Evidence for the Role of Gestural Overlap in Consonant Place Assimilation

Larissa H. Chen
Yale University and Haskins Laboratories, New Haven, CT, 06511, USA
E-mail: chen@haskins.yale.edu

ABSTRACT

Two hypotheses were tested: 1) Increasing the overlap between a coronal and a following non-coronal can result in the coronal gesture being heard as reduced in magnitude. 2) This relation does not hold for a non-coronal followed by a coronal. Computational models were used to generate and "listen" to several tokens of the test sequences "ba**d** **b**an" and "ba**b** **D**an," which varied in their combinations of reduction and overlap levels. The original generated articulatory configurations were compared to the recovered ones, and from this the relationship between overlap and reduction was examined, both as they were produced and as they were recovered Results support the hypotheses that overlap can sometimes masquerade as reduction, and that the effect is asymmetrical in a way that is consistent with a previously proposed hierarchy of constraints on place assimilation.

1.0 INTRODUCTION

Jun ([3]) has shown that there is a universal hierarchy of constraints that governs place assimilation in consonant clusters, such as the place and manner of trigger and target consonants. One such constraint states that, for any given language, the presence of velars as targets of assimilation implies the existence of labial targets, which implies further coronals ones. Based on this hierarchy, he hypothesizes that acoustically less salient sounds are targets of place assimilation more often than their more salient counterparts are. Steriade ([6]) postulates that speakers exploit their knowledge of what contexts are perceptually weak and therefore prone to assimilation, and reduce their articulations in just those weak environments. Jun showed that in at least some cases of perceived assimilation, speakers reduce their articulations of the first gesture in CC clusters. Using oral pressure experiments, he demonstrated for Korean that cases in which listeners perceived labials as assimilated to following velars in sequences such as /ipki/ corresponded to those in which the labial gesture was reduced in magnitude.

Jun argued that the reason for this gestural reduction is ease of articulation. But since the listener also needs to be able to understand what the speaker is saying, there are two conflicting groups of constraints: articulatorily motivated ones, which lead to gestural reduction and loss, and perceptually motivated ones, which lead to the preservation of all gestures. Jun provided an OT analysis to reconcile this conflict, thereby explaining the different cross-linguitic patterns of place assimilation, and presented his work as an argument against Browman & Goldstein ([1]). He claims that they overemphasized the role of gestural overlap in assimilation in C_1C_2 clusters, and that reduction, as described above, is the primary factor in consonant place assimilation.

However, if speakers are reducing their articulations only in environments in which they believe they will not be perceived by listeners (as in Steriade's proposal), they are clearly not succeeding. Listeners *do* notice when a gesture is significantly reduced in some context. If these events slipped by subtly, it is unlikely that they would be "phonologized" as categorical assimilations, or become part of the orthography. Under this account, speakers are trying to maintain the sound patterns of their language, which if possible would have led to them remaining unchanged over time. But this is not what has happened.

This paper attempts to provide an alternate explanation of these assimilations. Unlike in previous accounts, segmental mis-identification is not proposed to be the cause of diachronic sound change. Nor does the speaker deliberately change her articulation, either for ease of articulation or to improve perceptibility. Rather, speakers try to imitate the gestural patterns of the members of their speech communities. They engage in reduction when they hear others produce such reduction, but often this imitation target is in fact overlap that *sounds* like reduction due to the acoustical interaction of the overlapping gestures. A listener may imitate the first of these gestures as reduced, and such *produced* reduction is then further imitated, and gradually made more extreme. Over time these may become lexicalized or phonologized by speakers, resulting in changes in phonological representation. Although there are language-specific rules that govern permissible levels of overlap between CC or CV sequences, the speaker per se is not taking action to assist either the listener or herself; she is only trying to accommodate to the speech she hears.

2.0 BACKGROUND

Studies have shown that listeners are sensitive to the ways in which gestures overlap during speech. Yet the fact that listeners are somehow "aware" of overlap does not preclude these two physical events (reduction and overlap) from resulting in acoustics that are similar to one another. There is not a one-to-one mapping from the acoustic space

ISBN 1-876346-48-5 © 2003 UAB

to articulatory configurations, since one set of formants can result from an infinite number of area functions ([7]). Therefore, it is possible at times for one gesture or event to be mistaken for another by a listener.

Byrd ([2]) showed that there is an asymmetry between the effects of overlap on the assimilation of a coronal stop to a following labial one versus the assimilation of a labial to a following coronal. Using the same gestural model as the current study, she varied the amount of overlap between the stops in synthesized tokens of [b#b], [b#d], [d#b], and [d#d]. In a forced-choice discrimination task, listeners found coronals to assimilate to following labials more than the converse, with lesser amounts of overlap. Byrd cited two reasons: 1) the faster movement of the tongue tip is more easily hidden by the slower moving lips ([1]), and 2) C_2 coronals have a greater and more lasting effect on V_2 than do labials, so the V_2 formants of completely overlapped [bd]/[db] clusters sound more like canonical alveolars than they do bilabials.

This paper continues Byrd's research and aims to show that overlap can actually sound to a listener like gestural reduction. The model's behavior with respect to overlap alone is known, and is taken as a starting assumption. A gesture appearing (acoustically) not to reach its spatial target may be caused by gestural overlap rather than an actual reduction in magnitude. It is argued that this happens asymmetrically, such that the first consonant in highly overlapped coronal-labial clusters will sound more reduced than the first consonant in their labial-coronal counterparts. As Byrd demonstrated, the /d/ gesture does not have to be reduced to sound to the listener assimilated to the following /b/. Perhaps, then, the assimilation hierarchy can be better explained by asymmetries in articulator velocity and gestural timing than by some optimization scheme, conscious or not.

Despite the obvious problems with using a computational model rather than human subjects, it was necessary to do so in order to avoid the serious complications involved with using humans. The effects of native language on speech perception, as well as the naive speaker's lack of awareness of the relevant phenomena, would impede the subject's performance in a task involving overlap-reduction discrimination. A reasonable correspondence between natural speech articulations and those characterized by this model has already been demonstrated ([2]). Based on the results obtained here, it will be possible for future work to pursue this further using human speakers and listeners in imitation and similar tasks.

3.0 EXPERIMENT

The experiment is designed to test two hypotheses: 1) increasing the overlap between a coronal and a following non-coronal can result in the coronal gesture being heard as reduced in magnitude and 2) this relation does not hold for a non-coronal gesture followed by a coronal one.

3.1 Gestural Models

The "listener" model used is an algorithm that takes the first three formant frequencies of a speech signal as input, and determines the corresponding vocal tract area functions ([7]). It accounts for the fact that the vocal tract can only take a finite set of shapes, only a small fraction of which are used in speech, and even fewer that are dynamically plausible at any given point during speech.

The speech was created with a computational model of gestural structure called GEST ([1]), which generates gestural scores from an input phonetic transcription. A gestural score is a model of an utterance, which is comprised of individual gestures. Each tract variable set (velum, lips, tongue dorsum, tongue tip, and glottis) is specified for intervals of activation and for values of the dynamic parameters of target, frequency and damping which are fixed during these intervals. The values of these specifications were based on analysis of X-ray data ([1]). The gestural scores were input to the task dynamics model ([5]), and the resulting time functions of the model articulator degrees of freedom were input to the Haskins articulatory synthesizer ([4]).

3.2 Methods

Gestural scores ([1]) for the sequences *ba**d** **b**an* and *ba**b** **D**an* were generated with GEST, as described above. The additional stimuli were created by modifying these two initial scores, by adjusting in small increments the target value of C_1 and the relative timing between C_1 and C_2.

There were sixteen conditions for each of *ba**d** **b**an* and *ba**b** **D**an*. (All of the following was also done for the homorganic controls *ba**d** **D**an* and *ba**b** **b**an*, to test the validity of the model for this experiment. As the model behaved as expected for both controls, these results will not be reported here.) Four degrees of overlap between the medial /b/ and /d/ were crossed with four degrees of reduction in the first gesture, which was tongue tip raising for *ba**d** **b**an* and Lip Aperture (LA) for *ba**b** **D**an*.

Gestural reduction is a measure of magnitude, which for the TT refers to its distance from the alveolar ridge (*tongue tip constriction degree*, TTCD, measured in millimeters). The default value used by GEST is -3.5 mm, found empirically to be a typical TT height during a coronal stop in English ([1]). The TTCD values used were -3.5, -2.2, -0.9, and +0.4 mm. Thus at the lowest level of reduction, it is in fact a prototypical /d/. The TT is raised to where it not only makes contact with the palate, but also gets compressed, as it normally does during coronal stops in speech. At maximal reduction, the peak TT height is still *below* the point of contact; it never actually reaches the palate. To ensure that the two articulators were balanced in each word position, the same target values were used for LA in *ba**b** **D**an* as were used for TT in *ba**d** **b**an*, with the relevant measure being the distance between the upper and lower lips.

In addition to having a specification for magnitude, each gesture is also associated with a time, both internal (duration) and relative to surrounding gestures. A gesture's activation interval is the time for which the articulator is under active control. The total duration of each gesture was 110 ms, the default value. Overlap can be thought of as the amount of time for which the activation intervals for two gestures co-occur. In *bad ban*, for example, the smallest degree of overlap corresponds to the activation intervals of /d/ and /b/ being almost entirely separate, only co-occurring for 20 ms. At the greatest degree of overlap, they are simultaneous. The two intermediate levels used were 50 ms and 80 ms of overlap. Regardless of the amount of overlap, C_1's position in time remains constant, while C_2 gets shifted earlier.

The sound files output by the articulatory synthesizer for the 32 gestural scores were analyzed with a MATLAB function to extract the first three formant frequencies, estimated every 5 ms, during the VC transitions of the first word in the phrase (*bad* or *bab*). For each 5 ms formant vector, the recovery algorithm calculated a vocal tract area function, specified by 32 cross-sectional slices of the vocal tract, from the larynx to the lips. It was only necessary to examine the change in area for two sections: the lip region, which was simply the last section of the tube, and the TT region, which was the one whose area decreased towards zero over the successive frames during the VC transition into the coronal constriction. The articulator position values from the generated speech model output, rather than the formant frequencies, were used to calculate area functions for the "produced" tokens, used for comparison with the recovered ones.

For each token "recovered," the smallest apertures achieved by the tongue tip (distance between the TT and the palate) and the lips respectively during the VC transition were recorded, as were the smallest area functions recovered by the algorithm at those times. Presumably this would be around zero for a "normal," unreduced and minimally overlapped production, because there is complete, detectable closure. For other productions, however, a larger minimum aperture might be expected. Thus, when the articulator reaches full closure, or at least gets as close to it as it is going to get, the interesting part is what the listener perceives to be going on. Does it hear a stop closure, or something else? Note that because the algorithm is given only the formant frequencies of voiced VC transitions (i.e., before there is closure) as input, it will only recover positive values. It is important to keep this in mind when using a model intended for other purposes to study stop closures.

3.3 Results

For both *bad ban* and *bab Dan*, degree of reduction and degree of overlap were examined separately by plotting each against the minimum produced and recovered values, with the other held constant.

Fig. (1) contains two sample compilation graphs for the sequence *bad ban*. Fig. (1a) shows, for the case of maximal overlap, the minimum aperture values for both TT and LA that were *produced*, as well as the *perceived* apertures at those points in time. Fig. (1b) shows the effects of overlap alone on the minimum apertures, in the case of maximal reduction.

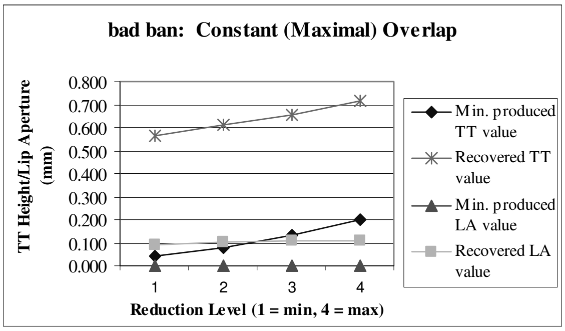

Figure (1a): Effect of Gestural Reduction on Minimum Produced and Recovered Constrictions for *bad ban*

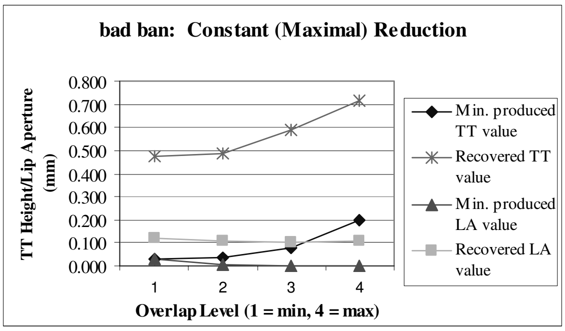

Figure (1b): Effect of Overlap on Minimum Produced and Recovered Articulator Constrictions for *bad ban*

There is a general upward trend, particularly in the recoveries, in both graphs. For constant overlap, this just means that decreasing the magnitude of a gesture has a substantial audible effect. Additionally, it can be seen in Fig. (1b) that for a given value of gestural reduction of C_1, as the amount of overlap is increased, the listener algorithm perceived less of a closure. These constant-overlap and constant-reduction graphs are quite similar, as were those for other levels of overlap and reduction; that is, both increasing reduction of $C_1$and increasing the overlap between C_1 and C_2 causes the recovery algorithm to recover less of a C_1 constriction.

The story is different for *bab Dan*. Again we want to see the effect of reduction of C_1 alone. For any given level of overlap, a change in the reduction level of C_1 (/b/) yields very little change in the minimum value recovered, as compared to the recovery of the coronal closure in the case of *bad ban*. Overlap alone in *bab Dan* has similarly little effect; increasing the amount of overlap between C_1 and C_2 has no appreciable effect on the recovery of the labial closure. Table (1) contains the minimum produced TT and LA values at the highest and lowest levels of overlap and reduction, as well as what the recovery perceived when those minima were reached. Note that the most relevant information here is for C_1 of each word: TT for *bad ban* and LA for *bab Dan*.

In summary, reduction of C_1 in coronal-labial clusters results in acoustics consistent with reduction. Overlap between C_1 and C_2 has a similar effect on these clusters, also making C_1 sound as if it were reduced. However, in

ISBN 1-876346-48-5 © 2003 UAB

labial-coronal clusters, a reduced C_1 sounds much more like a prototypical labial, and overlapping it with C_2 also has little influence on C_1's minimum recovered value.

			produced	**recovered**
TT	**min O, min R**	bad ban	0.000 mm	0.460
	min O, max R	bad ban	0.027	0.476
	max O, min R	bad ban	0.041	0.562
	max O, max R	bad ban	0.198	0.713
LA	**min O, min R**	bad ban	0.399	0.236
	min O, max R	bad ban	0.029	0.119
	max O, min R	bad ban	0.001	0.090
	max O, max R	bad ban	0.001	0.109
LA	**min O, min R**	bab Dan	0.001	0.001
	min O, max R	bab Dan	0.001	0.112
	max O, min R	bab Dan	0.001	0.090
	max O, max R	bab Dan	0.022	0.106
TT	**min O, min R**	bab Dan	1.273	1.091
	min O, max R	bab Dan	0.062	0.556
	max O, min R	bab Dan	0.041	0.562
	max O, max R	bab Dan	0.000	0.485

Table (1): Minimum produced and recovered values at high and low levels of overlap and reduction

The point is, the effects of overlap and reduction can be quite similar, but it is asymmetric; labials overlapped with preceding coronals greatly influence the coronals' perceptions, but the converse is not true. Overlap between the two gestures sounds like reduction in coronal-labial clusters, but not labial-coronal ones.

3.4 Discussion

The main expectation was that for coronal-labial clusters, increasing the amount of overlap would have a similar influence on the minimum recovered aperture as would increasing the amount of reduction, and that is to make it less constricted. The reason is that the second closure in a highly overlapped CC sequence will by definition begin before the first closure is complete. Therefore, the second stop will begin to have an influence on the V_1 formants before the first one has resulted in silence. The vocal tract will then sound like it did not close off completely for the first stop, just as when the first gesture is actually reduced.

As soon as either one of the gestures in a CC cluster reaches complete closure, it will presumably become impossible to tell what happens with the other gesture after that point. This is because the vocal tract has already been closed off, and there will be no formant pattern for the second closure to make its mark on. The minimum recovered value for this gesture will be its aperture at the point in time at which the first closure is achieved, and not necessarily the smallest constriction it actually reaches.

This means that the recovery is good at hearing what C_1 is doing. If at a given point in time it is not a complete closure, then this is what the recovery (and, presumably, a human listener) hears. It successfully hears the TT movement, whatever it is. And like the recovery algorithm, humans recognize when the vocal tract is not sealed off. Indeed, listeners are sensitive to the shape and state of the vocal tract over time. Human listeners attune to formant patterns (like the recovery model) as well as to other features of speech to determine the speaker's articulatory configuration. The lack of a one-to-one correspondence between acoustics and articulations means that we occasionally make mistakes. Sometimes a gesture is heard as reduced, when instead it was overlapped by the following gesture. Thus, support has also been found for Byrd's assertion ([2]) that "...assimilation need not be the result of operations or conditions altering the phonological representation but rather...can be the direct output of the representation itself" (p. 4). The place asymmetry comes in part from the fact that C_2 coronals have a greater effect on V_2 than do labials, and that at total overlap, the vowel offset formants are more similar to those of bilabials than coronals ([2]). The original motivation for hypothesizing it as such of course stems largely from the frequently-observed asymmetries in place assimilation ([3]).

4.0 FINAL THOUGHTS

The two models employed were created independently of both each other and of the current study, allowing the hypotheses to be tested independently of the design of either. If speech generated by a model in which overlap and reduction are orthogonal to one another can be used to show that a listener who is not specifically attending to either will recover overlap in this speech as reduction, then this provides some objective evidence confirming these hypotheses. Using such models may be an imperfect method to our means, but it is far preferable to using a design deliberately intended with that result in mind. In that case, it could not be said that the findings were merely an artifact of the physical nature of speech.

REFERENCES

[1] C. Browman and L. Goldstein, "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Kingston and Beckman, Eds. pp. 341-376. Cambridge Univ. Press, 1990.

[2] D. Byrd, "Perception of assimilation in consonant clusters: a gestural model," *Phonetica*, 49, pp. 1-24, 1992.

[3] J. Jun, "Perceptual and articulatory factors in place assimilation: an Optimality Theoretic approach," UCLA Dissertation, 1995.

[4] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *JASA* 70, pp. 321-328, 1981.

[5] E. Saltzman, "Intergestural timing in speech production: data and modeling," *Proceedings of the XIIIth ICPhS* 2, pp. 84-91, 1995.

[6] D. Steriade, "Directional asymmetries in place assimilation: a perceptual account," in *The Role of Speech Perception in Phonology*, E. Hume, K. Johnson, Eds., pp. 219-250. San Diego: Academic Press, 2001.

[7] H. C. Yehia, "A study on the speech acoustic-to-articulatory mapping using morphological constraints," Nayoga Univ. Dissertation, 1997.