Abstract: Statistical identification of critical parts of articulatory apparatus during an utterance. Evaluated EMA, PCA and LDA models, and synthesised motion.

DANSA

Roles in articulation for speech animation

Veena D Singampalli & Philip JB Jackson
CVSSP, University of Surrey, UK
v.singampalli, p.jackson@surrey.ac.uk
www.ee.surrey.ac.uk/personal/p.jackson/dansa

UNIVERSITY OF SURREY

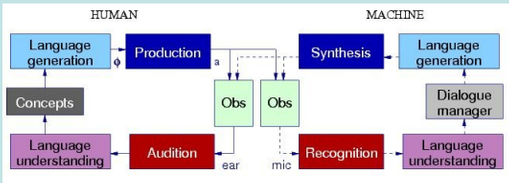
EPSRC Engineering and Physical Sciences Research Council

HOW IS SPEECH PRODUCED?

Speech is dynamic

Speech production forms part of an interactive dynamical system in the speech loop [8,12]

Feedback loops enable training, adaptation and control of speech



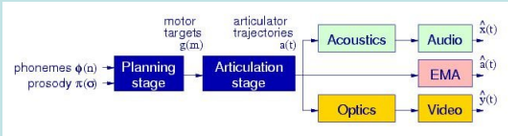
Coarticulation effects

Coarticulation is a major challenge for visual speech [1,3,17]

Coarticulation affects the *planning stage* of speech production to simplify commands for sequence of speech gestures [2,11]

Coarticulation affects the *articulation stage* from stiffness, damping and inertia, and effort minimisation [4,9,10,13,14]

Present work uses statistical models to determine a representation of gestural commands for each phone [15]



Rôles in articulation [6]

Critical: essential for a phone's production

Dependent: follows critical articulator due to connection with it

Redundant: free to move, unconstrained by current phone

WHEN IS A MOVEMENT CRITICAL?

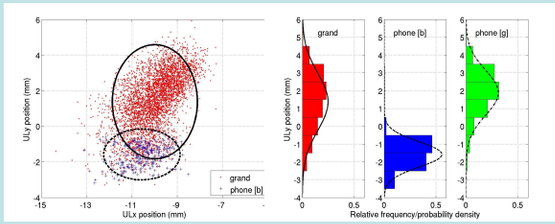
Symmetric Kullback-Leibler divergence (KLD)

Measures distance between two distributions in terms of average log-likelihood ratio [7]:

$$J = \int_{-\infty}^{\infty} f_1(x) \ln \frac{f_1(x)}{f_2(x)}\, dx + \int_{-\infty}^{\infty} f_2(x) \ln \frac{f_2(x)}{f_1(x)}\, dx$$
$$= \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_1 - \Sigma_2\right)\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right) + \mathrm{tr}\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^{\mathrm{T}}\right]$$

where $f_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $f_2 = \mathcal{N}(\mu_2, \Sigma_2)$.

Articulatory probability density functions (pdfs)



High KLD (9.1) for [b] marks ULy as critical for bilabial correlated LLy becomes dependent for [b]

Low KLD (0.2) for [g] shows that Uly is redundant for velar

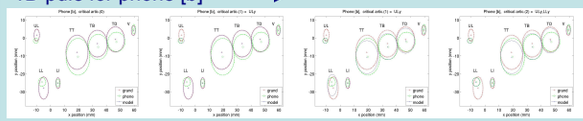
Φ	msak 1D	msak 2D	fsew 1D	fsew 2D
[f]	lly uly	ll	lly lbx	ll
[v]	lly uly	ll	lly lbx	ll
[th]	tbx tty lly	tt ll	tty ttx tby	tt tb ll li
[dh]	tbx tty	tt	tbx tty	tt
[s]	lly ttx tty	tt li	lly tty tbx	li tt
[z]	lly ttx tty	tt li	lly tty tbx	li tt
[sh]	tty tbx lly tdy	tt li	tty lly tbx	li tt
[zh]	lly tty tdy ttx lly	li tt td	tty lly tby tdx lbx	tt li tb
[ch]	lly tty tbx tby	tt li	tty lly tbx	li tt ul
[jh]	tty tby tbx lly	tt tb	lly tty tbx	li tt
[h]	-	-	tty	tt
[p]	uly l/y	ul	uly	ul
[b]	uly l/y	ul ll	uly lly	ul ll
[t]	tty	tt	tty l/y	tt
[d]	tty	tt	tty	tt
[k]	tdy	td	tdy	td
[g]	tdy	td	tdy	td
[m]	uly l/y vx	ul ll	lly uly	ll ul
[n]	tty vx	tt	tty	tt
[ng]	tdy vx	td v	tdy	td
[w]	uly	ul	uly tdy lly	ul td ll
[l]	-	-	-	-
[r]	tbx	tt	ttx	tt
[y]	tby	tb	tby	tb tt

Φ	msak 1D	msak 2D	fsew 1D	fsew 2D
[ii]	tdy lly ttx	tb	tby	tb td
[iy]	tdy	tb	tdy ttx	tb td
[uu]	tdy	-	tby	td
[i]	-	-	-	-
[u]	-	-	-	-
[@@]	lly	-	tty lbx	tb ll
[uh]	-	-	tby	tb
[oo]	lly tbx tty	tb ll	tdx lly tdy tby	td ll
[o]	tby	tb	tby tdx	tb
[aa]	lly tby tbx	ll tb	tby tdx	tb
[a]	lly	ll	lly tdy	ll td
[e]	lly	-	lly	ll

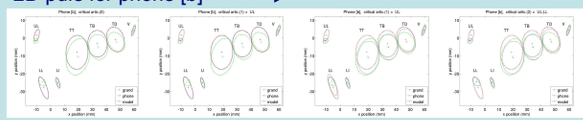
WHICH REPRESENTATION IS BEST?

Pdf convergence as critical articulators identified

1D pdfs for phone [b] →



2D pdfs for phone [b] →



Some strong correlations exist across articulators

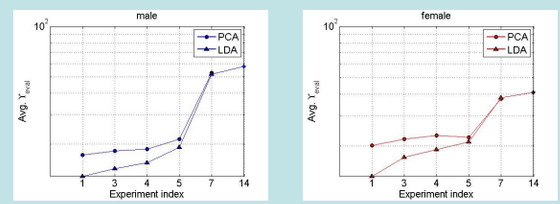
msak (male)

M	ULy	ULx	LLy	LLx	Lly	Llx	TTy	TBy	TDy	TTx	TBx	TDx	Vy	Vx
ULy	1.00	.51	.28	.24	.00	.00	.53	.24	.11	-.17	-.13	-.15	-.12	.00
ULx	.51	1.00	.18	.33	.20	.00	.00	.00	.00	.24	.22	.25	.17	.00
LLy	.28	.18	1.00	.79	.89	.64	.00	.00	.00	.13	.00	.00	.00	.00
LLx	.24	.33	.79	1.00	.58	.60	.38	.00	.16	-.12	-.25	-.18	-.16	-.14
Lly	.00	-.20	.66	-.58	1.00	.70	.63	.42	.00	.00	.00	.00	.00	.00
Llx	.00	.00	.48	.60	.70	1.00	.46	.37	.00	.00	.00	.00	.00	.00
TTy	.53	.00	.34	.38	.63	.46	1.00	.56	.11	.00	.16	.11	.00	.00
TBy	.24	.00	.00	.00	.42	.37	.56	1.00	.73	.00	.00	.00	.14	.00
TDy	.11	.00	.00	.16	.00	.00	.11	.73	1.00	.00	.24	.20	.29	.14
TTx	-.17	.24	.00	-.12	.00	.00	.00	.00	.00	1.00	.90	.85	.10	.12
TBx	.13	.22	.13	.25	.00	.00	.16	.00	.24	.90	1.00	.96	.00	.00
TDx	-.15	.25	.00	.18	.00	.00	.11	.00	.20	.85	.96	1.00	.00	.00
Vy	-.12	.17	.00	.16	.00	.00	.00	.14	.29	.10	.00	.00	1.00	.79
Vx	.00	.00	.00	.14	.00	.00	.00	.00	.14	.12	.00	.00	.79	1.00

fsew (female)

F	ULy	ULx	LLy	LLx	Lly	Llx	TTy	TBy	TDy	TTx	TBx	TDx	Vy	Vx
ULy	1.00	.11	-.27	-.21	.18	.00	.32	.22	.00	.00	.00	-.15	.00	.00
ULx	.11	1.00	.38	.55	.15	.00	-.12	.00	.00	.24	.28	.21	.00	.00
LLy	.23	.38	1.00	.88	.54	.22	.32	.00	-.15	.00	.00	.00	.00	.14
LLx	.21	.55	.88	1.00	.46	.53	.32	.00	.12	.00	.00	.11	.00	.00
Lly	.18	-.13	.54	.46	1.00	.48	.59	.33	.00	-.17	-.18	-.23	.00	.15
Llx	.00	.00	.27	.53	.48	1.00	.38	.00	.13	.19	.33	.37	.00	.30
TTy	.32	.12	.00	.32	.59	.38	1.00	.86	.30	.14	.00	.31	.16	.23
TBy	.22	.00	.00	.00	.33	.00	.86	1.00	.77	.00	-.12	.00	.00	.20
TDy	.00	.00	-.13	.12	.00	.13	.00	.77	1.00	.00	.00	.00	.17	.22
TTx	.00	.24	.00	.00	-.17	.19	.14	.00	.00	1.00	.86	.79	.00	.00
TBx	.00	.28	.00	.00	-.18	.33	.00	-.12	.00	.86	1.00	.96	.00	.14
TDx	-.15	.21	.00	.11	-.23	.37	.31	.00	.00	.79	.96	1.00	.00	.00
Vy	.00	.00	.00	.00	.16	.00	.16	.00	.17	.00	.00	.00	1.00	.58
Vx	.00	.00	.14	.00	.15	-.30	.23	.20	.22	.00	-.14	.00	.58	1.00

Goodness of fit with linear transformations, PCA and LDA



Null case (expt.14) equivalent to EMA coordinates

Various PCA and LDA schemes considered

Evaluation scale based on full 14D KLD

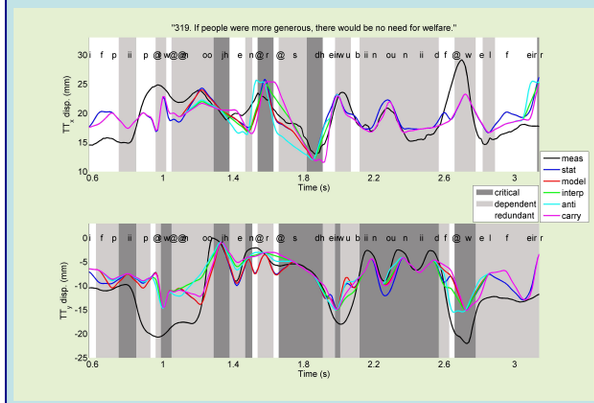
Expt.1 combined all coordinates, and LD1 performed best

WHAT ARE THE ANIMATION RESULTS?

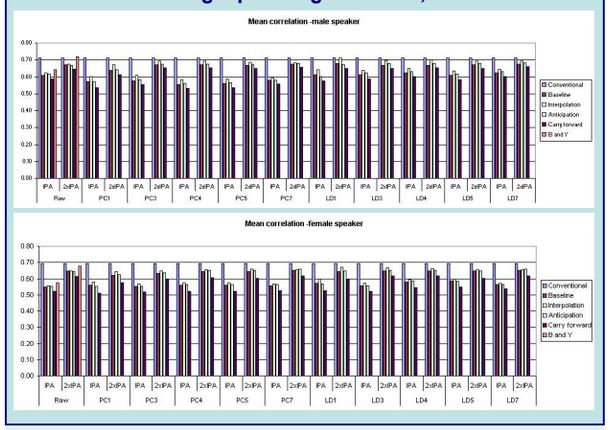
Measured vs. synthesised articulator motion

Articulator targets determined using knowledge of the critical, dependent and redundant roles

Trajectories produced by simple blending



Performance of target planning with EMA, PCA and LDA

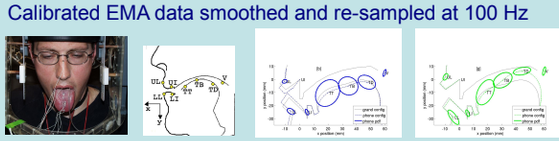


Electromagnetic articulography from MOCHA-TIMIT [16]

Male (msak) and female (fsew) British English subjects

460 sentences with phone transcription (manual correction)

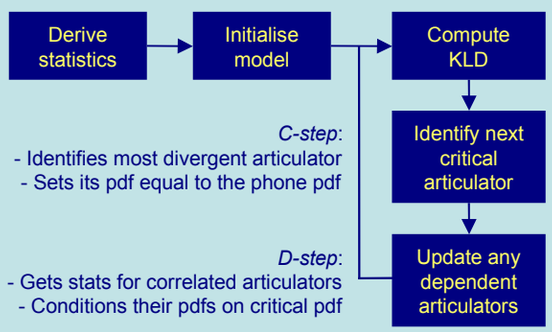
Calibrated EMA data smoothed and re-sampled at 100 Hz



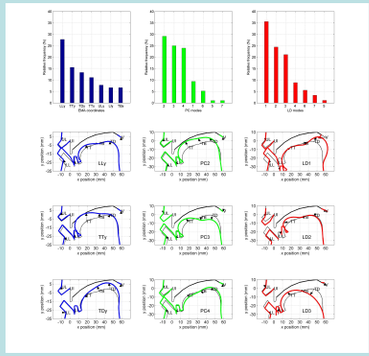
2D midsagittal traces of 7 glued coils:

UL - upper lip
LL - lower lip
LI - lower incisor
TT - tongue tip
TB - tongue blade
TD - tongue dorsum
V – velum
Nose & upper incisor used as ref.

Articulatory constraint identification algorithm (ACIDA)



Most-often critical dimensions



SUMMARY

Conclusions

Presented framework for use of critical, dependent and redundant articulator roles in speech animation

Explored various linear transformations of EMA data:
- both PCA and LDA improved fit at cost of interpretability
- LD1 gave best fit to phone pdfs

Preliminary results presented using roles to modify targets

Future perspectives

To extend the range of contexts in the planning stage

To investigate mode advanced motion synthesis methods in the articulation stage

References

[1] J. Beskow. Rule-based visual speech synthesis. In Proc. Eurospeech, p.299–302, 1995.
[2] N. Chomsky and M. Halle. The sound pattern of English. Harper & Row, 1968.
[3] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In Models and Techniques in Computer Animation, p.139–156. Springer-Verlag, 1993.
[4] J. Dang, M. Honda, and K. Honda. Investigation of coarticulation in continuous speech of Japanese. Acoust. Sci. & Tech., 25(5):318 – 329, 2004.
[5] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. Proc. ICASSP, 4:317–320, 2007.
[6] P. J. B. Jackson and V. D. Singampalli. Statistical identification of articulation constraints in the production of speech. Speech Comm., 51(8):695-710, 2009.
[7] S. Kullback. Information theory and statistics. Dover Pub., New York, 1968.
[8] W. Levelt. Speaking: From Intention to Articulation. Cambridge, MA: MIT Press, 1989.
[9] B. Lindblom. Spectrographic study of vowel reduction. J. Acoust. Soc. Am., 35:1773–81, 1963.
[10] A. Löfqvist. Speech as audible gestures. In W.J. Hardcastle and A. Marchal (eds.), Speech production and Speech Modeling. Dordrecht: Kluwer Academic Publishers, 1990.
[11] K. Moll and R. Daniloff. Investigation of the timing of velar movements during speech. J. Acoust. Soc. Am., 50(2):678–84, 1971.
[12] R. K. Moore. PRESENCE: A human-inspired architecture for speech-based human-machine interaction. IEEE Trans. on Computers, 56(9), 2007.
[13] S. Öhman. Numerical model of coarticulation. J. Acoust. Soc. Am., 41(2):310–20, 1967.
[14] E. Saltzman and K. Munhall. A dynamic approach to gestural patterning in speech production. Ecology Psychology, 1(4):333–82, 1989.
[15] V. D. Singampalli and P. J. B. Jackson. Statistical identification of critical, dependent and redundant articulators. Proc. Interspeech, Antwerp, p.70–73, 2007.
[16] A. Wrench. A new resource for production modelling in speech technology. Proc. Inst. of Acoust., Stratford-upon-Avon, UK, 2001.
[17] 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.

Poster presented at BMVA workshop on facial analysis and animation in Edinburgh on Wednesday 10 June 2009