos, "Auditory-visual speech analysis: In search of a theory," in *Proceedings of the 16th International Congress of Phonetics Sciences*, Saarbrcken, Germany, 2007, pp. 279–284.

[3] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212–215, 1954.

[4] E. T. Auer Jr., "Spoken word recognition by eye," *Scandinavian Journal of Psychology*, vol. 50, no. 5, pp. 419–425, 2009.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[6] L. D. Rosenblum, "Speech perception as a multimodal phenomenon," *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, 2008.

[7] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of facial and intraoral articulation from simultaneous measurements," in *15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, Spain, 2003.

[8] H. Kjellström, O. Engwall, and O. Bälter, "Reconstructing tongue movements from audio and video," in *Interspeech 2006*, Pittsburgh, PA, USA, 2006.

[9] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23–44, 1998.

[10] J. Jiang, A. Alwan, L. E. Bernstein, P. Keating, and E. Auer, "On the correlation between facial movements, tongue movements and speech acoustics," in *International Conference on Spoken Language Processing*, vol. 1, Bejing, China, 2000, pp. 42–45.

[11] G. Bailly and P. Badin, "Seeing tongue movements from outside," in *International Conference on Speech and Language Processing (ICSLP)*, Boulder, CO, USA, 2002, pp. 1913–1916.

[12] A. B. Youssef, P. Badin, and G. Bailly, "Can tongue be recovered from face? The answer of data-driven statistical models." in *Interspeech 2010*, 2010, pp. 2002–2005.

[13] A. Toutios and S. Ouni, "Predicting tongue positions from acoustics and facial features," in *Interspeech 2011*, 2011, pp. 2661–2664.

[14] C. Kroos, R. L. Bundgaard-Nielsen, and C. T. Best, "Now you see it, now you don't - frequency distribution of articulatory information reflected in speech face motion," in *SST 2012*, Sydney, Australia, 2012, pp. 117–120.

[15] ——, "Exploring nonlinear relationships between speech face motion and tongue movements using Mutual Information," in *International Speech Production Seminar 2014*, Köln, Germany, 2014, pp. 237–240.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[19] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[20] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[21] C. Kroos, "Evaluation of the measurement precision in three-dimensional Electromagnetic Articulography (Carstens AG500)," *Journal of Phonetics*, vol. 40, no. 3, pp. 453–465, 2012.

[22] J. Gower and G. Dijksterhuis, *Procrustes Problems*. New York: Oxford University Press, 2004.

[23] H. J. Woltring, "A Fortran package for generalized, cross-validatory spline smoothing and differentiation," *Advances in Engineering Software*, vol. 8, pp. 104–113, 1986.

[24] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, 1989.

[25] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, Pennsylvania: SIAM, 1992.

[26] B. Jawerth and W. Sweldens, "An overview of wavelet based multiresolution analyses," *SIAM Review*, vol. 36, no. 3, pp. 377–412, 1993.

[27] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, Massachusetts: Wellesley-Cambridge-Press, 1997.

[28] G. S. Sánchez, N. G. Prelic, and S. J. G. Galán, *Uvi_Wave. Wavelet Toolbox for use with Matlab*, 2nd ed., Departamento de Tecnoloxías das Comunicacións. Universidade de Vigo, Vigo, July 1996.