INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Investigation of Speed-Accuracy Tradeoffs in Speech Production Using Real-Time Magnetic Resonance Imaging

Adam C. Lammert[1], *Christine H. Shadle*[2], *Shrikanth S. Narayanan*[3], *Thomas F. Quatieri*[1]

[1]MIT Lincoln Laboratory, Lexington, Massachusetts, USA
[2]Haskins Laboratories, New Haven, Connecticut, USA
[3]Signal Analysis and Interpretation Laboratory, Los Angeles, California, USA
adam.lammert@ll.mit.edu, shadle@haskins.yale.edu, shri@sipi.usc.edu, quatieri@ll.mit.edu

Abstract

Motor actions in speech production are both rapid and highly dexterous, even though speed and accuracy are often thought to conflict. Fitts' law has served as a rigorous formulation of the fundamental speed-accuracy tradeoff in other domains of human motor action, but has not been directly examined with respect to speech production. This paper examines Fitts' law in speech articulation kinematics by analyzing USC-TIMIT, a large database of real-time magnetic resonance imaging data of speech production. This paper also addresses methodological challenges in applying Fitts-style analysis, including the definition and operational measurement of key variables in real-time MRI data. Results suggest high variability in the task demands associated with targeted articulatory kinematics, as well as a clear tradeoff between speed and accuracy for certain types of speech production actions. Consonant targets, and particularly those following vowels, show the strongest evidence of this tradeoff, with correlations as high as 0.71 between movement time and difficulty. Other speech actions seem to challenge Fitts' law. Results are discussed with respect to limitations of Fitts' law in the context of speech production, as well as future improvements and applications.
Index Terms: Fitts' law, articulatory difficulty, real-time MRI

1. Introduction

Motor actions associated with speech production are some of the most rapid and dexterous that humans execute. It is not necessarily possible, however, to attain high levels of speed and accuracy at the same time. The present paper examines one aspect of accuracy in speech actions: the kinematics of reaching for maximal articulatory targets. In speech production, there are potentially multiple domains in which accuracy may be demanded, ranging from articulatory and acoustic, to prosodic and communicative, with all of these demands being possibly simultaneous and overlapping. Kinematics are the present focus because many human motor actions exhibit a clear kinematic tradeoff between speed and accuracy. This tradeoff was first formulated rigorously by Paul Fitts in his classic 1954 paper [1]. Fitts posits a linear relationship between the difficulty of a motor task (with difficulty being defined, essentially, in terms of the precision required by task demands) and the time taken to complete that task. Commonly known as *Fitts' law*, this linear relationship has been used widely to model speed-accuracy tradeoffs in a variety of human movement domains. Example application domains include manual pointing and reaching (as in Fitts' original study), eye gaze [2], targeted foot movements [3] and computer device interaction [4]. Still, it is not well-established whether speech motor actions obey this pervasive law of human movement. Despite evidence that speech articulation obeys related tradeoffs among metrics of speed, distance and curvature [5, 6, 7], Fitts' law has not been directly examined in the context of speech production. [1]

It has been argued that speech motor actions vary considerably in terms of their difficulty. Hardcastle [8] asserted that the difficulty (or complexity, to use his terminology) of an articulatory action should be defined in terms of both the number of articulatory variables that are recruited over the course of that action, and in terms of the precision required for each of those variables. The issue of articulatory precision, and its kinematic consequences, is entirely compatible with Fitts' law. Hardcastle even makes direct reference to the speed-accuracy tradeoff in speech production, while arguing that fricatives require more precision than stop consonants: "One of the possible effects of this greater precision is that the articulators involved in the production of a fricative might move more slowly than for the production of a stop." Hardcastle notes that this may help to explain why vowels are often lengthened before fricatives (as originally suggested by MacNeilage [9]) and lower vowels are longer than higher vowels [10]. This is also a possible explanation for the observation that fricatives have longer durations, in general, than stops [11]. The notion of articulatory difficulty may also help to explain why fricatives tend to be acquired later than stops [12]. Differences in difficulty may aid in explaining why some productions are more quickly impacted when the condition of the motor system changes, as in the idea that sleepiness and alcohol intoxication lead to the salient changes in fricatives associated with "slurred speech" [13].

The purpose of this paper is two-fold. The primary goal is to analyze speech articulation from a large database of real-time magnetic resonance (rtMRI) data, in order to assess whether articulatory kinematics conform to Fitts' law. An associated goal is to address the methodological challenges inherent in performing Fitts-style analysis, including how to define the key variables of Fitts' law in the domain of speech articulation, and how to operationalize these definitions on complex and high-dimensional rtMRI data. Section 2 gives a brief introduction to the concepts and mathematics behind Fitts' law. Section 3 describes the data used in the present study, and the necessary preprocessing for the task being considered. Section 4 explains the

[1]This work is sponsored by the Assistant Secretary of Defense for Research & Engineering (ASD[R&E]) under Air Force contract #FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Funding also provided by the National Science Foundation (#1514544).

http://dx.doi.org/10.21437/Interspeech.2016-157

present approach to applying Fitts' law in the domain of speech production data. The results of applying the proposed methodology to rtMRI data, and a discussion of the results in terms of the goals of the paper, are given in Section 5.

2. Fitts' law

Given a *target* associated with a given task, as well as an *initial position* (also, *context*), key parameters of that action can be defined, and incorporated into a simple framework that represents the difficulty associated with that task. One parameter is the *distance* to the target from the initial position. Longer distances are assumed to make a task more difficult. The other parameter is the *width* of the target. A wider target is assumed to make a task less difficult, perhaps corresponding to more slack being permitted in declaring an action successful.

The ratio of the distance, D, and the width, W, are then associated with the *index of difficulty* (ID) in the following way:

$$ID = log_2(D/W + 1) \quad (1)$$

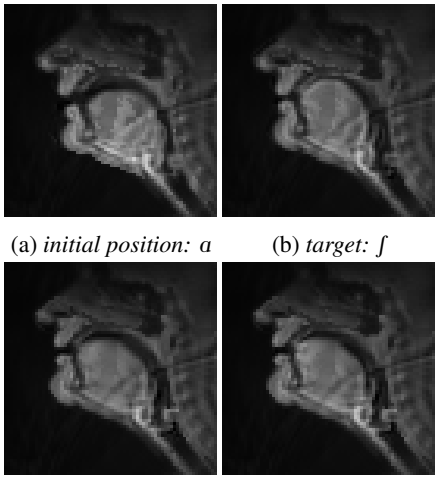
The ratio D/W constitutes one definition of the precision of a task. Taking the base-2 logarithm of this precision, then, gives the ID units that can be interpreted as bits, inspired by Claude Shannon's information theory [14]. The ID, having encapsulated a notion of precision of action, should then be related to the *movement time* (MT) associated with a given task, under the hypothesis that a tradeoff exists between speed and accuracy of that task. This relationship, Fitts' law, is commonly formulated as a simple, linear one:

$$MT = a \cdot ID + b \quad (2)$$

Fitts' law has been derived in various ways since the original formulation [15, 16, 17].

Note that, whereas the distance associated with a task is typically fairly straightforward to define given an initial position and a target (e.g., the Euclidean distance), there have been many definitions presented of the width parameter. Fitts' original experiments included targets with a literal, physical width of varying size, but many experimental setups have only a point target (as assumed in many human actions). In the domain of speech production, however, one is faced with an added complication stemming from a lack of consensus regarding how an articulatory target should be defined, or indeed whether an *articulatory* target (as opposed to acoustic) exists at all. In the present work, it is assumed that articulatory targets do exist, following the specific definition explained below.

To apply Fitts-style analysis to speech production data, it is necessary to operationally define the targets of articulation in space and time. To that end, it is assumed that a single articulatory target is associated with each phoneme. Targets might not be reached during continuous speech for a variety of reasons, including undershoot, misarticulation, or tolerance of the controller to some deviation from the target. However, it is assumed that the action associated with a given phone comes closest to achieving its target at the temporal center of the associated phone interval. Thus, each targeted task in continuous speech can be conceptualized as movement from one phoneme target to another, constituting a specific diphone. Tasks conceptualized this way can also be referred to by diphone, which represents a context-target task pair. It is further assumed that the target of a given phoneme is a vector in high-dimensional articulatory space. The location of that vector is estimated as the mean of all tokens with a given phoneme label. The initial position for a given task is assumed to be the target immediately preceding the current one. All these notions will be defined formally below.



(a) *initial position: a* (b) *target: ʃ*

(c) *initial position: i* (d) *target: d*

Figure 1: *Example high- and low-ID tasks for subject M2. The top row, (a)-(b), represent one of the highest ID tasks, while the bottom row (c)-(d) represents one of the lowest. Images were reconstructed from the L articulatory features in Z (see text).*

3. Real-Time MRI Data & Pre-Processing

Data used here are from the USC-TIMIT database [18]. USC-TIMIT is a publicly-available collection of speech production data from male and female speakers of American English. Speech articulation data were gathered for the database using rtMRI, as well as electromagnetic articulography. Resolution of the rtMRI data is 68 by 68 pixels, with pixels 2.9 by 2.9 mm in size, at a frame rate of the 23.18 frames/s. Audio was simultaneously recorded at 20 kHz sampling frequency, and later subjected to noise cancellation [19]. The rtMRI data from two male and two female subjects from the database (i.e., M1, M2, W1 and W2) were used in the present analysis. Forced phoneme alignment was carried out using SAIL-Align [20]. Subjects were analyzed separately, due to concerns about the proper method of combining articulatory features across subjects.

The analysis presented here began by treating the gray-scale intensity values of each pixel in the image plane as a candidate articulatory feature [21, 22]. These candidate features were pre-processed and recombined prior to analysis, in order to produce new features that are fewer in number and more specific to speech articulation (details below). Such a pixel-wise approach may seen unintuitive, but it provides the opportunity to analyze data about the entire midsagittal plane, while making minimal assumptions about what information might be important for describing articulation. Pixel-wise analysis is also relatively robust compared to a more traditional edge-detection and boundaries-extraction approach when applied to low-contrast, low spatial-resolution rtMR images [23].

The rtMRI image sequences were pre-processed to facilitate further analysis, in particular to (a) isolate frames of interest, and (b) reduce the high dimensionality of the data to a manageable number. Analysis began with an image sequence, X, of the form $X = [I_1 I_2 I_3 \dots I_n]^T$, comprising all n image frames I_m in the corpus from a single subject, where the images I_m are

vectorized in column format. That is, pixels located at (i, j) in rectangular r by c image format are now located at $c(i-1)+j$ in the vector I, and I is of length rc. A retrospective intensity correction scheme was employed, incorporating a nonparametric, monotonically increasing estimate of coil sensitivity, which was derived from all pixel values in the image sequence [22]. Image intensity correction results in a matrix X^c of corrected image vectors.

Pixels that are unrelated to vocal tract action were eliminated by a simple threshold procedure. Pixels representing the air around the head, or representing static spinal or brain tissue, have intensities that change very little over the image sequence. These pixels can be identified by calculating the variance along columns of X^c, and selecting only columns with highest variance. Such pixels represent approximately 75% of all pixels in the images analyzed in the present work, as identified by visual inspection of the images. Therefore, the matrix X^c_{sub} was formed, which contained only those columns of X^c with variance above the 74^{th} percentile across all columns.

The matrix X^c_{sub} is therefore n by $\frac{rc}{4}$ in size, but only a subset of the n data vectors represent vocal tract configurations temporally close to an articulatory target. Using the above operational definition of articulatory targets, the row vectors in X^c_{sub} corresponding to the temporal centers of phones are identified and extracted. From the forced alignment, each phone is assigned a starting boundary A_m, and an ending boundary B_m, both in seconds. From these, the temporal center of a phone can be calculated as $\Gamma_m = \frac{A_m+B_m}{2}$, and the corresponding image frame is $arg_m min(\Gamma_m - \tau_m)^2$ for timestamps $\tau_1, \ldots, \tau_n$ associated with each original image frame. In this way, a new matrix Y is formed, which is P by $\frac{rc}{4}$ in size, where P is the total number of phones represented in the image sequence, $P \approx 15121$.

Principal Component Analysis (PCA) was employed to further reduce the data dimensionality. $Z = YC_L$ was computed, where C is the matrix whose columns are eigenvectors of YY^T, and C_L is a matrix containing only L columns that represent eigenvectors with the highest eigenvalues (i.e., the largest principal components). The magnitude of L was chosen so as to retain $\geq 85\%$ of the variance for each subject being analyzed. Across the subjects analyzed in the present study, L was approximately equal to 50. The resulting P by L matrix Z, which contains a reduced-dimension representation of each vocal tract configuration nearest to an articulatory target, was used for all subsequent analyses. Example images reconstructed from the matrix X are shown in Figure 1.

4. Distance and Width Calculations

For the purposes of analysis, a phoneme vector Π is defined, which is of length P. The p^{th} element of Π, Π_p, is a numerical index from 1 to 35, uniquely specifying an American English phoneme, and representing the phoneme associated with row p of Z. The vector S^g, which is of length P, is associated with a given phoneme index g from 1 to 35. S^g_p = 1 whenever $\Pi_p = g$, and 0 elsewhere. The mean configuration vector associated with the phoneme indexed by g is

$$F^g = \frac{\mathbf{1}^T diag(S^g) Z}{\parallel S^g \parallel_1} \tag{3}$$

where $\mathbf{1}$ is a vector of ones. The vector F^g represents our operationally-defined articulatory target associated with the phoneme indexed by g.

For every pair of phoneme indices g and h, it is now possible to state precisely the spatial distance between the associated

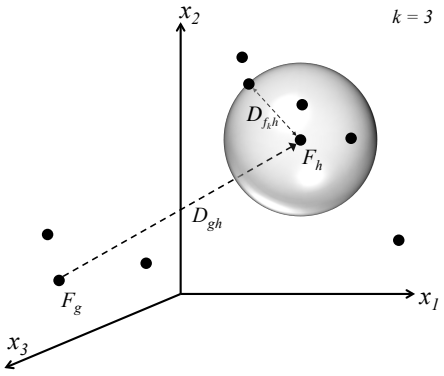


Figure 2: *Illustration of the key components of ID (variable names taken from text). Target vectors are defined in articulatory space, represented here by 3 features, instead of L. The articulatory target vector F_g is the initial position of the current movement, and the target is F_h. Distance to the target is the Euclidean distance between F_g and F_h. Width is calculated with respect to a hypersphere around F_h, used to estimate the density of other target vectors near the current one.*

phonemes. Using the Euclidean distance in the L-dimensional articulatory space, the distance $D_{gh} = \parallel F_g - F_h \parallel$. A graphical representation of this can be seen in Figure 2.

To calculate the time to reach phoneme h from g (indices), assume that S_{gh} is a vector that is 1 whenever both $\Pi_p = h$ and $\Pi_{p-1} = g$. Similarly, S_{hg} is 1 whenever both $\Pi_p = g$ and $\Pi_{p+1} = h$. The mean time, then, between the phonemes indexed by g and h across all instances is

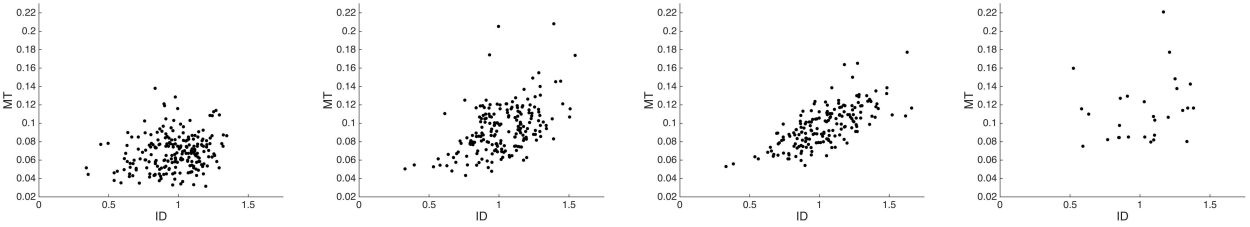
$$T_{gh} = \frac{\mathbf{1}^T diag(S_{gh})\Gamma - \mathbf{1}^T diag(S_{hg})\Gamma}{\parallel S_g \parallel_1} \tag{4}$$

There are many possible definitions for the width of a targeted speech production task. There are no hard physical limits around the target, as in Fitts' original experiments, which necessitates exploring other definitions. Width could be defined in terms of variability about the target, as in later measures of "effective" width [24, 25]. Other definitions have been based on the amount of under/overshoot associated with a particular movement [16]. However, the nature of speech being such that phonetic contrasts can be made with very small changes in vocal tract configuration, allows the possibility for another definition based on the density of targets in articulatory space. Consider the distance values D_{fh} for a given h and all $f = 1, \ldots, 35$. These distance values with respect to h can be sorted and ranked, and – given a parameter k – we can select the distance between F_h and the k^{th} closest vector F_{f_k}. That distance can be used as the basis for a high-dimensional k-nearest-neighbor density calculation. The probability density of configuration vectors in the neighborhood of F_h will be:

$$Q_h = \frac{k}{35 \frac{\pi^{L/2}}{\Gamma(\frac{L}{2}+1)} D^L_{f_k h}} \tag{5}$$

where $\Gamma(x)$ is the gamma function and 35 is the number of phonemes under consideration (24 consonants and 11 vowels, with no diphthongs or rhoticized vowels). The width can be calculated from this probability density as $W_g = -log_2(Q_g)$. Note that the final width value does not depend on the context.

Fitts' law can be calculated directly using D_{gh}, T_{gh} and W_h for any phoneme indexed by h, and presented in the context of another phoneme g. Applying Equation 1, it is possible



(a) CC (b) CV (c) VC (d) VV

Figure 3: *Movement time (MT) vs. index of difficulty (ID) for subject M2. All context-target tasks are shown, divided by diphone type.*

Context		M2	M1	W1	W2
C-C	r	**0.207**	-0.123	**0.303**	**0.252**
n = 220	p	0.002	0.069	< 0.001	< 0.001
C-V	r	**0.486**	0.074	**0.349**	**0.490**
n = 195	p	< 0.001	0.301	< 0.001	< 0.001
V-C	r	**0.711**	0.121	**0.536**	**0.678**
n = 178	p	< 0.001	0.104	< 0.001	< 0.001
V-V	r	0.205	0.205	-0.054	0.205
n = 28	p	0.296	0.315	0.789	0.327

Table 1: *Pearson's r (and p-values) between movement time (MT) and index of difficulty (ID) for all context-task tasks, divided by diphone type. Correlation coefficients significant at the $p < 0.01$ level are highlighted in bold.*

to calculate $ID_{gh} = log_2(D_{gh}/W_h + 1)$. Furthermore, by Equation 2, we expect that $T_{gh} = a \cdot ID_{gh} = b$, for some coefficients a and b.

5. Results and Discussion

Figure 3 plots the relationship between MT and ID for all subjects. Each plot represents a different context-target task type. For instance, CV represents all vowel targets for which the initial position was a consonant. The correlation values (i.e., Pearson's r) between MT and ID – divided into the same context-target types – are shown in Table 1.

Results suggest that the difficulty associated with targeted articulatory kinematics is highly variable in speech production. ID ranges from approximately 0.25 to 1.75 bits for all subjects. A few general patterns in the distribution of ID can be noted. Tasks involving back vowels and/or fricatives and affricates tend to have the highest IDs, while tasks involving short vowels and stop consonants tend to have the lowest IDs. For example, VC tasks with the highest IDs in the current analysis included ɑ-d͡ʒ, ɑ-t͡ʃ, ɑ-ʃ and ɑ-ʒ for subjects M2, W1 and W2, with the raised variants ʊ-t͡ʃ, ʊ-ʃ and ʊ-ʒ for M1. VC tasks with the lowest IDs included i-t, i-d for all four subjects. Figure 1 shows low- and high-ID tasks for subject M2.

Results also suggest that certain types of actions exhibit a clear tradeoff between speed and accuracy. Significant correlations can be seen in the data that correspond to the relationship between MT and ID predicted by Fitts' law. The strength of that relationship varies across context-target type, and across subjects. The strongest such relationships are seen for VC context-target tasks, with CV tasks showing nearly as strong correlations. CC tasks also generally show significant, but much weaker, relationships between MT and ID, whereas VV tasks were not significant for any subject analyzed. Note that many fewer VV tasks exist, as compared to other context-target types. Significant correlations between MT and ID were observed for three of our four subjects. It is clear already from those three subjects that inter-subject variability exists in terms of the strength of MT-ID correlations. However, subject M1 showed no significant correlations.

Despite several significant correlation values between MT and ID, the correlations observed in the present analysis are relatively modest compared to those observed in other domains of human movement. Correlation coefficients above 0.9 are commonly reported in the literature [26]. In general terms, targeted speech kinematics do seem to obey Fitts' law, but with caveats depending on the speaker and the specific actions analyzed. One question raised by such a results is how this seemingly fundamental tradeoff, that has been well-established in other motor domains, can be sometimes weakly obeyed or ignored altogether in speech. A potential explanation is that Fitts' law does not incorporate factors that are crucial to speech production. As mentioned above, speech has multiple levels in which accuracy may be demanded. Speech motor actions have communicative and prosodic goals, in addition to kinematic requirements. Temporal constraints exist as part of those goals, both at the level of phonetic segments (e.g., lengthening as a phonemic contrast) and suprasegmentally (e.g. accenting). A modification of Fitt's law is needed to account for these various levels of task requirements, and associated timing requirements.

It should be noted that there are many sources of variability in the present analysis that may have impacted the correlation values, and may limit the generality of these results. One limitation relates to the accuracy of finding a video frame near to the temporal center of a given phone, which is limited by the temporal resolution of rtMRI and the quality of forced phoneme alignment. Recent advances in rtMRI protocols may alleviate this limitation [27, 28]. Additional variability may stem from non-Gaussian noise on pixel intensity values that rtMRI images often contain. Added variability in the data/analysis would have the clearest impact on the VV diphone correlation results, due to their much smaller number. Data are also limited to a midsagittal view of the speech articulators, meaning not all kinematic aspects are captured in the data.

A possible application of the present work is in designing new speech features for assessing neurocognitive changes. Features based on timing and duration, such as speaking rate, have been widely adopted as part of clinical assessments of neurological conditions. Moreover, phoneme-level speaking rate has been proven highly effective for predicting neurological changes due to conditions ranging from depression to Parkinson's disease [29, 30, 31]. Mean phoneme duration is highly correlated with mean MT in the current data set (e.g., for subject M2, Pearson's $r = 0.681$, with $p \ll 0.001$, n=899). If phoneme durations are sensitive to task demands, it may help to explain why phoneme durations are good predictors of neurocognitive change.

6. References

[1] P. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology*, vol. 47, no. 6, pp. 381–391, 1954.

[2] C. Ware and H. H. Mikaelian, "An evaluation of an eye tracker as a device for computer input," in *ACM SIGCHI Bulletin*, vol. 17, no. SI. ACM, 1987, pp. 183–188.

[3] C. G. Drury, "Application of Fitts' Law to foot-pedal design," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 17, no. 4, pp. 368–373, 1975.

[4] S. K. Card, W. K. English, and B. J. Burr, "Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT," *Ergonomics*, vol. 21, no. 8, pp. 601–613, 1978.

[5] A. Lofqvist and V. L. Gracco, "Lip and jaw kinematics in bilabial stop consonant production," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 4, pp. 877–893, 1997.

[6] P. Perrier and S. Fuchs, "Speed–curvature relations in speech production challenge the 1/3 power law," *Journal of Neurophysiology*, vol. 100, no. 3, pp. 1171–1183, 2008.

[7] T. Kato, S. Lee, and S. Narayanan, "An analysis of articulatory-acoustic data based on articulatory strokes," in *International Conference on Acoustics, Speech & Signal Processing*, 2009, pp. 4493–4496.

[8] W. Hardcastle, *Physiology of Speech Production: An Introduction for Speech Scientists*. London: Academic Press, 1976.

[9] P. MacNeilage, *Speech physiology*. University of Texas, 1972.

[10] I. Lehiste, *Suprasegmentals*. MIT Press, 1970.

[11] H. Kuwabara, "Acoustic properties of phonemes in continuous speech for different speaking rate," in *ICSLP 96*, vol. 4, 1996.

[12] M. Templin, *Certain language skills in children; their development and interrelationships.* University of Minnesota Press, 1957.

[13] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states: A review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, vol. 28, no. 2, pp. 346–374, 2014.

[14] C. Shannon and W. Weaver, *The mathematical theory of information*. University of Illinois Press, 1949.

[15] E. Crossman and P. Goodeve, "Feedback control of hand movement and Fitt's law," *Quarterly Journal of Experimental Psychology*, vol. 35A, pp. 251–278, 1983.

[16] D. Bullock and S. Grossberg, "Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation," *Psychological Review*, vol. 95, no. 1, pp. 49–90, 1988.

[17] D. Beamish, S. A. Bhatti, I. S. MacKenzie, and J. Wu, "Fifty years later: a neurodynamic explanation of Fitts' law," *Journal of The Royal Society Interface*, vol. 3, no. 10, pp. 649–654, 2006.

[18] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[19] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[20] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[21] A. C. Lammert, M. I. Proctor, S. S. Narayanan *et al.*, "Data-driven analysis of realtime vocal tract MRI using correlated image regions." in *INTERSPEECH*. Citeseer, 2010, pp. 1572–1575.

[22] A. C. Lammert, V. Ramanarayanan, M. I. Proctor, S. Narayanan *et al.*, "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis." in *INTERSPEECH*, 2013, pp. 959–962.

[23] A. Lammert, L. Goldstein, V. Ramanarayanan, and S. Narayanan, "Gestural control in the English past-tense suffix: an articulatory study using real-time MRI," *Phonetica*, vol. 71, no. 4, pp. 229–248, 2014.

[24] A. Welford, *Fundamentals of Skill*. Methuen, 1968.

[25] P. Fitts and J. Peterson, "Information capacity of discrete motor responses." *Journal of experimental psychology*, vol. 67, no. 2, p. 103, 1964.

[26] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *Human-computer interaction*, vol. 7, no. 1, pp. 91–139, 1992.

[27] S. Lingala, B. Sutton, M. Miquel, and K. Nayak, "Recommendations for real-time speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[28] S. Lingala, Y. Zhu, Y. Kim, A. Toutios, S. Narayanan, and K. Nayak, "A fast and flexible MRI system for the dynamic study of vocal tract shaping," *Magnetic Resonance in Medicine*, 2016.

[29] A. Trevino, T. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech*, 2011.

[30] J. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 65–72.

[31] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. Ciccarelli, and D. D. Mehta, "Segment-dependent dynamics in predicting Parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.