

Roles in articulation for speech animation

Veena D Singampalli and Philip JB Jackson*

Centre for Vision, Speech and Signal Processing, University of Surrey, UK.

Coarticulation modelling is important for visual speech synthesis to generate smooth and convincing articulation. Minimisation of effort in planning and production cause coarticulation, with the overlap of slow and continuous articulator movements. The visual realisation of a phone is affected by the neighbouring segments due to coarticulation, e.g., for anticipation of lip rounding. An accurate model of speech articulation can help in generating realistic facial animation. Coarticulation in visual speech has been modelled using various rule based techniques [1], theories of motor planning and speech production [2], and machine learning algorithms [5]. In articulatory control models, the constraints have been incorporated using phonetic knowledge.

We developed a statistical and data-driven algorithm for identification of articulatory roles from measured articulatory data [3]. The articulation constraint identification algorithm (ACIDA) was used to analyse different articulatory representations generated from the captured articulatory motion. The Electro-Magnetic Articulograph (EMA) data from MOCHA-TIMIT database [4] was used for this work. The data has 14 channels representing the horizontal (x) and vertical (y) movements of 7 articulatory points with the upper incisor and bridge of the nose as reference points. The articulatory points were located on upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD and velum V (Figure 1). The recordings were made as the speakers (one male and one female) uttered 460 sentences in English. The EMA data were smoothed and downsampled to 10 ms frames. Various principal components analysis (PCA) and linear discriminant analysis (LDA) representations were derived from the EMA data. For each representation, articulatory models for every phone were estimated using the ACIDA algorithm. The algorithm uses statistically significant correlations amongst the articulators and spatial correlations within articulators in identification of roles and in estimation of model articulatory distributions. The Gaussian articulatory models thus generated were found to be more compact than the conventional statistical representations. The goodness of fit of models from each representation to the actual phone distributions was analysed using evaluation scale Υ_{eval} , i.e., the KL divergence between the model distributions and actual phone distributions (with full 14D covariance matrix). The results (Figure 1) show that the LDA based representations generated better fitting models (smaller Υ_{eval} values).

The articulatory role information was also used in generating synthetic articulatory trajectories. The results for synthesis, which includes tongue, teeth, jaw and lip motion were compared with the recordings. The synthesis work presented in this paper is aimed at generating realistic speech movements from the parsimonious representations and role information. Our algorithm and methods of analysis are transferable and could readily be applied to facial data (e.g., marker or mesh vertex coordinates). Gestures in articulatory control models for synthesis can also be prioritised in a data-driven way using ACIDA.

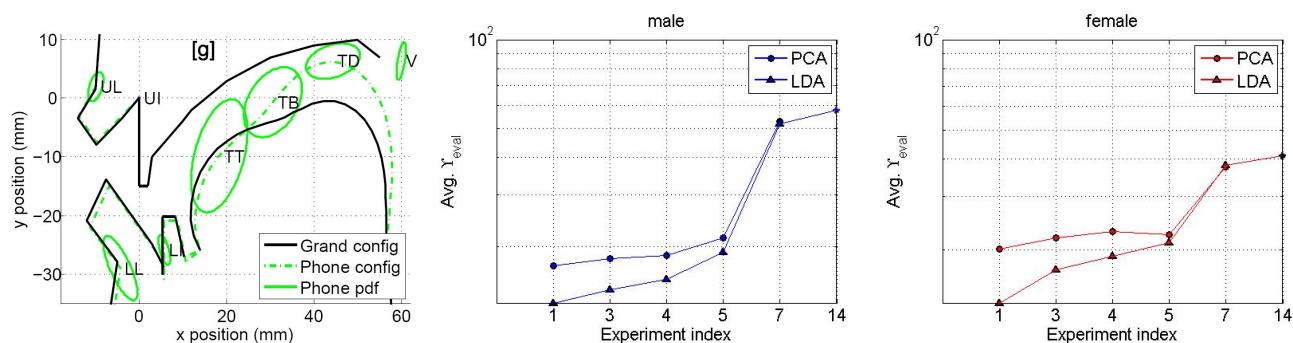


Figure 1: Midsagittal display of [g] (left); average evaluation divergence for male (middle, blue) and female (right, red) speakers for different PCA and LDA experiments at IPA level threshold.

*This research was funded by EPSRC (GR/S85511/01) under the DANSA project [www.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/].

References

- [1] J. Beskow. Rule-based visual speech synthesis. In *Proc. Eurospeech*, pages 299–302, 1995.
- [2] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, 1993.
- [3] P. J. B. Jackson and V. D. Singampalli. Statistical identification of articulation constraints in the production of speech. *Speech Communication*, In press, 2009.
- [4] A. A. Wrench. A new resource for production modelling in speech technology. *Proc. Inst. of Acoust., Stratford-upon-Avon, UK*, 23(3):207–217, 2001.
- [5] J. Xue, J. Borgstrom, J. Jiang, L. Bernstein, and A. Alwan. Acoustically-driven talking face synthesis using dynamic bayesian networks. *IEEE Int. Conf. on Multimedia & Expo*, 0:1165–1168, 2006.