 [DOI: 10.1121/1.3213452] Published Online 19 October 2009

Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation

Vikram Ramanarayanan and Erik Bresch
Department of Electrical Engineering, Speech Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, California 90089
vramanar@usc.edu, bresch@usc.edu

Dani Byrd and Louis Goldstein
Department of Linguistics, University of Southern California, Los Angeles, California 90089
dbyrd@usc.edu, louisgol@usc.edu

Shrikanth S. Narayanan
Department of Electrical Engineering and Department of Linguistics, University of Southern California, Los Angeles, California 90089
shri@sipi.usc.edu

Abstract: It is hypothesized that pauses at major syntactic boundaries (i.e., grammatical pauses), but not ungrammatical (e.g., word search) pauses, are planned by a high-level cognitive mechanism that also controls the rate of articulation around these junctures. Real-time magnetic resonance imaging is used to analyze articulation at and around grammatical and ungrammatical pauses in spontaneous speech. Measures quantifying the speed of articulators were developed and applied during these pauses as well as during their immediate neighborhoods. Grammatical pauses were found to have an appreciable drop in speed at the pause itself as compared to ungrammatical pauses, which is consistent with our hypothesis that grammatical pauses are indeed choreographed by a central cognitive planner.

PACS numbers: 43.70.Fq, 43.72.Ar [DO]
Date Received: February 17, 2009 **Date Accepted:** July 30, 2009

1. Pauses during spontaneous speech

Pausing in natural speech can be considered from a listener perspective—how do pauses aid or impair speech understanding—or from a speaker perspective—how do pauses reflect the speech planning process, either operating well or encountering difficulties. In this paper, we use real-time magnetic resonance imaging (MRI) to provide a noninvasive view of the entire length of the moving vocal tract and to examine pauses from a speech production perspective. Pauses can be broadly categorized into planned or grammatical pauses and unplanned or ungrammatical pauses. (For our purposes, we make no distinction between planned and grammatical, and likewise unplanned and ungrammatical pauses.) Grammatical pauses generally occur at the boundary of a clause, presumably due to the need to parse and plan the sentence. Ungrammatical pauses can indicate a breakdown in composing the speech stream and occur at inappropriate locations as the planning, production, and/or lexical access process is disrupted (see O'Shaughnessy, 1992, Rochester, 1973).

The framework of articulatory phonology (Browman and Goldstein, 1992, 1995) in conjunction with the prosodic-gesture model (Byrd and Saltzman, 2003) of phrase boundaries offers one approach for considering the nature of grammatical and ungrammatical pauses in articulation. In this framework, the act of speaking is decomposable into a.u. of vocal tract action—gestures—that can be defined as an equivalence class of goal-directed movements, such as those by a set of articulators in the vocal tract (see the task dynamics model, Saltzman

and Munhall, 1989). Byrd and Saltzman (2003) viewed phrase junctures as phonologically planned intervals of controlled local slowing of speech timing around a phrase edge, with the articulatory slowly increases as the boundary approaches and the speech stream resumes speed as the boundary recedes (i.e., immediately postboundary). This "clock" slowing, at its extreme, can be understood to result in a pause, as the clock controlling articulation slows to a near-stop and then speeds up again as the postpause interval is initiated. In contrast, ungrammatical pauses (which may be filled or unfilled depending on the state of voicing) abruptly interrupt the execution of the planned speech stream interfering with the vocal tract articulators reaching their targets. Under this approach, a grammatical pause, then, is viewed as a planned event under cognitive control with explicit consequences for the spatiotemporal behavior of the articulators over an interval; consequences that are distinct from ungrammatical pauses that abruptly perturb articulation. In this paper, we will examine direct articulatory evidence for this hypothesis.

Although pauses can contribute information about speech planning, few joint acoustic and articulatory studies of pausing behavior have been carried out, one reason being the difficulty of acquiring data on vocal tract movement during running speech. Recent progress in real-time MRI (Narayanan *et al.*, 2004) allows for a more comprehensive investigation of pauses in speech than does study of the acoustic signal alone. Since the technique allows for a complete view of the moving vocal tract, providing synchronized audio in conjunction, it is possible to examine the supraglottal articulators during not only the spoken portion but also the silent portions of the speech stream.

2. Data acquisition and preparation

The data we examined comprise spontaneous speech utterances and the corresponding time-synchronized movies of the moving vocal tract, elicited in response to queries from the experimenter. Seven healthy native speakers of American English were asked to answer simple questions on general topics such as "what music do you listen to…," "tell me more about your favorite cuisine…," etc.) while lying inside a MRI scanner. For each of the stimulus questions, time-synchronized audio responses and MRI videos of speech articulation were recorded for 30 s. Further details regarding the recording/imaging setup can be found in Narayanan *et al.*, 2004; Bresch *et al.*, 2006. Midsagittal real-time MR images of the vocal tract were acquired with a MR pulse repetition time of TR=6.5 ms on a GE Signa 1.5 T scanner with a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3 mm. A sliding window reconstruction at a rate of 22.44 frames/s was employed. The field of view was adjusted depending on the subject's head size, so that images covered an area of 18.4×18.4 cm^2 at a resolution of 68×68 pixels.

For the manual annotation of the audio waveform for this experiment, a grammatical pause was defined to be a silent or filled pause that occurred between overt syntactic constituents (including sentence end). Examples include pauses at (1) clause boundaries such as relative clause boundaries, (2) subject-verb or verb-object boundaries, and (3) prepositional phrases offset from another constituent. Any pause other than the above, i.e., generally those occurring within a clause, was marked as an ungrammatical pause. Such pauses are atypical in this natural speech and do not mark the juncture between obvious syntactic or semantic word groups in the sentence; they do not appear to encode linguistic information. For each speaker's utterances, grammatical and ungrammatical pauses were manually annotated by the first author according to this definition and verified by a linguist for accuracy.

3. Analyses

In order to examine the articulatory characteristics at and around pauses, the extraction of a "gradient energy" measure that captures the speed of articulatory motion (of all articulators) from image sequences is employed. In order to study the time evolution of vocal tract shaping, for each set of image sequences, the air-tissue boundary of the articulatory structures needs to

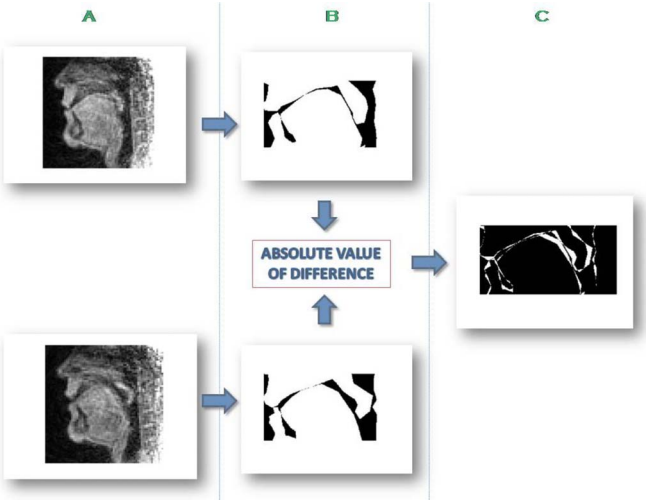

Fig. 1. (Color online) An illustration of the gradient energy calculation process: first, contour outlines are obtained from the MRI images in panel A and are then converted to binary masks (panel B); these are then used to compute the "gradient" images (panel C), the energy of which is then calculated by a simple addition operation (of all white pixels).

be clearly delineated. This contour tracing process is time consuming and tedious when carried out by a human, so an algorithm using Fourier region segmentation to automatically carry out the task was used (see Bresch and Narayanan, 2009).

In order to observe articulatory effects of pausing behavior more comprehensively, it is important to study articulator dynamics not only in the pause frames but also in the interval preceding and following the pause, particularly since models such as the prosodic-gesture model (Byrd & Saltzman, 2003) predict spatiotemporal effects during neighboring intervals. Since most appreciable effects, including construction of a rough plan for the utterance (Kochanski *et al.*, 2003), occur in a time window of 500 ms before and after the pause, neighborhoods of that order were analyzed for global range of movements of articulators. Since in our experimental setup, the frame rate is about 22.44 frames/s; this approximately translates to neighborhoods consisting of about 12 frames. Thus, for analysis, although the length (in number of frames) of each pause was variable, the analysis neighborhoods before and after the pause were of fixed lengths.

Once the contour outlines have been extracted from the MR images, they are used to create binary mask images, with all pixels enclosed by these contour outlines assigned a normalized value of 1, and the rest, 0, such that the midsagittal section of the vocal tract appears white on a black background (see Fig. 1). A gradient energy measure was calculated for every pair of *contiguous mask images* in a pause/neighborhood frame sequence, by subtracting them, taking the absolute value of the difference, and computing the "pixel energy" of the result (by finding the number of pixels of value "1"). The overall gradient energy value for a pause/neighborhood is then computed by averaging over all gradient energies obtained during the pause/neighborhood period. This is done to obtain an entropy measure that can capture variability in articulator movement and thus give an estimate of the speed of articulatory motion during such periods, which this measure does well on a global level.[1] In a similar manner, one can compute delta gradient measures that will capture the acceleration of the articulators during pauses/neighborhoods.

A two-factor parametric analysis of variance (ANOVA) was conducted on the dependent variable of gradient energy with data pooled across speakers and with the factors: site (levels: prepause, pause, and postpause) and grammaticality (levels: grammatical and ungrammatical). It should be noted that since the number of occurrences of grammatical and ungrammatical pauses (especially the latter[2]) were too small for a repeated measures ANOVA, analyses

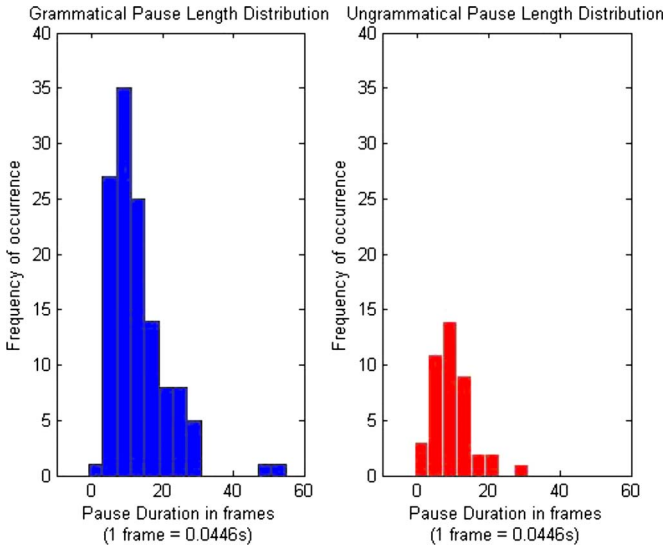

Fig. 2. (Color online) Pause length distributions for grammatical and ungrammatical pauses.

were carried out with data pooled across speakers.[3]

4. Results

Histograms of grammatical and ungrammatical pause durations across all speakers were plotted, and while these pauses cannot be reliably separated based on duration values alone, a statistical t-test indicated that grammatical pauses tended to be significantly longer on average ($p \leq 0.01$)—the mean and standard deviation of pause durations were found to be 13.62 frames and 8.2 frames, respectively, for grammatical, and 9.76 frames and 5.8 frames, respectively, for ungrammatical pauses (1 frame = 0.0446 s) (see Fig. 2).

The time-normalized average gradient frame energy for each pause and for the neighborhoods before and after it, pooled across all 7 speakers, is plotted as a bar graph in Fig. 3. Corresponding time-normalized average local phone rates are also plotted to the right of each of these graphs. The derived measure of mean gradient frame energy captures articulator speeds well and is a good indicator of the local phone rate (assuming that articulator speeds directly inform the local phone rate to a certain extent).

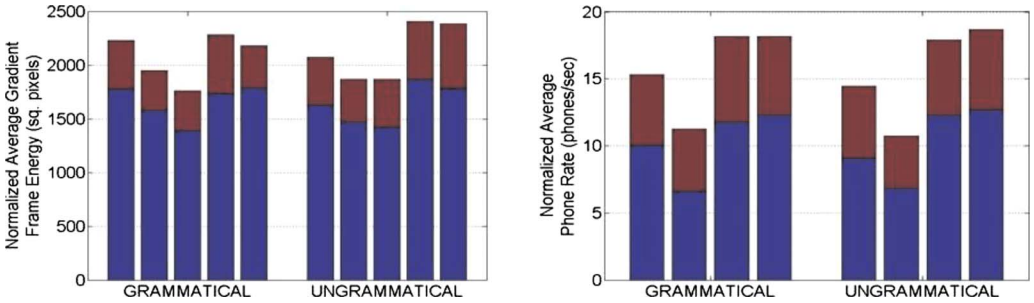

Fig. 3. (Color online) Time-normalized average *gradient frame energies* (in squared pixels) of grammatical and ungrammatical pauses and their neighborhoods pooled across all seven speakers (standard deviation bars are plotted on top of each energy bar in a lighter color). Corresponding average *local phone rates* (phones/s) are also shown to the right of the gradient frame energy panel. Each panel consists of two pause groups on the *x*-axis: (1) grammatical and (2) ungrammatical. Group 1 consists of, in order, bars for two neighborhoods immediately before the grammatical pause (~250 ms), followed by one bar for the pause itself (*not shown for phone rate graph*), followed by two bars for the neighborhoods following the pause (~250 ms); this set of five bars is followed by a parallel sequence of five bars for the ungrammatical pauses (Group 2).

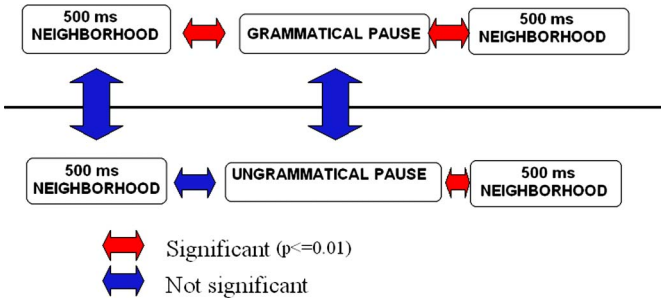

Fig. 4. (Color online) A schematic depicting the levels (grammatical and ungrammatical) and sites (prepause, pause, and postpause) at which the ANOVA statistical analyses were performed.

In order to examine the effects of speech planning on the structure of pauses, we have to examine how the gradient frame energies vary moving into and out of the pause. Figure 4 schematically summarizes the statistical comparisons (double-headed arrows) performed on the data. Significant differences ($p \leq 0.01$) were found between the means of the gradient energies in the (prepausal) neighborhood before *grammatical* pauses and those of the pauses themselves. That is, the gradient energy means of the grammatical pauses themselves were *significantly lower* than these prepausal neighborhood gradient energy means. In contrast, *ungrammatical* pauses displayed *no significant differences* between the means of the prepausal and pausal gradient energies. However, there was a significant increase in the gradient energy for the neighborhood immediately following both grammatical and ungrammatical pauses, which was often slightly higher than the prepausal energy value. There were no significant differences in the means of the grammatical and ungrammatical (i) pause gradient energies or (ii) prepausal neighborhood gradient energies. However, ungrammatical pauses showed a significantly higher variation in the values of gradient energies compared to the grammatical case, especially during and after the pause (see Table 1), which is expected, since such a pause would hypothetically serve to interrupt the flow of speech (irrespective of the speech rate) and hence would have a much higher gradient energy variance compared to grammatical pauses.

For both grammatical and ungrammatical pauses, there was no trend found that distinguished filled from unfilled pauses, as the gradient energies of these cases can be highly context dependent. In some cases, the filled pause gradient energies for some speakers were much higher than their unfilled counterparts, while the opposite effect was found for other speakers. Furthermore, there was no observed trend of gradient frame energy variation *within* a pause, be it grammatical or ungrammatical, filled or unfilled, suggesting that instantaneous values of these gradient energies may be context dependent.

The results obtained suggest that grammatical pauses are part of a more globally choreographed plan of articulatory movement, since at the pause, the speed of the articulators drops (as indicated by the reduction in mean gradient energy), which, finally, toward the end, increases to around the level where it was at the start of the pause. Also, the results suggest that ungrammatical pauses are essentially unplanned, with the articulator speed dropping slightly (but not significantly) early into the pause, following which there is a sudden jump in the gradient en-

Table 1. Standard deviation (in squared pixels) of the gradient energies for grammatical and ungrammatical pauses and their neighborhoods pooled across all speakers (here the two 250 ms neighborhoods before and after the pause are pooled together to get one 500 ms neighborhood before and after).

Grammatical pauses			Ungrammatical pauses		
500 ms before	Pause	500 ms after	500 ms before	Pause	500 ms after
366.54	369.25	423.57	374.95	445.37	538.11

ergy. In addition, there is a large variance associated with the gradient energy values in this case (Table 1). Such pauses in spontaneous speech, which occur without the linguistic structuring of the speaker, are characterized by a sudden increase in articulator speed on a global level when the speaker eventually succeeds in lexical access or planning.

5. Conclusions

In this paper, our principal hypothesis that grammatical pauses are a result of a higher-level cognitive plan of articulatory movement, while ungrammatical ones are not, has been validated via direct observation of articulatory behavior. Measures that help distinguish between the two in the articulatory domain were developed.

It has long been recognized that pauses are relevant to cognitive processing and are related to effect, style, and lexical and grammatical structure (e.g., Lehiste, 1970; Rochester, 1973; Kutik *et al.*, 1983; Zellner, 1994). Direct observation of articulation along the entire vocal tract offers an important new source of data for investigation of speech planning, since it allows a view of how the speech flow is altered in either a cognitively planned way or interrupted by a perturbation when normal planning fails. It can also inform as to how much time it takes to "recover" from the effect of a sudden unplanned pause that perturbs the linguistic structural integrity of the utterance.

Acknowledgments

The authors thank Ed Holsinger and Krishna Nayak. Work described in this paper was supported by NIH Grant Nos. DC007124 and DC03172, the USC Imaging Sciences Center, and the USC Center for High Performance Computing and Communications (HPCC).

References and links

[1]By "global," we mean that the gradient energy measure is an indicator of the net motion of all vocal tract articulators.

[2]Some utterances of some speakers were found to not contain any ungrammatical pauses.

[3]Also, due to the same reason, we cannot directly assume that the data are parametric, although the results obtained using a nonparametric data distribution assumption are similar to those obtained using parametric analysis, and so only the latter results are reported.

Bresch, E. and Narayanan, S. (**2009**). "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Trans. Med. Imaging **28**, 323–338.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. (**2006**). "Synchronized and noise-robust audio recordings during realtime MRI scans," J. Acoust. Soc. Am. **120**, 1791–1794.

Browman, C. P. and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155–180.

Browman, C. P. and Goldstein, L. (**1995**). "Dynamics and articulatory phonology," in *Mind as Motion: Dynamics, Behavior, and Cognition*, edited by R. Port and T. van Gelder (MIT, Cambridge, MA), pp. 175–193.

Byrd, D. and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," J. Phonetics **31**, 149–180.

Kochanski, G., Shih, C., and Jing, H. (**2003**). "Quantitative measurement of prosodic strength in Mandarin," Speech Commun. **41**, 625–645.

Kutik, E. J., Cooper, W. E., and Boyce, S. (**1983**). "Declination of fundamental frequency in speakers' production of parenthetical and main clauses," J. Acoust. Soc. Am. **73**, 1731–1738.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge, MA)

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776.

O'Shaughnessy, D. (**1992**). "Recognition of hesitations in spontaneous speech," in Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, pp. 521–524.

Rochester, S. R. (**1973**). "The significance of pauses in spontaneous speech," J. Psycholinguist. Res. **2**, 51–82.

Saltzman, E. L. and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological Psychol. **1**, 333–382.

Zellner, B. (**1994**). "Pauses and the temporal structure of speech," in *Fundamentals of Speech Synthesis and Speech Recognition*, edited by E. Keller (Wiley, Chichester), pp. 41–62.