Coarticulatory constraints determined by automatic identification from articulograph data

Philip JB Jackson and Veena D Singampalli

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
E-mail: p.jackson@surrey.ac.uk

Abstract

A statistical technique for identifying critical, dependent and redundant articulators in English phones was applied to 1D and 2D distributions of articulatograph coordinates. Results compared well with phonetic descriptions from the IPA chart with some interesting findings for fricatives and alveolar stops. An extension of the method is discussed.[1]

Keywords: speech production, articulatory gesture, coarticulation model, critical articulator

1 Introduction

Coarticulation remains a central challenge in speech research. In the production of speech sounds, e.g., from a specified sequence of phonemes, coarticulation spreads their influence across the utterance so that substitution of one phoneme for another (or the change of one distinctive feature) modifies the corresponding phone segment and its neighbours. The primary cause is that human speech articulators (jaw, tongue, lips, etc.) have limited freedom to move, interconnections, stiffness, damping and inertia. Speech is thus planned in a coordinated sequence of relatively slow and overlapping muscle gestures expending minimal energy. So, it is difficult to match linguistic units with acoustic data and account for all the variability observed. State-of-the-art TTS and ASR systems use ever longer units and models to accommodate coarticulatory effects without using explicit knowledge of the articulatory constraints that convert a phoneme string into speech.

Coarticulation theories tend to focus on one or other of the two main stages in the linguistic-to-acoustic realization. Feature spreading theory deals with effects in the planning stage, converting phoneme sequences into distinctive articulatory features, and gestural theory with the motor control and physical dynamics of the articulators. In the feature spreading approach, a distinctive set of bipolar phonetic features encodes the place, manner and voicing information of a phone [1], while non-critical features are left unspecified. Anticipatory coarticulation then spreads features to unspecified segments, [7]. Yet, speech articulation is actually a continuous process, in time and space.

Using vowel formants to investigate coarticulation with consonants [5], Lindblom estimated locus equations in CVC utterances. Öhman's hand-crafted model [8] distinguished the fast and localised consonant gestures from vocalic ones, and recognized that those articulators not actively involved in producing the consonantal gesture (i.e., not critical) were most influenced by vowel context. The idea of a crucial or critical articulator has been defined as being resistant to contextual effects and having maximum coarticulatory influence on its neighbours [2]. Coarticulation results from the co-production of gestures [9, 6], prescribed according to phonetic rules.

The present work identifies articulatory roles, defined as critical, dependent and redundant, from quantitative and statistical measurements. Our statistical approach [10] identifies these constraints in the articulatory domain. If an articulatory gesture is essential for the production of a phone, it is considered to be *critical*. As in previous work, the critical articulators were associated with smaller variance compared with others. Here, an articulator can be critical if a significant shift in mean from its neutral average state arises. As the form of articulatory

[1] DANSA project was funded by EPSRC (GR/S85511/01), http://personal.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/.

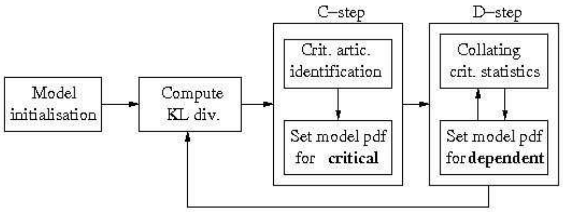
C–step
D–step
Model initialisation
Compute KL div.
Crit. artic. identification
Set model pdf for critical
Collating crit. statistics
Set model pdf for dependent

Figure 1: Critical articulator identification algorithm.

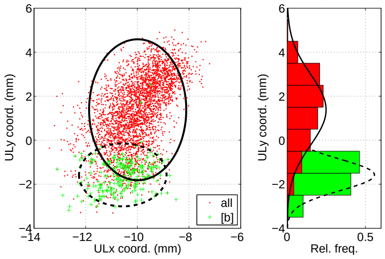

Figure 2: (a): scatter plot of grand (red ·) and phone [b] (green +) upper lip coordinates (ULx and ULy) at the centre of each phone label; ellipses show ±2 standard deviations of the corresponding grand (solid) and phone (dashed) normal distributions, which encompass 95% of the points. (b): histograms with fitted Gaussian pdfs of ULy for grand (red/solid) and phone [b] (green/dashed).

variation in space could characterise a phone's production, we included covariance changes too. A *dependent* articulator is defined as one that follows a critical articulator due to bio-mechanical correlation between them.[2] A *redundant* articulator is free to move, its position unaffected by the current phone, so coarticulation effects from neighbouring phones are strongest. The change in the spatial distribution of an articulator is used to identify whether it is critical or not for a given phone. The procedure is entirely data driven and generates parsimonious representations of the articulatory target configurations for each phone. The next sections outline the method and a phonetic analysis of the results, followed by a discussion and conclusion.

2 Identification of articulatory constraints

The block diagram in Figure 1 shows the main steps in the identification procedure for each phone. Electro-magnetic articulograph (EMA) data for two speakers (male msak, female fsew) from the MOCHA-TIMIT database [11] provided horizontal (x) and vertical (y) midsagittal measurements for 460 phonetically-balanced British English sentences at 7 points: upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD and velum V. Smoothed samples selected at phone midpoints yielded phone distributions, whose mean and covariance defined Gaussian phone pdfs. The data set's overall mean and covariance for each articulator defined the grand pdf. For each phone, critical articulators were identified based on the symmetric Kullback-Leibler divergence (KLD) between grand and phone pdfs [4], which incorporates changes in mean and covariance. Fig. 2 shows grand and phone [b] distributions from ULy. The KLD was high (9.1), flagging ULy as critical for labial [b], whereas the divergence was low (0.2) for velar [g]. The pdfs of dependent articulators were conditioned on the critical articulator pdfs using grand inter-articulator correlations, while redundant articulators were unaltered. For example, the vertical movement of the tongue tip TTy was correlated with that of the jaw (LIy) but not the back of the tongue (TDy). The algorithm was implemented for 1D (independent x and y) and 2D (including covariation) pdfs.

Figure 3 illustrates the 1D critical articulator identification algorithm for phone [g]. A model described by a univariate Gaussian distribution was allocated to each articulatory coordinate for each phone. In **model initialisation** (Fig. 3a), the model means and variances were set as the grand distributions. The KLD between model and phone pdfs was calculated in the **compute KL div.** stage for each articulatory coordinate. In the critical identification step (**C-step**), the articulatory coordinate with maximum divergence was identified as critical, and its model distribution set to the phone distribution. For phone [g], TDy was identified (Fig. 3b).

In the dependent update or **D-step**, the distributions of the correlated articulators were conditioned on the critical articulator's new distribution using their grand correlation with it. For phone [g], the effect of TDy's position on the distribution of TBy can be seen in Fig. 3c, but there was no change in the TT distribution since it had no substantial correlation with TDy. To identify all the critical articulatory coordinates for each phone, the algorithm functions iteratively, until all the divergences fall below the critical threshold, θ_C. Fig. 3d shows that Vx was identified as a second critical corrdinate for [g].

[2] This kind of gesture is often called a passive gesture owing to correlation with an active tract variable [9].

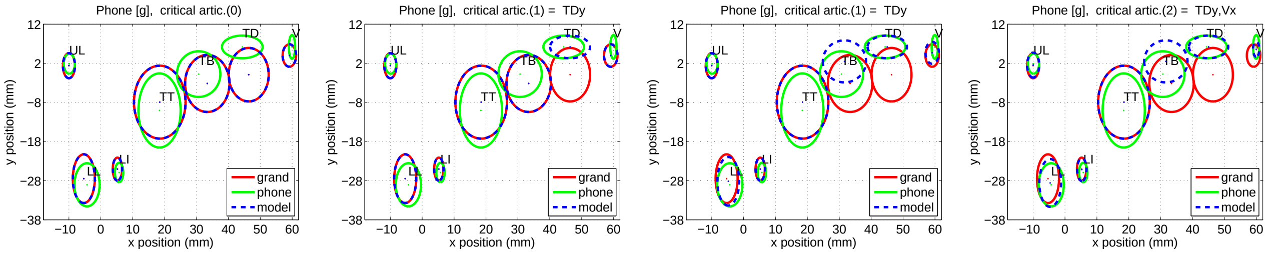

Figure 3: (a,b,c,d): Convergence of 1D distributions as each critical articulator was identified: grand (red), phone (green) and model (dashed blue) pdfs.

3 Phonetic analysis of results

Critical articulator lists obtained for each consonant and vowel were compared to international phonetic alphabet (IPA) descriptions. Although IPA gives a notational standard for phonetic transcription, it provides a good basis for comparison because it is a long-standing, thoroughly-debated summary of speech sound categories in the world's languages. It was designed (i) to capture linguistic distinctions, (ii) for human speech, (iii) observed by phoneticians. Increasingly, speech technologies need phonological descriptions that can (i) model typical phone characteristics within a voice/dialect, (ii) include implicit effects found in human phoneme-to-phone realisation, such as coarticulation, and (iii) incorporate other knowledge sources, e.g., from X-ray, MRI and EMA. For comparison, IPA-based pdfs were generated using manner and place attributes to define a set of critical articulators for each phone. The algorithm's critical threshold was adjusted to identify the same number of critical articulators in total across all phones ($\theta_C = 2.2/2.4$ for male/female).

3.1 Consonants

No critical articulation was identified for 20% of consonants: [h] and the alveolars [t,n,l] characterised by rapid movement of the tongue tip. The velum was more often identified in the closed than open position, for stops and fricatives rather than nasals, and fricatives incorporated a secondary articulation. Despite a few insertions, substitutions and speaker differences, the algorithm produced plausible constraints to fit the data in broad agreement with IPA descriptions and highlighted implicit or absent details.

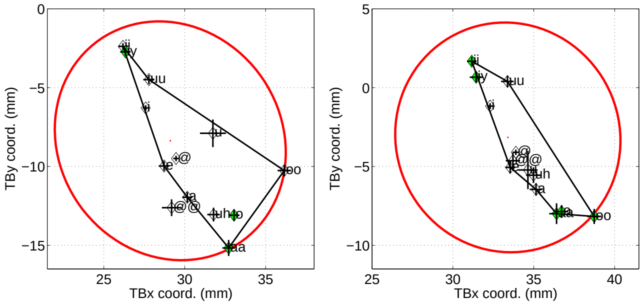

Figure 4: Tongue blade (TB) phone means for (a) female and (b) male EMA vowel data, with grand covariance ellipse (red) and vowel quadrilateral (black).

In a more detailed analysis of alveolar obstruents [t,d,n], we investigated why no critical articulators were identified (except LIy for female [n]). Though TT had the highest divergence, it was below the chosen θ_C (at IPA level). Likely factors are: (i) the labelling accuracy, (ii) the choice of sampling time, and (iii) non-Gaussianity of articulator coordinate distributions.

Outliers in articulator distributions for alveolars occurred through labelling errors in the database: (i) incorrect word labels (1%), and (ii) elision in the continuous utterances (5%). Here, the TTy traces showed no indication of alveolar articulation. The removal of these unreliable data gave a clear improvement: TTy was identified as the primary articulator for [t,d,n] (male and female). The male [n] also identified the velum (Vx).

Samples were taken at the phone midpoints assuming that the articulator approached its target then. The time of maximum TTy coincided with the start of the phone for more than 70% of the plosives [t,d] on average, which suggests that further improvements could be achieved by taking the middle of the closure (before release) as the salient time instant.

3.2 Vowels

Since the IPA vowel quadrilateral is part acoustic and part articulatory (tongue height, backness and lip rounding), the critical articulators were less clearly defined, as in Fig. 4. Tongue blade and dorsum dominated the identified articulator lists, especially for open-back and close vowels, but no critical articulation was identified for 25%, e.g., [ə, i, ʌ], whose configurations were more susceptible to coarticulation. Vowel openness led to identification of the jaw LI and rounding sometimes identification of LL, although it was not always clear in these midsagittal data. While vowels were harder to predict, the identified critical articulators did pattern across the quadrilateral, and verified the different strength of the constraints, especially for short and reduced vowels. Thus, the results of the algorithm offer some interesting statistical insights into specification of vowel targets in the production of fluent speech.

4 Discussion

One potential extension of the present algorithm concerns the assumption that the grand and phone distributions are Gaussian, which simplifies the calculation of KLD and makes it analytically tractable. A single-sample Kolmogorov-Smirnov test (with Lilliefor's correction) for goodness of fit of the samples to normal distributions was performed (at level of significance $\alpha = 0.05$). The grand distributions failed the KS test for all articulators, as did 60% of the phone distributions. Note that the sample size and outliers from labelling errors affect the KS test results. Using a mixture of Gaussians would give better models, but significantly complicate the KLD calculation, although numerical methods such as Monte Carlo simulations can be used [3].

5 Conclusion

Using an algorithm for identifying critical, dependent and redundant roles played by articulators during speech production, we have presented results of application to the MOCHA-TIMIT corpus of EMA coordinate data. Results were compared to IPA descriptions of the speech sounds. For consonants, the role of the velum in fricative and stop production was highlighted, with secondary roles for the teeth and lips acting as obstacles in fricatives. An investigation of the alveolar consonants showed how inaccurate labelling of articulatory events can influence the identification of critical articulators. Analysis of vowels showed a distinction between full and reduced vowels, whose configurations were more susceptible to coarticulation.

Many potential ways to embed models of coarticulation and trajectory generation in speech synthesis and recognition could exploit the link between phonemes and their realisation, as explicit articulatory models or used implicitly, e.g., join cost. In phonetics, articulatory roles can help explain coarticulation effects, and derive phonetic inventories for different languages, speakers and dialects. Further work would develop a coarticulation model to describe dynamic constraints by offering distributions of articulatory trajectories from a planned sequence of phonemes.

References

[1] N. Chomsky and M. Halle. *The sound pattern of English.* Harper & Row, New York, 1968.

[2] J. Dang, M. Honda, and K. Honda. Investigation of coarticulation in continuous speech of Japanese. *Acoust. Sci. & Tech.*, 25(5):318 – 329, 2004.

[3] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. *Proc. ICASSP*, 4:317–320, 2007.

[4] S. Kullback. *Information theory and statistics.* Dover Pub., New York, 1 edition, 1968.

[5] B. Lindblom. Spectrographic study of vowel reduction. *JASA*, 35:1773–81, 1963.

[6] A. Löfqvist. Speech as audible gestures. *In W.J. Hardcastle and A. Marchal (Eds.), Speech production and Speech Modeling. Dordrecht: Kluwer Academic Publishers*, 1990.

[7] K. Moll and R. Daniloff. Investigation of the timing of velar movements during speech. *JASA*, 50(2):678–84, 1971.

[8] S. Öhman. Numerical model of coarticulation. *JASA*, 41(2):310–20, 1967.

[9] E. Saltzman and K. Munhall. A dynamic approach to gestural patterning in speech production. *Ecology Psychology*, 1(4):333–82, 1989.

[10] V. D. Singampalli and P. J. B. Jackson. Statistical identification of critical, dependent and redundant articulators. *Proc. Interspeech*, Antwerp, pages 70–73, 2007.

[11] A. Wrench. A new resource for production modelling in speech technology. *Proc. Inst. of Acoust.*, Stratford-upon-Avon, UK, 2001.