THE ACOUSTICS AND KINEMATICS OF REGULARLY TIMED SPEECH: A DATABASE AND METHOD FOR THE STUDY OF THE P-CENTER PROBLEM

Aniruddh D. Patel[*], Anders Löfqvist[†], and Walter Naito[‡]
[*]*The Neurosciences Institute, San Diego, CA, USA,* [†]*Haskins Laboratories, New Haven, CT, USA,*
University of Tokyo Medical School, Japan

ABSTRACT

The physical cues underlying the perception of temporal intervals in speech have yet to be firmly established. It has long been known that when speakers are asked to produce "evenly-timed" sequences of alternating syllables (e.g. ba-la-ba-la...) they introduce systematic deviations from onset-to-onset isochrony, showing that the perception of temporal intervals in speech is based on some cue other than syllable onsets. Identification of this cue would aid in measuring speech timing in a perceptually meaningful manner, benefiting both psycholinguistic and synthesis research. This study gathered a diverse set of regularly-timed syllable sequences from six speakers, and tested four acoustic and kinematic candidates for temporal iscohrony using a simple statistical method. While none of the cues showed isochrony, the database of speech and kinematic data (available from http://www.nsi.edu/users/patel) and method of analysis can be used by other researchers to evaluate proposed candidates for timing cues ("P- centers") in speech.

1. INTRODUCTION

1.1. General Background

A key problem in studying suprasegmental timing patterns is deciding how to measure duration in a way which reflects the perception of timing by listeners and the timing strategies of speakers. Measuring intervals between syllable onsets is unsatisfactory because abundant evidence shows that speakers and listeners, when asked to attend to the timing of syllables, are not concerned with the timing of onsets. Rather, they seem to care about a point within the syllable that *perceptually* feels like the syllables "moment of occurrence" [1, 2. 3, 4]. This point has been called the "P-center" in psychological studies of language and its physical correlate has yet to be firmly established.

The P-center phenomenon can be demonstrated by asking speakers to produce regularly-timed sequences of alternating syllables (e.g. *bad-nad-bad-nad...* or *bad-sad-bad-sad...* spoken "like a metronome"). Speakers will invariably introduce systematic deviations from onset-to-onset isochrony in order to achieve perceptual regularity [5]. In the sequences just mentioned, "nad" and "sad" would be timed so that their acoustic onset happened "too early" from the standpoint of onset-to-onset isochrony. The amount of anisochrony depends on the particular onset consonants of the syllables involved. Longer consonants (e.g. /s/) are associated with greater anisochrony than shorter consonants (e.g. /n/). Listening experiments show that these adjustments by speakers produce a sense of perceived regularity in other listeners [6]. Furthermore, the P-center effect is not specific to English: Hoequist [7] has documented the effect in Spanish and Japanese.

1.2. Previous theories

Howell [8, 9, 10] suggested that the amplitude envelope of a signal was the key factor in determining P-center location, and proposed a syllabic "center of gravity model" on this basis. More recent models by Pompino-Marschall [11] and Harsin [12] are based on loudness functions or modulation envelopes within different auditory critical bands. All of these models require measurement of energy from the entire syllable before a P-center can be computed. This contradicts subjective experience, in which a syllables "moment of occurrence" or "beat" is felt before the entire syllable is heard or spoken. To overcome this problem, Scott [4] proposed that the P-center corresponds to a point of rapid energy rise in within the first formant frequency band. Scott's proposal is based on acoustic analysis of only two syllables, however (the spoken digits "one" and "two"), and needs a more thorough test. The above proposals may be contrasted with a suggestion by Tuller and Fowler [13] that the perceived regularity is not necessarily in the acoustic signal, but in the kinematic signal which underlies it. Fowler [14] suggests that the regularly-timed event is the onset of the vowel gesture. Defining this point in kinematic data is not a straightforward matter, however. So far, no clear articulatory correlate of the P-center has been found [15].

1.3. Goals of the current study

The current study sought to evaluate two acoustic and two kinematic candidates for the P-center, and to develop a database and method for evaluating future candidates. With this goal, a diverse set of regularly-timed syllable sequences were collected from six different speakers, along with kinematic data from three of these speakers.

1.3.1. Acoustic hypotheses. Points of maximally rapid amplitude rise in both the first formant (F1) and the fundamental frequency (F0) range were examined as possible acoustic cues for the P-center. Rapid onsets are very salient to the auditory system, and both F0 and F1 have the ability to dominate the activity of neurons along large sections of the basilar membrane [16]. This is due to the fact that neurons which are tuned to higher frequencies than F0 or F1 retain sensitivity to these lower frequencies, and will often phase-lock their firing patterns to F0 and F1 due to the substantial energy of these low frequencies. This results in a *temporal* representation of F0 or F1 in the firing pattern of neurons across a broad range of auditory nerve fibers. This spatially distributed temporal

coding can offer a good deal of noise-resistance [17], an advantageous feature for a signal used in perceptual timing processes.

1.3.2. Kinematic hypotheses. Velocity maxima were measured in the kinematics of primary articulators (e.g. the lips for /b/ and the tongue tip for /l/), and of the jaw. Velocity was chosen because muscle spindles are known to have primary endings with velocity sensitivity [18], suggesting that rate of muscle stretch can be detected as a proprioceptive signal. If the basis of regularly-timed speech is kinematic rather than acoustic, points of maximum articulator velocity are a possible candidate for temporal control.

2. METHODS

2.1. Database: subjects, procedures, syllables

Six adult native English speakers (three females and three males) participated in the study. Each speaker produced the syllable /ba/ in alternation with other specified syllables (e.g., "ba-la-ba-la . . ." or "sa-ba-sa-ba . . ." etc.) [SOUND 0613.WAV]. Speakers were instructed to produce syllables as evenly spaced in time as possible ("like a metronome"). To ensure that the subjects understood the instructions, a brief practice trial was included in which subjects produced the syllable /ba/ in time with a metronome (metronome rate = 2 beats/sec).

Fifty-two syllable sequences were recorded. For each sequence, the speaker was informed of the syllable to produce in alternation with /ba/, and then given a signal to begin. The subject was told not to count syllables internally (as this might distract attention from the task). The experimenter kept track of the number of syllables and signaled the speaker to stop after eight pairs of syllables had been produced. A sequence was repeated if there was a hesitation, breath pause, or misarticulation during production. On half the trials, the subject was instructed to begin the sequence with "ba," and on the other half, with the other syllable.

The syllables chosen for alternation with /ba/ are referred to as "target syllables" (note that /ba/ can serve as its own target syllable, i.e. a sequence consisting of "ba-ba-ba-ba . . ."). Thirteen target syllables were chosen for production in alternation with /ba/. Nine were chosen to span a range of onset consonant classes: affricate, aspirate, fricative, glide, liquid, nasal, and stop (/cha/, /ha/, /sa/, /ya/, /la/, /ma/, /ba/, /pa/, /ta/). The remaining syllables were chosen to include a syllable with a coda consonant (/lad/), a complex onset (/spa/), an unstressed affix (/dela/), and a vowel maximally different from /a/ (/li/). Each speaker produced four sequences for each target syllable.

Acoustic data were collected from all subjects, and kinematic data from three subjects. Details of acoustic and kinematic data acquisition are given below. The database is available from the first author's URL (see abstract).

2.2. Method for testing P-center candidates

This section describes general principles of data extraction and P-center testing which apply to both kinematic and acoustic data.

From each spoken sequence of 16 syllables, a continuous string of 11 syllables was selected for analysis, subject to the following constraints: the string had to begin and end with /ba/ and could not include either the first two or last two syllables of the 16-syllable sequence. These constraints were imposed in order to give all sequences a uniform alternating structure and to avoid any timing effects peculiar to the beginning or end of spoken utterances. The resulting sequences always had the form:

Ba-tg-Ba-tg-Ba-tg-Ba-tg-Ba-tg-Ba

where "tg" indicates the target syllable. For three of the subjects, each acoustic syllable sequence was accompanied by several time-aligned kinematic traces, each containing movement data for a different articulator (e.g., lower lip, tongue tip).

For a given subject and target syllable, timing data were extracted and P-center candidates evaluated in the following way. First, a single sequence was examined, and a particular time in each syllable corresponding to an event of interest was identified (e.g., the point of maximum amplitude envelope slope in the first formant rage). This event was identified in both the /ba/ syllables and the target syllables, yielding eleven time points. The temporal intervals between successive time points were then computed via simple subtraction, yielding ten intervals per sequence (measured in milliseconds). This procedure was repeated with the other three sequences containing the same target syllable, resulting in four sets of ten intervals for each target syllable. These intervals were pooled[1] and then divided into two populations: "ba-target" intervals (between /ba/ and the target syllable), and "target-ba" intervals (between the target syllable and /ba/). A statistical comparison of these two populations using a t-test for the difference between two means [19] allows one to test the null hypothesis that the event of interest is produced in an isochronous fashion, i.e. to assign a probability that the two sets of intervals are drawn from the same underlying population.

2.3 Acoustic measurements

Speech was recorded in a quiet room using either a miniature condenser microphone (Audio-Technica AT803b, at Harvard) or a Sennheiser MKH 816T microphone (at Haskins Labs). The acoustic signal was anti-alias filtered at 4 kHz and digitized at 10 kHz. All acoustic measurements and manipulations described below were performed using SIGNAL (Engineering Design, Belmont, MA) on a personal computer.

The point of onset of each syllable was recorded via waveform or spectrogram-based measurements. For /ba/, syllable onset was taken to be the time of the release burst. For target syllables, onset was determined via waveform-based measures, using a gating algorithm to detect regions where absolute waveform amplitude exceeded the noise floor (typically 1/4th - 1/7th of signal RMS). These gate limits were checked by eye for each syllable; erroneous time markings were corrected by hand.

To study the amplitude envelope slope in the first formant range, the mean time-varying spectral amplitude of a frequency

band extending from 390 Hz to 1015 Hz was derived from wide-band spectrograms (64 point FFT w/Hanning window apx. every 2.5 msec). For the vowel /i/, which has a lower first formant, a frequency range of 234 - 859 Hz was used. This amplitude curve was smoothed twice with a 30 msec window and then differentiated. The resulting curve was smoothed once again with a 20 msec window. Times of positive peaks in the derivative were recorded.

To study the amplitude envelope slope in the fundamental frequency range, the mean time-varying spectral amplitude of a frequency band matched to each speaker's F0 was derived from a "medium-band" spectrogram (256 point FFT w/Hanning window apx. every 5 msec). For the female speakers, the frequency band extended from 98-215 Hz, or from 136-254 Hz. For the male speakers, the band extended from 58-176 Hz. The choice of a narrower-band spectrogram was made so that energy from the fundamental frequency could be distinguished from the energy of F1. Smoothing and differentiation procedures were the same as those for F1 amplitude curves.

2.4. Kinematic measurements

Kinematic data were collected for three of the subjects using an electromagnetic midsagittal articulometer (EMMA) system [20], which allows measurement of horizontal and vertical displacement for selected articulators in a standardized coordinate system. Movements were measured from the upper lip, lower lip, jaw, tongue tip, tongue blade, tongue body, and tongue rear. The coordinate frame (units = centimeters) was centered at the upper incisors. The "tongue tip" transducer was not truly on the tip of the tongue (as this would interfere with articulation), but roughly 1/2 cm behind the tip.

Tangential velocity curves for primary articulators and for the jaw were derived for each syllable sequence. Basic horizontal and vertical displacement data were used to compute tangential velocity curves via differentiation and vector addition (velocity = $\sqrt{(dx/dt^2 + dy/dt^2)}$) .

3. RESULTS

3.1. Intervals between syllable onsets

Figure 1 shows the pattern of syllable-onset timing in the different sequences. The ratio of "ba-target" interval duration is expressed as a fraction of the duration between alternate /ba/ syllables. (Perfect isochrony would yield a ratio of 50 %). It can be seen that for all target syllables except /ba/ itself, syllable onset timing is anisochronous. A t-test comparing "ba-target" and "target-ba" onset-to-onset intervals was computed for each of the thirteen different target syllables, on a subject-by-subject basis. For every subject, these intervals differed significantly when /ba/ was produced in alternation with another syllable ($p < 0.05$), and did not differ significantly when /ba/ served as its own target. Furthermore, for every syllable and subject, temporal intervals between *alternate* /ba/ syllable onsets and *alternate* target syllables do not differ significantly in duration ($p > 0.05$, t-test for unequal sample sizes). This shows that deviations from onset-to-onset isochrony is in these syllable sequences are systematic and stable, suggesting that speakers have a clearly defined focus in their timing strategy. (Note that in Figure 1, the "onset" of /spa/ was measured from the /p/, in order to contrast it with /pa/).

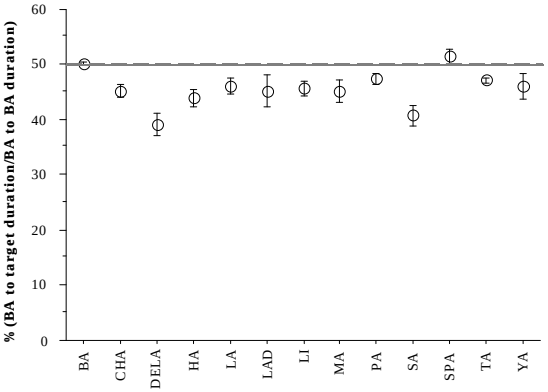

Figure 1. The amount of temporal deviation from isochrony (represented by 50%) for intervals between **syllable onsets** in the different syllable sequences, across six speakers. Error bars are 95% confidence intervals.

3.2. Intervals between point of first-formant amplitude slope maxima, and fundamental frequency slope maxima

Temporal intervals between slope maxima in first-formant (F1) amplitude were measured for 5 target syllables: /ha/, /la/, /pa/, /sa/, and /ya/. These five were chosen as a subgroup because their initial consonants are produced in quite different ways and have very distinct acoustic/phonetic properties (aspirate, liquid, stop, fricative, and glide). They provide a good first testing ground for any proposed acoustic correlate of regular timing. The same syllables were used to examine temporal intervals between fundamental-frequency (F0) amplitude slope maxima. Intervals between these points were computed for each of the five syllables, and the resulting data were treated in the same manner as in the preceding section. For all subjects and syllables, the intervals between F1 amplitude slope maxima were significantly different for ba-target intervals and target-ba intervals ($p < 0.05$). For F0 amplitude slope maxima, ba-target intervals also differed significantly from target-ba intervals for all syllables and speakers except /ha/, where these points were evenly timed in 4 out of 5 subjects ($p > 0.05$).

3.3. Intervals between articulator velocity maxima

Figure 2 shows the pattern of timing between velocity maxima for primary articulators for the different sequences, and are organized in the same way as Figure 4. The figure shows that in "ba-ba" sequences, the mean "ba-target" interval is almost exactly 50% of the interval between alternate /ba/s (velocity-maxima intervals were isochronous for all three subjects in these sequences). It is notable that the 50% figure is approached in many other sequences as well. However, statistical tests reveal that only a few target syllables other than /ba/ showed kinematic isochrony ($p > 0.05$). Subject JB showed isochrony of primary articulator intervals for /li/, and of jaw intervals for /lad/, and subject LK showed isochrony of primary articulator intervals for /ma/ and /spa/ and of jaw intervals for /la/, /ma/, and /spa/.

Overall, however, articulator velocity maxima do not appear to be evenly timed.

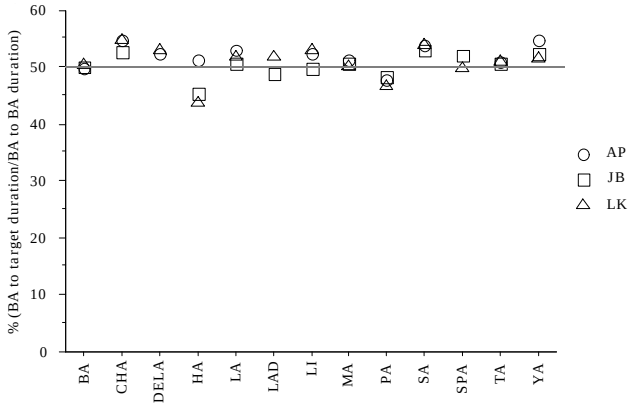

Figure 2. The amount of temporal deviation from isochrony (represented by 50%) for intervals between **primary articulator velocity maxima**. Each subject is represented by a different symbol.

4. CONCLUSION

Over twenty years ago speech researchers observed that speakers and listeners gauge temporal intervals in speech on some basis other than syllable onsets, and posed the "P-center" problem for speech. An acoustically-based solution to this problem would allow the automatic extraction of temporal intervals from speech in a manner that reflects the perception of timing by listeners and the creation of timing by speakers. The possibility also exists that the perceived regularity is reflected not in the acoustics, but in the kinematics of speech. This study examined acoustic and kinematic candidates for the P-center in regularly-timed syllable sequences across a range of syllable types from different speakers. None of these candidates appear to be the cue underlying the P-center. The acoustic and kinematic data from this study (available from http://www.nsi.edu/users/patel) may be useful for testing other candidates for the P-center. Candidates which pass the isochrony test can be examined in synthesis and perception experiments to validate their identity as the cue(s) underlying the perception of temporal intervals in speech.

ACKNOWLEDGEMENTS

We thank Evan Balaban, Kim Beeman, Jennifer Burton, Jim and Leslie Costa, Fuzz Crompton, Farish Jenkins, Jr., John Ohala, Joe Perkell, Stefanie Shattuck-Hufnagel, Kenneth Stevens, and Edward O. Wilson for valuable comments and support. We also thank the subjects who volunteered for this study. A. Patel was supported by a grant from the Arthur Green Fund of Harvard University. This work was supported in part by Grants DC-00865 and DC-02717 from the National Institute of Deafness and Other Communicative Disorders to Haskins Laboratories, and by Neurosciences Research Foundation.

NOTES

1. Before combining these forty intervals into a single population (to study interval variability), it was necessary to include a rate-correction step. This is because average rate differences *between* sequences would lead to inflated measurements of interval variability once the data were pooled. The average duration of the four target-syllable sequences was computed, and for each sequence, a rate correction factor was calculated by dividing average sequence duration by specific sequence duration. (Sequence duration was measured as the time between the onset of the first and last /ba/.) Intervals extracted from a particular sequence were multiplied by the rate correction factor for that sequence. Once this operation was performed, interval data could be pooled without contamination of variance measurements from average rate differences.

REFERENCES

[1] Morton, J., Marcus, S. and Frankish, C. 1976. Perceptual centers (P-centers). *Psychological Review*, 83, 405-408.

[2] Marcus, S. 1981. Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, 30, 247-256.

[3] Cooper, A.M., Whalen, D.H. and Fowler, C.A. 1986. P-centers are unaffected by phonetic categorization. *Perception and Psychophysics*, 39**,** 187-196.

[4] Scott, S. 1993. *P-centers in Speech: An Acoustic Analysis.* Ph.D. Thesis, University College, London.

[5] Fowler, C. and Tassinary, G. 1981. Natural measurement criteria for speech: the anisochrony illusion. In Long, J. and Baddeley, A. (eds.), *Attention and Performance IX*. Hillsdale, NJ: Erlbaum.

[6] Fowler, C. 1979. "'Perceptual centers" in speech production and perception. *Perception and Psychophysics*, 25, 375-388.

[7] Hoequist, C.E. 1983. The perceptual center and rhythm categories. *Language and Speech*, 26**,** 367-376.

[8] Howell, P. 1984. An acoustic determinant of perceived and produced anisochrony. In Van den Broecke, M.P.R. and Cohen, A. (eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences.* Dordrecht: Foris.

[9] Howell, P. 1988a. Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics*, 43, 90-93.

[10] Howell, P. 1988b. Prediction of P-center location from the distribution of energy in the amplitude envelope: II. *Perception and Psychophysics*, 43**,** 99.

[11] Pompino-Marschall, B. 1989. On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, 17**,** 175-192.

[12] Harsin, C. A. 1997. Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics*, 59**,** 243-251.

[13] Tuller, B. and Fowler, C.A. 1980. Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27**,** 277-283.

[14] Fowler, C.A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General*, 112, 386-412.

[15] deJong, K. J. 1994. The correlation of P-center adjustments with articulatory and acoustic events. *Perception and Psychophysics*, 56, 447-460.

[16] Greenberg. S. 1996. Auditory processing of speech. In Lass, N.J. (ed.), *Principles of Experimental Phonetics*, St. Louis:Mosby.

[17] Ghitza, O. 1992. Auditory nerve representation as a basis for speech processing. In Furui, S. and Sondhi, M.M. (eds.), *Advances in Speech Signal Processing*. New York: Marcel Dekker.

[18] Gordon, J. and Ghez C. 1991. Muscle receptors and spinal reflexes: the stretch reflex. In Kandel, E.R., Schwartz, J.H. and Jessell, T.M. (eds.), *Principles of Neural Science*. New York: Elsevier.

[19] Sokal, R. and Rohlf, F.J. 1995. *Biometry (3rd Ed).* New York: W.H. Freeman.

[20] Perkell, J.S., Cohen, M.H., Svirsky M.A. and Matthies, M. 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America* 92, 3078-3096.