Now you see it, now you don't - frequency distribution of articulatory information reflected in speech face motion

Christian Kroos[1], *Rikke L. Bundgaard-Nielsen*[1], *Catherine T. Best*[1,2]

[1]MARCS Institute, University of Western Sydney, Australia
[2]Haskins Laboratories, USA

c.kroos@uws.edu.au, rikkelou@gmail.com, c.best@uws.edu.au

Abstract

Although the increase in intelligibility of spoken language when watching the speaker's face is well documented, the characteristics and the distribution of phonetic information over the frequency range of visual speech remain largely unknown. For this study face motion and tongue movements were measured simultaneously for three speakers of American English. The motion signals were subjected to a multiresolution analysis using spline wavelets and partial least squares regression was applied to estimate tongue movements from face motion. The amount of recovered variance was found to be small (below 30%) compared to previous studies and more pronounced in the lower frequencies.

Index Terms: auditory-visual speech, face motion, electromagnetic articulography, partial least squares, wavelets

1. Background

We have previously proposed [1] to analyse visual speech production within the framework of Articulatory Phonology [2]. As stated in [1], Articulatory Phonology (AP) does not need special extension or modification to cover speech face motion since it does not assume acoustic segments to be the basic units of speech, rather articulatory gestures. These gestures are implicitly multimodal by definition, due to their grounding in the physical reality of vocal tract configuration changes. Thus, in theory, AP can be directly applied to visual speech. There is, however, a marked difference in the degree and detail of gestural information that can be recovered by human perceivers via the visual or the auditory modality [3]. For instance, the constriction degree of velar stop consonants and fricatives might be difficult (or even impossible) to identify visually and the activity of the velum is hidden in general.

Independent of theoretical perspective, it remains unclear how much phonetic/articulatory detail is preserved in visible face motion. Obtaining simultaneous measurements of face motion and the movements of the non-visible speech articulators 'hidden' in the vocal tract has been forbiddingly difficult until now due to technical limitations.

Nevertheless there have been a few studies on the association between vocal tract articulations and speech face motion (e.g., [4, 5]). In particular, [6] found high correlations between the two modalities, which accounted for 72% to 91% of the variance in the data sets. Because of the aforementioned technical difficulties, the researchers were not able to record articulatory data and face motion data simultaneously and were forced to use 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). [7] confirmed the results investigating two male and two female speakers of American English with a slightly larger stimulus set. They registered average correlation values in the range of 0.74 to 0.83 for the four speakers. All data were simultaneously recorded, however, their 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 their cross-modality estimation was based on CV syllables only. [8] used several linear modelling steps to combine tongue traces from cineradiographic data with face motion data recorded with a large number of markers. Both types of data were not recorded simultaneously but linked via estimated vocal tract target configurations. '[M]ost analysed targets were hyperarticulated sustained articulations' [8]. The recovered variance of four parameters capturing tongue motion ranged from 37% to 71%.

With the exception of some of the findings of the last study the results seem to be at odds with human performance in silent speech reading. For instance, [9] 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. The difference between technical estimation and speech reading abilities in humans might be simply due to the fact that not all physically available information is perceived, e.g., some changes in the facial surface might be too small to be detected. The greatly improved performance of exceptionally good speech readers [10] points in this direction. Also, in the above studies only two-dimensional tongue movements were estimated, not, for instance, additionally velum activity or lateral tongue shape differences. The necessity to recover these articulatory movements from face motion certainly contributes to the higher difficulty encountered by human speech readers in a word recognition task. However, the studies cited above apply concurrent sample-based methods that do not take into account dynamic information and non-linear relationships that could be exploited by a human speech reader. Thus, under the assumption of modality-independent articulatory gestures as building blocks of phonetic information, it stands to reason that speech reading should be more accurate even when no top-down linguistic processes are able to support the recognition process and/or the environmental and speaker conditions are not ideal.

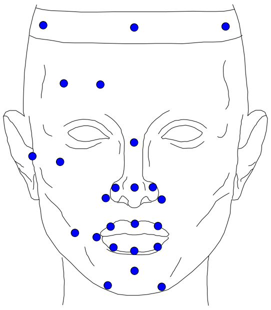

Figure 1: *Face marker locations.*

The aim of the current study is to examine the amount of shared variance of observable speech face motion and mostly unobservable tongue movements using simultaneously acquired three-dimensional measurements of natural continuous speech. In particular, the distribution across different frequencies is investigated. Given the relatively high speaking rate and natural speaking style of our speakers (see below) and in line with theoretical reasoning above, we hypothesise that substantially less information about tongue movements can be recovered from face motion than found in previous studies, explaining the relatively poor performance of human speech readers.

2. Method

2.1. Data acquisition

We recorded six female speakers, three speakers of American English and three speakers of Australian English. Three-dimensional Electromagnetic Articulography (*Carstens AG500*) - EMA hereafter - was used to obtain flesh point measurements of the tongue (3 sensors) and jaw (1 sensor) and to track head motion (sensors at the mastoid processes of the left and right ears, and at the maxilla). The orientation of the tongue sensors was chosen such that the sensor axis was aligned with the mid-saggital axis; thus allowing recovery of the tongue orientation in the saggital plane at the sensor location from the sensor orientation angle measurements.

Face motion was measured using the optical *Vicon* (Oxford Metrics) motion capture system (OPT hereafter) with 8 MX40 cameras, with 4 cameras 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). Twenty-one half-spherical 3-mm markers were attached to the facial surface of the participant, primarily on the right side, since the wires of the EMA sensors led out from the mouth to the left and were attached with micropore tape to the participant's left cheek. Seven face markers were placed around the vermilion border of the lips, 3 on the chin, 5 on the cheek, 4 on the nose (wings, tip and bridge), and 2 at the right eye brow. Finally, three 9-mm spherical markers were sewn to a head band the speaker wore, allowing us to also track head motion with the Vicon system. Figure 1 shows a schematic with the target locations of the markers.

To be able to compensate for potential EMA cube movements relative to the static OPT coordinate system, the EMA cube was tracked with three 14-mm markers fixed to the front of the cube with plastic screws.

Both systems sampled measurements with a rate of 200 Hz. To enable temporal synchronisation in the post-processing the trigger signal from the *AG500 Sybox* was recorded with the *Vicon* analog signal recording unit *MX Control*.

2.2. Post-processing

As a first step, face motion, articulatory and acoustic signals were temporally aligned using the synchonisation signal mentioned above. As a second step both types of measurement data were spatially aligned by determining the global offsets between the two coordinate systems. Special trials in which four EMA sensors were wrapped with reflective tape - thus becoming OPT markers - provided the coordinates of four points in both the EMA and the OPT coordinate reference frame (details in [11]). The offsets were then computed using conventional pose estimation via the General Procrustes Method [12]. As a third step, the impact of head movements was removed from the motion measurements, i.e., the speaker's head was computationally stabilised using the methods proposed in [13]. Both, residual and smoothness analyses indicated that the OPT tracking using the head band markers yielded the most reliable results and it was employed to estimate head movements sample-wise using again pose estimation. All data were then rotated and translated to compensate for the estimated head movements. Finally, the motion signals were downsampled to 50 Hz.

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 [14]. Due to the time-consuming nature of the data correction we will present here early results limited to the three American English speakers and one of the stories, the traditional children's story 'Chicken Little'. It exhibits repetitions of specific phrases due to the unfolding of the story but at the same time is phonetically rich. The story was recorded divided into seven passages comprising about 6 - 9 sentences each and was read in a very lively manner by the participants. Altogether 31,738 concurrent samples were obtained. Because of repeated episodes of measurement failure the tongue dorsum sensor signal of speaker 2 had to be excluded entirely. All speakers included in this study were right-handed.

2.3. Analysis

In order to obtain the signal at the desired temporal frequency subbands a multiresolution analysis [15] was applied using a discrete wavelet transformation (DWT) [16]. Wavelet transformations represent functions (or signals) in terms of 'small' base functions at different scales and positions [17]:

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

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

While the Fourier transformation expands signals (or functions) in terms of sines and cosines (or equivalently in terms of complex exponentials) that are infinite, wavelet transformations use 'small waves', wavelets, that have their energy concentrated around a point in time. They are thus localised in time dependent on the frequency range: at higher frequencies the trade off a relatively poor localisation in the frequency domain for a relatively good resolution in the time domain, but the trend is gradually reversed when moving towards lower frequencies. This property makes them very suitable for the analysis of biological motion. The discrete wavelet transformation can be implemented with a set of cascading digital halfband filters [18]. 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.

In this study filters were used corresponding to a biorthogonal scheme with cubic spline wavelets (see [19] for details about the algorithm) as implemented in the Uvi_Wave wavelet toolbox for Matlab (The Mathworks, Inc.). A multiresolution analysis entails as the 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 band-limited to one octave (given the usual dyadic DWT) at different scales: the first one containing 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. We deemed four wavelet levels sufficient for our purposes, 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.

We used concurrent sample-based Partial Least Squares [20] to quantify the relations between the subsets of the visually observable (OPT system) and the non-observable (EMA system) measurement points. Partial Least Squares (PLS, also known as Projection on Latent Structures) extracts components that maximise the squared covariance between two data sets and can be applied to data sets with high multi-collinearity, such as ours. A PLS analysis returns a set of regression coefficients that allow estimation of each data set from the respective other. We estimated the EMA data from the OPT data and computed the root mean squared (RMS) error between the original and the reconstructed signal as measure for goodness of fit and the amount of recovered variance. These measures were computed for the full signal and each of the subbands. The data were randomly divided into a training set (80%) and a test set (20%) and the estimation process was repeated 5 times with different divisions. All results presented are averaged over the 5 repetitions.

3. Results

The percentage of recovered variance of the tongue movement data from the test set against the number of latent components included in the PLS model is shown for the full signal in Figure 2 for each speaker separately. As can be seen with an increasing number of latent variables the amount of recovered variance increases, too, until all the shared variance is exhausted. The maximum percentages remain, however, below 30%.

Figure 3 depicts the results with respect to the five subbands averaged over the three speakers. Clearly, the three low frequency subbands contain more shared variance and within this group D4 (1.56 - 3.13 Hz) contains slightly more shared variance than the other two.

The RMS error is tabulated in Table 1 split according to speaker and EMA sensor. They appear small ranging from 1.31 to 1.71 mm. Note, however, that these are the RMS error values calculated with normalised signals and signal estimates with a standard deviation of 1 mm. Otherwise the values on different scales cannot be compared since the higher frequency subbands usually exhibit lower global variance values. Relative to the standard deviation of the signals the RMS errors are, of course, substantial. Tongue tip movements are slightly better recovered than tongue dorsum and tongue back which is most likely due to the stronger influence of jaw position on the tongue tip.

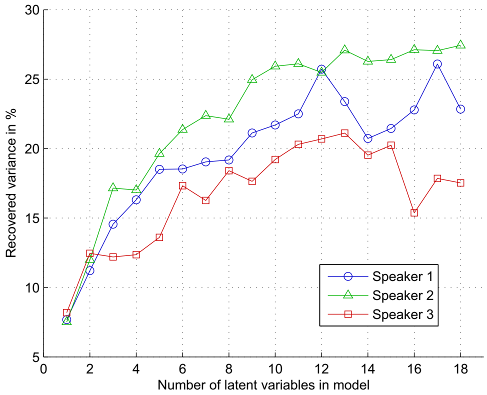

Figure 2: *Recovered variance in percent dependent on the number of latent variables included in the PLS model. Entire signal.*

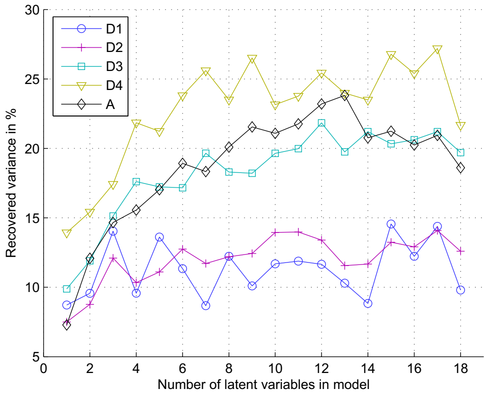

Figure 3: *Recovered variance in percent dependent on the number of latent variables included in the PLS model. Multiresolution subbands averaged over speakers.*

4. Discussion

The current study failed to find the high shared variance or correlation values between face motion and tongue movements of previous studies. As hypothesised the obtained values are in general low. A couple of likely causes can be put forward to explain the difference:

- In contrast to [6] and [7] our EMA data were three-dimensional. However, we observed only small amplitudes in the lateral dimension.
- Facial marker positions might have not been ideal for the purpose of the study. The locations were chosen according to the requirements of the face synthesis software *FaceRobot* (Softimage). *FaceRobot* models soft tissue deformation of the facial surface and needs the used locations to drive animations with motion capture data. It is unclear how much attention the software developers paid to speech face motion (as compared to emotional face expressions). However, the location differences to

Table 1: *RMS prediction error using the optimal model (speaker 1 and 2: 17 LVs, speaker 3: 13 LVs) for tongue tip (TT), tongue dorsum (TD), and tongue back (TB) shown for the five mutilresolution subbands (from higher to lower frequencies: D1, D2, D3, D4, A) and the entire signal (F).*

	Subband	TT	TD	TB
Speaker 1	D1	1.673	1.679	1.712
	D2	1.594	1.644	1.675
	D3	1.383	1.561	1.628
	D4	1.306	1.499	1.592
	A	1.404	1.541	1.590
	F	**1.377**	**1.513**	**1.569**
Speaker 2	D1	1.545	-	1.569
	D2	1.567	-	1.626
	D3	1.481	-	1.575
	D4	1.441	-	1.477
	A	1.424	-	1.548
	F	**1.427**	**-**	**1.528**
Speaker 3	D1	1.636	1.662	1.692
	D2	1.583	1.656	1.690
	D3	1.472	1.595	1.626
	D4	1.484	1.583	1.593
	A	1.484	1.562	1.563
	F	**1.475**	**1.569**	**1.568**

[6] and [7] appear not pronounced enough to cause such strikingly different results. We also had to place the majority of markers on the right hand side of the face but with right-handed speakers this side is assumed to show speech face motion more clearly and exhibit less interference from emotional face expression.

- Our source data are notably different from previous studies. In our case, continuous speech was used, uttered in a lively manner as if reading the story of 'Chicken Little' to a child. The resulting face motion can be assumed to differ in various ways from the lab speech used in the previous studies, e.g., showing traces of emotional expressions.

The results from the current study, however, appear to be more in line with the performance of human speech reading than the high values from previous studies. They also might explain why audio-visual automatic speech recognition provides only gradual improvements over acoustic-only speech recognition instead of a breakthrough augmentation that could be expected from adding information from a different modality with noise unrelated to the noise in the base modality.

Putting the findings within a more general context it has to be mentioned that PLS is a linear method and it might be that the relationship between face motion and tongue movements is predominantly non-linear. Specifically, it can be hypothesised that the relationship is only linear over short time ranges with no face motion from other source than speaking interfering and non-linear when analysed in wider contexts. We thus intend to employ *Mutual Information* methods as a next step.

Results from those non-linear methods have to be awaited. Nevertheless, it can be speculated that the increase in intelligibility when watching speaker's face might be (aside from top-down linguistic processes) limited almost entirely to what is directly visible in speech articulation: jaw and lip movements and occasionally the tongue tip. There might be no globally independent tongue movement information in face motion as movements from other sources interact with speech face motion.

5. Acknowledgments

The authors would like to thank three anonymous reviewers for their insightful comments. We acknowledge the support of NHMRC/ARC grant TS0669874, and from a UWS International Research Initiative grant.

6. References

[1] 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.

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

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

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

[5] H. Kjellstrm, O. Engwall, and O. Blter, "Reconstructing tongue movements from audio and video," in *Interspeech 2006*, Pittsburgh, PA, USA, 2006.

[6] 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.

[7] 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.

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

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

[10] U. Andersson and B. Lidestam, "Bottom-up driven speechreading in a speechreading expert: The case of AA (JK023)," *Ear & Hearing*, vol. 26, no. 2, pp. 214–224, 2005.

[11] 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.

[12] J. Gower and G. Dijksterhuis, *Procrustes Problems*. Oxford University Press, 2004.

[13] C. Kroos, "Using sensor orientation information for computational head stabilisation in 3D Electromagnetic Articulography (EMA)," in *Proceedings of Interspeech 2009*, Brighton, UK, 2009, pp. 776–779.

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

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

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

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

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

[19] 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.

[20] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, "The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.