Planning Compensates for the Mechanical Limitations of Articulation

Greg P. Kochanski & Chilin Shih

The University of Oxford & University of Illinois at Urbana-Champaign
greg.kochanski@phon.ox.ac.uk & cls@uiuc.edu

Abstract

We explore a simple model of speech articulation. The model consists of an articulator combined with the ability to remember and improve the neural drive signal for the articulator. Over many productions, the system learns a neural drive signal that provides an accurate match for acoustically-defined targets. In fact, the match can be better than expected, yielding narrower regions of coarticulation than the intrinsic muscle response time. Further, despite the time delay introduced by the muscle, the articulatory response has no time delay, because the learned neural drive signal occurs in advance of changes in the acoustic targets. Finally, we test the model against tonal production data from Mandarin conversation, and show that it can represent non-trivial surface intonation patterns with simple and linguistically reasonable targets.

1. Introduction

In this paper, we will investigate a simple model of articulation that we call the **M**inimal **A**rticulatory **L**earning **M**odel (MALM). It is designed to be a simple schematic representation and yet capture some of the essential details of the physiology and language learning process.

The model simulates the learning process of improving motor skills through practice and evaluation. It applies to commonly repeated speech fragments where the speaker has the opportunity to try different productions and observe the effect. Presumably, speakers make an effort to improve their articulation from production to production to ensure that their listeners can understand and to meet social norms.

The model consists of a controller ("brain"), an articulator and some acoustic targets. The controller follows a Markov-Chain Monte-Carlo algorithm [9, 4]: it takes the stored pattern of neural impulses, and produces a randomly modified version. If the modified version gives a production that is a better match to the targets than the stored pattern, the modified version is accepted and stored. It will then be the basis of future modifications. Otherwise, the modified pattern is discarded and a differently modified pattern will be tested in the next iteration.

The articulator is a critically-damped impulse response, which can optionally have an exponential relationship between muscle length and the resulting f_0 [2]. We investigate the behavior of the model, and compare it with examples of tonal production data from Mandarin conversation.

2. Model Details

At each iteration, the model is defined in terms of n_t, which is the rate of nerve firing as a function of time. We sample this (and other signals) at 10 ms intervals. This signal then excites the articulatory muscle, which has a critically damped response:

$$G(\Delta t) = \begin{cases} \frac{\Delta t}{\tau} \cdot e^{-\Delta t/\tau} & \text{if } t > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We choose $\tau = 100$ ms for these examples, to approximately match the rate at which people can adjust their pitch [11]. The results presented here are neither critically dependent on the exact choice of the articulatory response nor τ.

We take the produced f_0 to be equal to a nonlinear function H of the the convolution of n_t and $G(\Delta t)$:

$$f_t = H\left(\sum_i n_i G(t-i)\right). \quad (2)$$

This corresponds to a simple linear superposition of the force from each nerve impulse, and H provides a convenient encapsulation of the anatomy and aerodynamics of speech.

It can be argued that the correct relationship between muscle response and f_0 (i.e. H) is nonlinear. [2] has proposed an exponential relationship between muscle length and f_0, while one might alternatively expect an f_0 to be proportional to the square root of the tension in the laryngeal fold to the extent a vibrating string model is appropriate. For Figure 1, we choose H to be a Fujisaki exponential, but for simplicity in the remaining figures we take $H(y) = y$. We find little qualitative difference in the results for any reasonable choice of H.

The model gradually learns the optimal n_t via a simple iterative algorithm.[1] It randomly changes n_t, computes f_t via Equation 2, measures of the error between f_t and the acoustic targets, then accepts the change if the error has decreased. Otherwise, it rejects the change and repeats. Each iteration of this process corresponds to producing an utterance then deciding if it sounded right, was understood, or was socially approved.

Specifically, it computes a measure of the difference between the acoustic output and the desired target. Most simply, this can be

$$R_a = \sum_t (f_t - y_t)^2 \cdot S_t^2, \quad (3)$$

which measures how far the produced frequency is from the target frequency. It can alternatively be

$$R_s = \sum_t (\dot{f}_t - \dot{y}_t)^2 \cdot s_t^2, \quad (4)$$

where the dots show differentiation vs. time; this measures how much the produced slope differs from the intended slope. We make a combination error measure

$$R = R_a + R_s, \quad (5)$$

[1] Note that there is a different n_t for each combination of tones. The model implies that people store perhaps thousands of neural drive patterns, each learned separately.

which can act like R_a, R_s, or anywhere in between depending upon s and S. The two sets of coefficients, s_t and S_t control which aspects of the target are important: If S_t is small, then the absolute frequency is unimportant near time t, and if s_t is small, then the slope of f_t is unimportant near time t. If both s and S are small, then the target will have little influence on the resulting computed f_0.

Then we generate a randomly modified neural drive signal, $\tilde{n}_t$ from n_t, compute $\tilde{f}_t$ and f_t from them, then $\tilde{R}$ and R, which are the errors between the desired target and the f_0 pattern produced by $\tilde{n}_t$ and n_t, respectively. If $\tilde{R} < R + X$, we accept the change and assign $n_t \leftarrow \tilde{n}_t$, where X is a random variable chosen from a exponential distribution with a mean of one.

The algorithm accepts all steps that decrease R, but also accepts some steps that make small increases in R. The algorithm never terminates; it rolls down-hill to a solution for n_t that gives close to the minimum possible R. Then, if there are many control signals that give nearly optimal values of R (which there normally are), the algorithm will randomly walk from one near-optimal instance of the control signal to another.

The random modification is generated by adding a random variable chosen from a zero-mean Gaussian probability distribution to each component of n_t. The variance of this Gaussian is increased by 10% whenever a change is accepted, and decreased by 2.5% whenever a change is rejected. This adjustment of σ adjusts the step size so that approximately 25% of all steps are accepted.

If we take our model to represent the muscles in the vocal folds themselves, then they cannot push, so no nerve impulses yield a relaxed muscle, loose vocal folds, and a correspondingly low oscillatory frequency. The rate of nerve impulses cannot go below zero, so we constrain the possible values of n_t to be positive. If the random step would cause $n_t < 0$ for some $t = \tilde{t}$, we replace $n_{\tilde{t}} \leftarrow -n_{\tilde{t}}$.

Similarly, there is a maximum possible firing rate L, and if $n_t > L$ for some $t = \tilde{t}$, we replace $n_{\tilde{t}} \leftarrow 2L - n_{\tilde{t}}$. The existence of minimum and maximum allowable values for n_t is crucial. It can be shown that if there are no limits, then the overall system can learn to perfectly reproduce any targets with no error; coarticulation would then be nonexistent.

2.1. Articulatory Targets

In the models we use, an "articulatory target" is the articulatory invariant of a phonological feature. It is a theoretical construct and may not be precisely realized, especially if there are other nearby and/or conflicting targets. A target can be a combination of a articulator position and velocity, that manages to communicate the phonological feature to the listener.

An analogy can be made to throwing darts. For a dart to reach its target, the hand and arm must move through certain positions at the right speeds. Trade-offs may occur between position and speed or between the motions of one joint (analogous to articulator) and another. So, the articulatory invariant of a throw that hits the target is not defined by a single position, but is instead a range of positions, with a specific velocity for each position.

3. Behavior of the Model

Figure 1 shows what the model is capable of learning. We set the target f_0 target to be four Mandarin rising tones, represented in the figure as the saw-tooth pattern of black dots. In this example, the targets constrain the f_0 values but not the slopes: $s = 0$

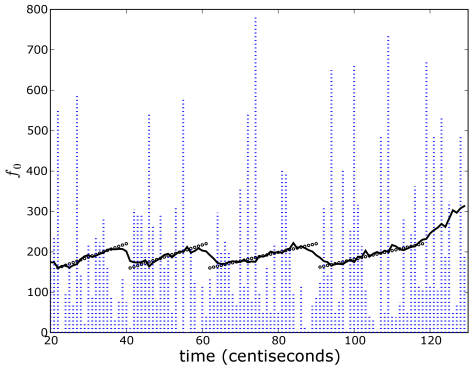

Figure 1: Planning in the form of anticipatory coarticulation compensates for limitations of muscle response. The vertical axis is f_0, and the horizontal axis is time in centiseconds. The acoustic target is shown as the sloping lines of dark black dots, the predicted model output as the black curve, and n_t as the vertical blue (gray) bar chart in the background.

and $S > 0$. We set the maximum nerve impulse rate $L = 800$, about four times the average of n_t, and we use a moderately nonlinear $H(x) = 190 \cdot e^{(y-190)/100}$. The pattern of nerve impulses is then optimized to minimize the difference between the resulting f_0 and a desired f_0 pattern.

After 5000 steps of the learning algorithm, the MALM model output matches the target f_0 with a mean-squared frequency error of 8.6 Hz. In two respects,the performance is better than what might be expected given the muscle response:

- The response time of the overall system is about 30 ms in this example. This is faster than the 100 ms response time of the muscle.
- The overall system response is coincident with steps in the target, not delayed by the muscle's response time of about 100 ms.

The system achieve this result by anticipating the movement through iterative learning. It compensates for the smoothing effect of the muscle's response by making the neural drive spiky (see the blue (gray) bars in Figure 1); it compensates for the mechanical delay by shifting the neural commands early. (For instance, near the step at $t = 60$ centiseconds, n_t goes to zero before the target steps down and becomes nonzero again as soon as the new target begins.)[2] This leads to coarticulation that is approximately symmetric, in that the anticipatory and carry-over smoothing on opposite sides of a step in the target have the same magnitude and duration. The anticipatory move leads to the smallest mean-squared error given the muscle speed.

In speech production, coarticulation is sensitive to articulatory strength. It has been reported that stress affects both the magnitude and range of coarticulation. Many papers show that a stressed vowel is less affected by coarticulation effect then unstressed vowel [8, 10, 3]. Schwa, the unstressed vowel in English, shows strong coarticulation effect with neighboring vowels [5, 1]. We model the effect of stress on coarticulation by

[2]This behavior is much different from models where a dynamical system is driven by a signal that is proportional to the error. In such systems, the signal analogous to n_t would follow after the change in target and the mechanical response (analogous to f_t) would follow even later, after the n_t-analog.

changing the weights on articulatory targets (i.e. s and S), following [6, 7].

4. Examples

We test the simulation model using Mandarin conversational data with words carrying four consecutive falling tones, or tone 4. Tone 4 starts high and falls throughout the syllable. In the simple concatenation model that is the basis for defining coarticulation, a sequence of 4-4-4-4 then gives a saw-tooth pitch pattern. At the end of each syllable, the pitch needs to jump up to the beginning point of the next.

Examples of 4-4-4-4 tone patterns were extracted from a corpus of two hours of conversational speech evenly divided among four speakers. Two of the speakers did not produce any 4-4-4-4 patterns. The other two speakers produced two each. In the following, pei4-dian4 she4-ji4 *electronic design* and shi4-jie4 ri4-bao4 *World Journal* were from one speaker and the two instances of lu4-lu4-xu4-xu4 *continuously* were from another speaker. These four natural productions show different degrees of coarticulation effects, from a mild case in Figures 2, to a severe case in Figure 5, and intermediate stages in the other two productions. The f_0 data are plotted by circles on the y-axis as a function of time in centiseconds. Solid vertical lines mark syllable boundaries and dashed lines marked the boundaries between the initial consonant and the vowel.

The models are shown after 1000 iterations; there are no systematic changes beyond that point. We model the observed pitch movement in these four cases with the same pitch target on every syllable; it decreases from 220 Hz to 160 Hz over the duration of the syllable. Likewise, we always chose $s > S$ so that the target constrains the slope of f_0 more strongly than it constrains f_0 to any particular value. (Figures 2, 4, 5 have $s/S = 90$ and Figure 3 has $s/S = 23$.) These uniformities can be interpreted as the articulatory target or invariant of tone 4.

The strengths (s and S) on the tone target for each syllable are adjusted together to match the data. The strength values of the four syllables in each case are: Figure 2: $s = 0.9, 0.9, 0.45, 0.9$; Figure 3: $s = 0.9, 0.45, 0.45, 0.9$; Figure 4: $s = 0.9, 0, 0, 1.8$; Figure 5: $s = 0.09, 0.09, 0.9, 0.27$. In the models, multiple lines represent the results of different runs.[3] Linguistically, one can justify different values of s for different syllables on several bases: words can be semantically more or less important, they can be in a stressed position in a word, or the word can be under focus.

The strength values reflect segmental information as well. For example, in Figure 5, the third syllable has a strong weight and the other syllables are weak in the model representation. This weight assignment is consistent with the weakening effect that is found in the segmental channel of the spoken data, where consonant lenition occurs on the second and the fourth syllables.

Much of the behaviour of the model is driven by the size of the inter-vowel gaps (unvoiced regions). A small gap pretty much forces the f_0 on one syllable to be continuous with the next, while a large gap breaks that constraint and lets the S part of the target (R_a) take over.

5. Conclusion

In this paper, we present a simulation model *MALM* which learns patterns of neuron firing to drive the muscle to produce

[3]Note that this is intrinsically a probabilistic model, and it has intrinsic variability. The solution chosen is ultimately the result of the sequence of random steps chosen in the learning process.

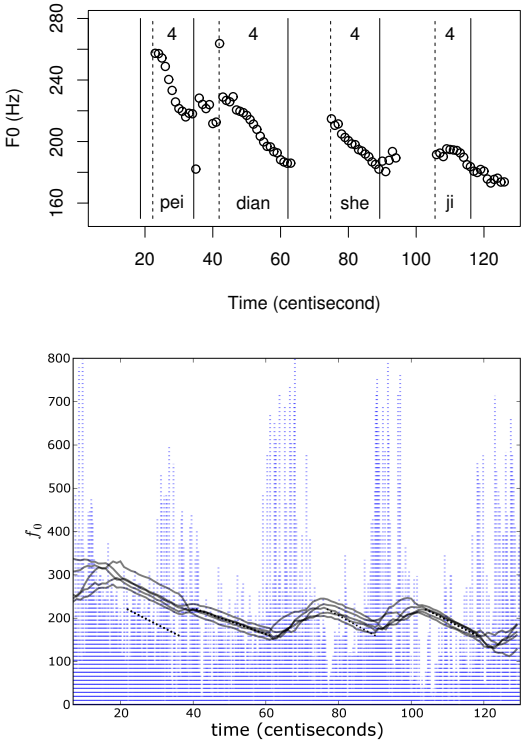

Figure 2: Sequence of falling tones (4-4-4-4) produced with very little coarticulation, and a MALM model that produces a similar result.

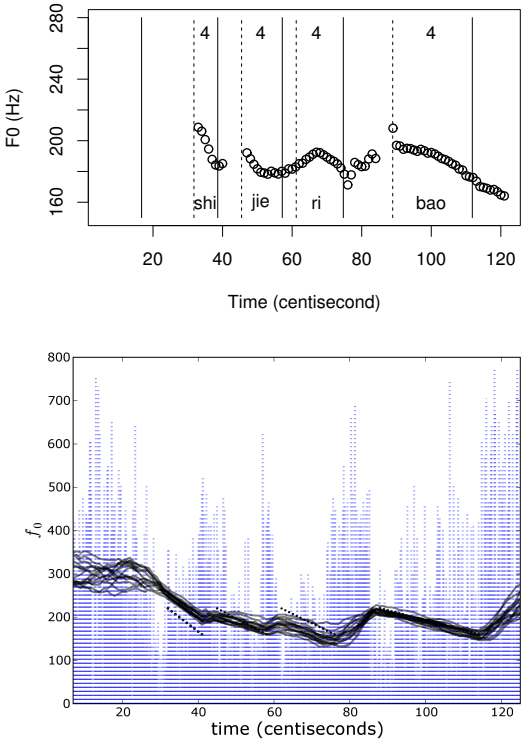

Figure 3: Sequence of falling tones produced with mild coarticulation, and a similar MALM model.

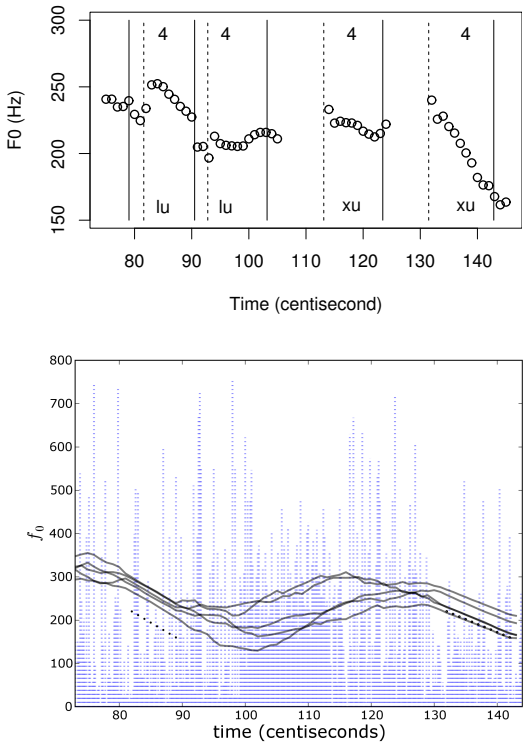

Figure 4: Sequence of falling tones produced with some coarticulation and a similar MALM model.

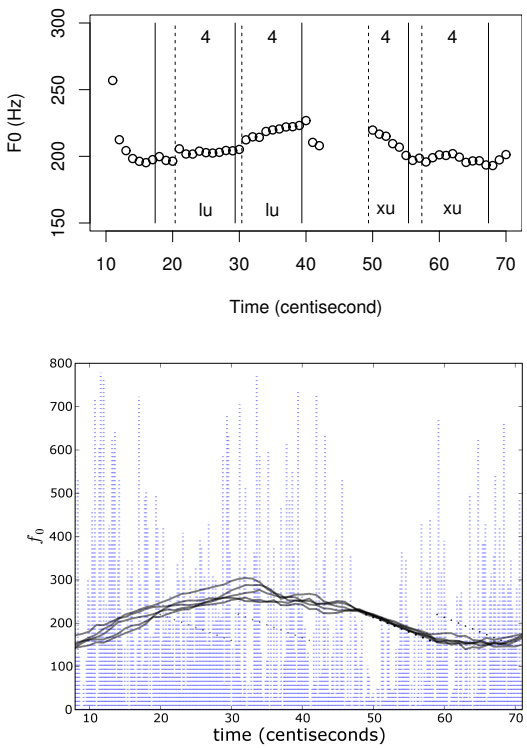

Figure 5: Sequence of falling tones produced with strong coarticulation and a related MALM model.

intended tone patterns. The model improves its motor skills through practice and evaluation. Eventually, it learns to compensate for the mechanical delay of the muscle response by shifting the neural commands early. This leads to coarticulation that is approximately symmetric. The anticipatory neural drive leads to the smallest error given the muscle speed.

We show a variety of surface intonation patterns from conversational Mandarin that were all generated from the same underlying 4-4-4-4 phonological sequence. The MALM model is able to plausibly reproduce the surface intonation patterns using linguistically plausible targets. We suggest that differences in surface intonation pattern from syllable to syllable occur because some syllables provide tighter constraints on the local intonation than others. This, combined with interactions of syllables with their neighbors can explain many different surface patterns.

6. References

[1] C. P. Browman and L. Goldstein. "Targetless" schwa: an articulatory analysis. In G. J. Docherty and D. R. Ladd, editors, *Papers in Laboratory Phonology*, volume II, pages 26–65. Cambridge University Press, 1992.

[2] H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage, editor, *The Production of Speech*, pages 39–55. Springer, New York, 1983.

[3] J. Harrington, J. Fletcher, and C. Roberts. Coarticulation and the accented/unaccented distinction: evidence from jaw movement data. *Journal of Phonetics*, 23(3):305–322, 1995.

[4] W. K. Hastings. Monte Carlo samplint methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[5] P. A. Keating. The window model of coarticulation: articulatory evidence. In J. Kingston and M. Beckman, editors, *Papers in Laboratory Phonology I. Between the Grammar and Physics of Speech*, pages 451–470. Cambridge University Press, Cambridge, 1990.

[6] Greg Kochanski and Chilin Shih. Prosody modeling with soft templates. *Speech Communication*, 39(3–4):311–352, 2003.

[7] Greg Kochanski, Chilin Shih, and Hongyan Jing. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication*, 41(4):625–645, 2003.

[8] H. Magen. The extent of vowel-to-vowel coarticulation in English. *J. Phonetics*, 25:187–205, 1997.

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21: 1087–1091, 1953.

[10] Joseph S. Perkell and Melanie L. Matthies. Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. *Journal of the Acoustical Society of America*, 91(5):2911–2925, 1992.

[11] Yi Xu and Xuejing Sun. Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. America*, 111(3):1399–1413, 2002.