Research Article

# Engineering Innovation in Speech Science: Data and Technologies

Christina Hagedorn,[a,b] Tanner Sorensen,[c] Adam Lammert,[d] Asterios Toutios,[e] Louis Goldstein,[c] Dani Byrd,[c] and Shrikanth Narayanan[f]

**Purpose:** As increasing amounts and types of speech data become accessible, health care and technology industries increasingly demand quantitative insight into speech content. The potential for speech data to provide insight into cognitive, affective, and psychological health states and behavior crucially depends on the ability to integrate speech data into the scientific process. Current engineering methods for acquiring, analyzing, and modeling speech data present the opportunity to integrate speech data into the scientific process. Additionally, machine learning systems recognize patterns in data that can facilitate hypothesis generation, data analysis, and statistical modeling. The goals of the present article are (a) to review developments across these domains that have allowed real-time magnetic resonance imaging to shed light on aspects of atypical speech articulation; (b) in a parallel vein, to discuss how advancements in signal processing have allowed for an improved understanding of communication markers associated with autism spectrum disorder; and (c) to highlight the clinical significance and implications of the application of these technological advancements to each of these areas.
**Conclusion:** The collaboration of engineers, speech scientists, and clinicians has resulted in (a) the development of biologically inspired technology that has been proven useful for both small- and large-scale analyses, (b) a deepened practical and theoretical understanding of both typical and impaired speech production, and (c) the establishment and enhancement of diagnostic and therapeutic tools, all having far-reaching, interdisciplinary significance.
**Supplemental Material:** https://doi.org/10.23641/asha.7740191

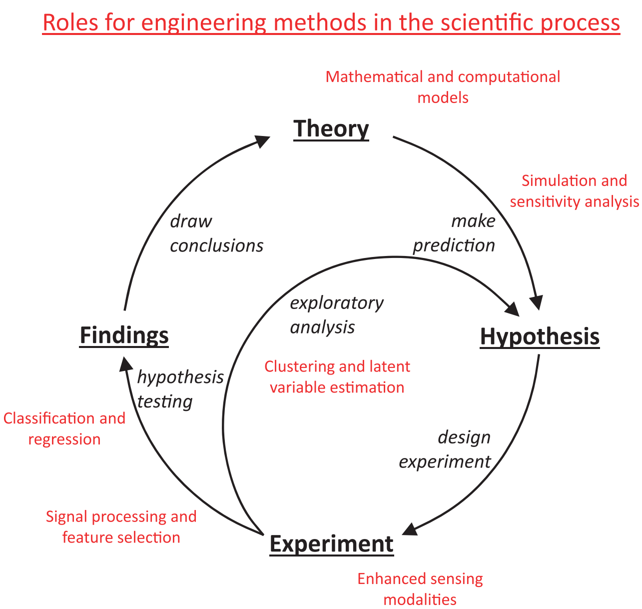
As increasing amounts and types of speech data become accessible, health care and industry increasingly demand quantitative insight into speech content. The potential for speech data to provide insight into behavior, cognitive, and other psychological states crucially depends on the ability to integrate these data into the scientific process. Current engineering methods for acquiring, analyzing, and modeling speech data present the opportunity to integrate speech data into the scientific process (see Figure 1). For instance, novel sensors enable accurate measurement of signals in controlled experiments, in the clinic, and at home. Additionally, machine learning systems recognize patterns in data that can facilitate hypothesis generation, data analysis, and statistical modeling. Accordingly, these advancements in hardware development, software development, and signal processing have led to an improved understanding of a variety of communication disorders. The goals of the present article are (a) to review developments across these domains that have allowed real-time magnetic resonance imaging (rtMRI) to shed light on aspects of atypical speech articulation; (b) in a parallel vein, to discuss how advancements in signal processing have allowed for the identification of communication markers associated with autism spectrum disorder (ASD) using quantitative approaches; and (c) to highlight the clinical significance and implications of these technological advancements in each of these areas.

[a]Linguistics, College of Staten Island, City University of New York, NY
[b]Linguistics, The Graduate Center, City University of New York, NY
[c]Linguistics, University of Southern California, Los Angeles, CA
[d]Bioengineering Systems & Technologies, MIT Lincoln Laboratory, Lexington, MA
[e]Electrical Engineering, University of Southern California, Los Angeles, CA
[f]Signal and Image Processing Institute, University of Southern California, Los Angeles, CA
Correspondence to Christina Hagedorn: christina.hagedorn@csi.cuny.edu

Editor: Alison Behrman
Received May 20, 2018
Revision received September 10, 2018
Accepted September 13, 2018
https://doi.org/10.1044/2018_PERS-SIG19-2018-0003

**Disclosures**
*Financial: The authors have no relevant financial interests to disclose.*
*Nonfinancial: The authors have no relevant financial interests to disclose.*

**Figure 1.** The role for engineering methods in the various stages of the scientific process.




## The Development of Engineering Techniques and Tools for the Acquisition and Analysis of Speech Production Data Using Magnetic Resonance Imaging

Through the years, engineering has played a critical role in the development of techniques and tools allowing for the investigation of typical and atypical speech production using magnetic resonance imaging (MRI). Static structural MRI and cine MRI, used to detect and differentiate levels of hydrogen concentration in the soft tissues of the human body, have been used to investigate the state of the vocal tract at individual points in time during speech production (e.g., Narayanan, Alwan, & Haker, 1995; Stone et al., 2001; Story, Titze, & Hoffman, 1996; Takemoto, Honda, Masaki, Shimada, & Fujimoto, 2006). MRI is a particularly useful tool for the investigation of speech production, given that (a) it is minimally invasive, requiring the participant only to read phrases from a projected screen or speak spontaneously in response to a spoken or textual cue, (often) lying supine in the scanner bore, without any sensors being adhered to their articulators; (b) it subjects the participant to no ionizing radiation; and (c) it allows for visualization of the entire vocal tract, including the velum, pharynx, and larynx, in various planes (e.g., midsagittal, coronal). Static MRI provides rich spatial information about the articulators at only one point in time and so is conducive to studying only sustained articulations, such as vowels, liquids, and fricatives (Narayanan et al., 1995; Story et al., 1996). Based on the same technology, cine MRI attempts to reconstruct the dynamics of articulation by collecting static configurations of the vocal tract over several productions of a single item, sequencing them to render an approximation of real-time production (e.g., Stone et al., 2001; Takemoto et al., 2006).

Over the past two decades, rtMRI, elaborating traditional medical MRI, has played a critical role in studying a variety of biological movement patterns, including cardiac motion (Nayak & Hu, 2005) and joint kinematics (Zhang, Gersdorff, & Frahm, 2011). Engineers and speech scientists have teamed up to apply rtMRI technology to the study of speech production (Bresch, Kim, Nayak, Byrd, & Narayanan, 2008; Y. Kim, Narayanan, & Nayak, 2009; Narayanan, Nayak, Lee, Sethy, & Byrd, 2004). In addition to imaging advances, this includes, for example, the development of noise cancellation technology (Bresch, Nielsen, Nayak, & Narayanan, 2006; Vaz, Ramanarayanan, & Narayanan, 2018), which has been essential in applying rtMRI to the study of speech production, given the intensity of scanner noise and the importance of acquiring reliable acoustic data for analysis alongside articulatory data. Likewise, automatic tongue and vocal tract contour tracking tools that rely on identification of air-tissue boundaries (Bresch & Narayanan, 2009; J. Kim, Kumar, Lee, & Narayanan, 2014; Proctor, Bone, Katsamanis, & Narayanan, 2010) have played a critical role in both quantitative and qualitative analyses of speech production patterns in rtMRI data. In contrast to static MRI and cine MRI, rtMRI does not require participants to produce several repetitions of each token but rather allows for fast acquisition rates (e.g., 83 frames per second or higher; Lingala et al., 2017) on single tokens. rtMRI for speech has been shown to effectively shed light on a wide variety of phenomena in both typical and disordered speech production that would not be possible to investigate using tools providing more limited spatiotemporal information about vocal tract shaping (Carignan, Shosted, Fu, Liang, & Sutton, 2015; Feng et al., 2018; Perry, Kuehn, Sutton, & Fang, 2017).

### *Applications of rtMRI to the Study of Typical Speech Production*

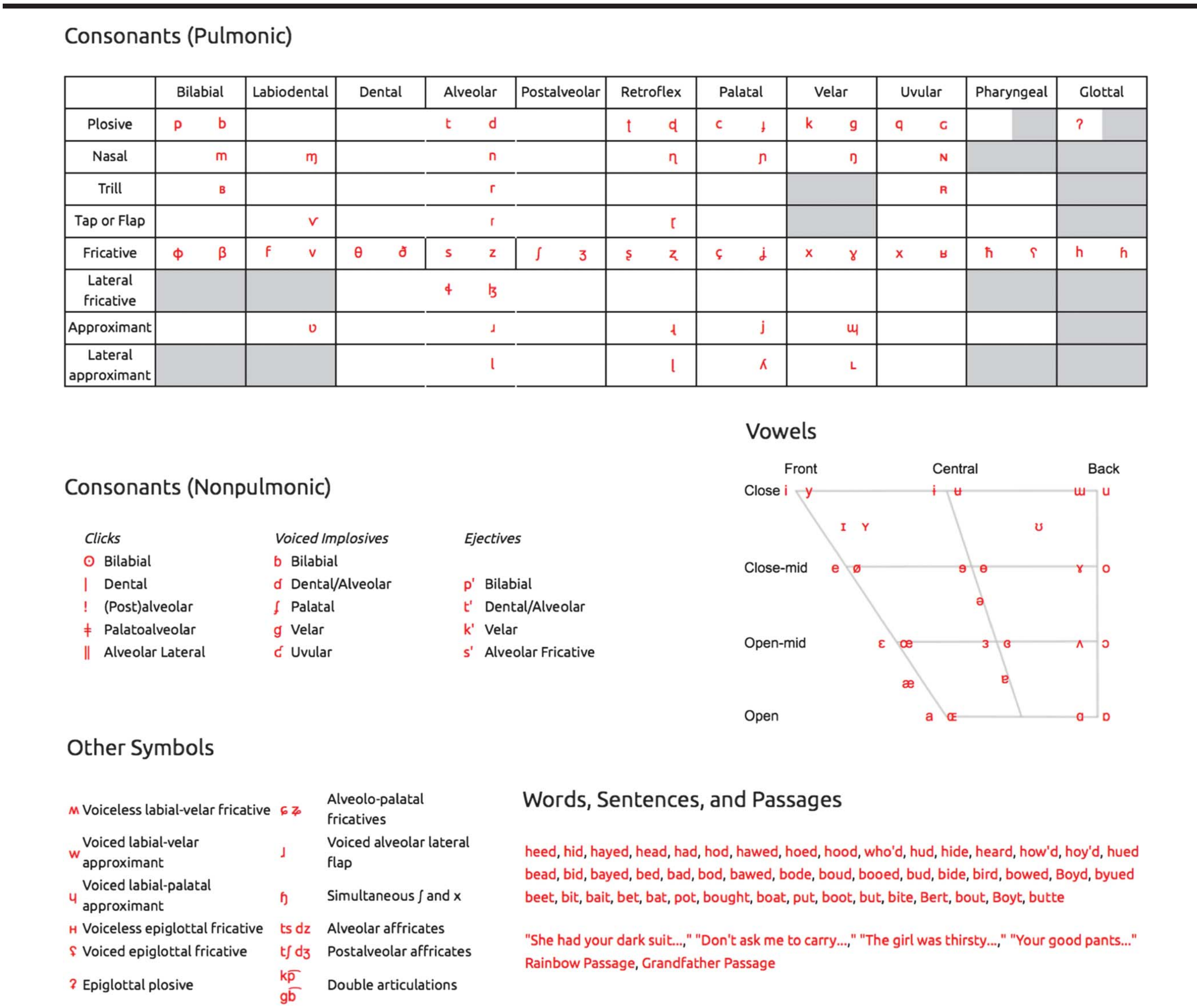
rtMRI has been used to investigate typical speech production, focusing on articulatory behavior that underlies speech sounds produced in isolation (Toutios et al., 2016), in systematically controlled phonetic environments (Byrd, Tobin, Bresch, & Narayanan, 2009; Lammert, Goldstein, Ramanarayanan, & Narayanan, 2015; S. Lee, Potamianos, & Narayanan, 2014; Y. Lee, Goldstein, & Narayanan, 2015), and in naturalistic, running speech (Narayanan et al., 2014; Proctor, Lo, & Narayanan, 2015).

Toutios et al. (2016) illustrate an organized collection of rtMRI data available for public use that comprises data from four distinguished phoneticians producing each sound of the International Phonetic Alphabet (IPA). All vowel sounds, elicited in isolation, and all sounds in the Pulmonic Consonants, Nonpulmonic Consonants, and Other Symbols sections of the IPA chart, elicited in the context of low back vowel /a/ (i.e., [aCa]), are included. Additionally, monosyllabic words containing all vowels and diphthongs of American English of the form [hVd] are included, as well as four phonetically rich sentences and the Rainbow and Grandfather passages. Figure 2 displays the clickable IPA symbols, words, sentences, and passages, each linking to its respective production clip for each of the four speakers. Pronunciation training, as occurs in phonetics instruction, second language acquisition, and therapeutic intervention for speech disorders, typically relies on close listening and participants' visual observation of the instructor's extraoral articulators, with less visual information regarding lingual movement being available to the learner. The rtMRI data, becoming increasingly available, have the potential to serve as a rich complement to existing pronunciation training tools, allowing learners to observe target movement patterns of articulators that are typically hidden within the vocal tract.

### *Applications of rtMRI to the Study of Disordered Speech Production*

In addition to its utility in characterizing typical speech articulation, rtMRI has been used to investigate speech produced by individuals with speech impairments secondary to a variety of conditions, including aglossia (McMicken et al., 2017; Toutios, Byrd, Goldstein, & Narayanan, 2017), oral cancer (Hagedorn et al., 2014; Hagedorn, Kim, et al., 2017; Lander-Portnoy, Goldstein, & Narayanan, 2017), and

**Figure 2.** Snapshot from the web resource, illustrating the stimulus set. On the web resource, symbols, words and phrases link to real-time MRI videos of their productions. Reprinted from "Illustrating the Production of the International Phonetic Alphabet Sounds using Fast Real-Time Magnetic Resonance Imaging" by A. Toutios et al., 2016, INTERSPEECH, p. 2428. Copyright 2016 by INTERSPEECH. Reprinted with permission.



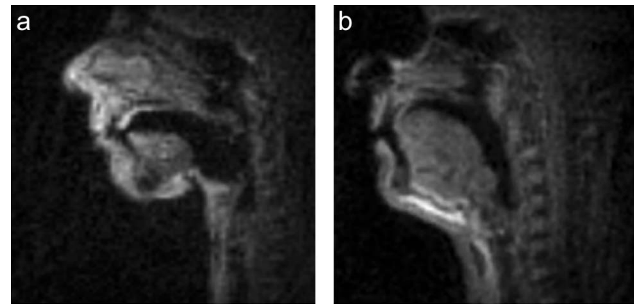
acquired apraxia of speech (AoS; Hagedorn, Proctor, et al., 2017).

rtMRI reveals that speakers with congenital aglossia, a syndrome in which a person is born without a (complete) tongue, and speakers with oral cancer who have undergone glossectomy, in which part of the tongue is surgically resected, articulatorily compensate to produce sounds that they are not able to produce in the same way as individuals having typical speech. Specifically, McMicken et al. (2017) found that a speaker with congenital aglossia, who lacked a tongue tip, produced target coronal constrictions using bilabial closure. Upon recognizing that this speaker's compensatory (labial) productions for /t/ and /d/ were perceptually similar to targets /t/ and /d/, despite being produced bilabially, researchers further investigated the underlying cause of the perceptual distinction between the speaker's target bilabials and target coronals that were also produced bilabially. Work by Toutios et al. (2017) leveraged the global view of the vocal tract that rtMRI provides to demonstrate that the speaker's target coronals (produced bilabially) differed from target bilabial stops in that they were produced with an elongated bilabial constriction and widening of the pharynx (Supplemental Material S1). An acoustic simulation experiment carried out by the authors provided evidence that this specific combination of compensatory strategies was indeed required to produce the perceptual differences between the speaker's production of target bilabials and target coronals.

In a similar vein, work by Hagedorn et al. (2014) and Hagedorn, Kim, et al. (2017) demonstrated that individuals with oral cancer who underwent resection of the oral tongue can, in some circumstances, compensate for consonant targets using articulators other than those typically used. Specifically, rtMRI revealed that one individual who underwent resection of the oral tongue produced target coronal stops as labiodental stops, while producing target coronal fricatives using the tongue dorsum (Supplemental Material S2). Investigation of the speech produced by an individual who underwent resection of the oral and base of tongue revealed that target coronal stops were produced using simultaneously the tongue tip to create incomplete closure against the alveolar ridge and the lips to create complete occlusion (Supplemental Material S3). Target coronal fricatives were compensated for by creating both alveolar (using the residual tongue tip) and labial constrictions (see Figures 3a and 3b). rtMRI also shed light on vowel production in postglossectomy speech, revealing not only that vertical jaw position and lingual constriction degree are more highly correlated in patients than in typical speakers but also that larger differences in constriction degree are achieved, per unit jaw height, in patients than in typical speakers, suggesting that patients use modulation of jaw height in order to compensate for reduced lingual mobility.

In addition to its application for the study of speech impairment consequent to atypical articulatory structures, as in cases of aglossia or glossectomy, rtMRI has also been used to shed light on the nature of speech disorders arising from breakdown at the planning and programming levels

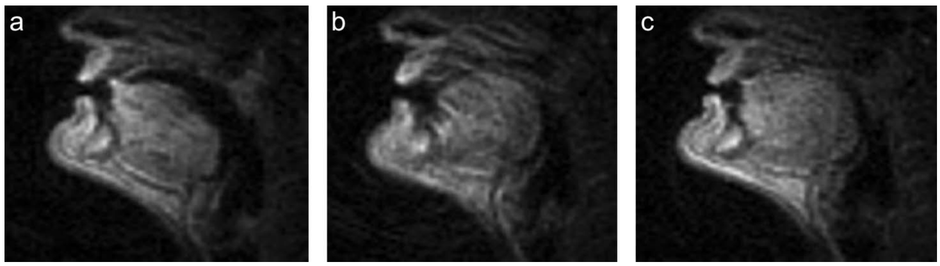
**Figure 3.** Glossectomy patient (a) produces /d/ by creating both labial and coronal constrictions; typical speaker (b) produces /d/ by creating only a coronal constriction. Note the reduced lingual mass in the glossectomy patient as compared with the typical speaker.




of speech motor control, as occurs in acquired AoS. Using rtMRI, Hagedorn, Proctor, et al. (2017) characterized covert articulation in AoS, whereby two articulatory movements are coproduced—one corresponding to the target sound and the other as an erroneous articulatory movement that is not typical for the target sound. These covert, and often auditorily undetectable, "intrusion" errors were shown to occur not only in word-pair repetition tasks (e.g., "cop-top-cop-top-cop-top…") but also in repeated sentences (e.g., repeating "I can type bow know five times."; Supplemental Material S4). Further, covert multiple initiation gestures (i.e., instances of unphonated articulatory groping) were shown to occur more frequently in segments requiring the coordination of multiple vocal tract articulators (e.g., the lips and tongue as for /w/) than in those requiring a single vocal tract gesture (e.g., the lips as for /b/; see Figures 4a–4c).

The capability of rtMRI to shed light on disordered speech has translational significance, in that the findings of these studies can be used to refine therapeutic intervention techniques. The findings of McMicken et al. (2017), Toutios et al. (2017), Hagedorn et al. (2014), and Hagedorn, Kim, et al. (2017) provide insight into spontaneous compensation strategies (i.e., those produced on one's own, without the direction of a clinician) that can give rise to typical or near-typical target acoustics in individuals with congenital aglossia and in those who have undergone lingual resection. These strategies could be used by clinicians in order to fine-tune therapy programs for those who do not (sufficiently) compensate spontaneously, requiring therapy to remediate articulation disorders caused by structural impairments. The findings of Hagedorn, Proctor, et al. (2017) suggest that therapeutic intervention for individuals with AoS might be fine-tuned to focus on the production of complex segments that require the coordination of multiple supralaryngeal articulators. Many of these preliminary findings using rtMRI to investigate disordered speech emphasize that clinicians ought to be aware that the breakdowns, deficiencies, and/or atypicalities in articulatory coordination can manifest silently—that is, with no perceptually salient acoustic trace—or as coproduction errors, for example, with two vocal tract movements (one target and one erroneous) being

**Figure 4.** Speaker with apraxia of speech produces target /t/ (a), target /k/ (b), and intrusion error (c), in which coronal and dorsal gestures for /t/ and /k/ are coproduced.




produced simultaneously, rather than an erroneous movement completely replacing the target movement.

Although rtMRI has successfully been used to shed light on several aspects of both typical and atypical speech production, its implementation comes not without challenges. For example, the application of rtMRI to the study of speech in certain clinical populations is oftentimes difficult due to constraints of patient availability and willingness to participate. Apart from these limitations, the presence of metal implants (e.g., reconstructive rods or plates placed during surgery) or pacemakers may preclude participants from being eligible for participation in studies relying on magnetic fields due to risk of electromagnetic interference, leading to artifacts that render quality of the data suboptimal and, more importantly, physical risk to the participant. Moreover, individuals with claustrophobia who might otherwise be willing to participate in such studies may experience difficulty remaining in a confined space (such as an MRI scanner bore) for the time that such studies often require. Similarly, both pediatric and adult participants, and particularly those in the patient population, may experience difficulty remaining stationary yet alert for the duration of the experiment, due to fatigue.

In summary, rtMRI and the novel analytical approaches described herein have aided not only in characterizing disordered speech in a manner not possible using other speech production research tools but also in informing speech scientists and clinicians of ways in which fine-tuning speech intervention programs may be beneficial to patients' speech outcomes. Engineering techniques, propelled by advances in signal processing and machine learning, are also enabling new possibilities both in helping illuminate the scientific underpinnings of various mental and behavioral health conditions and in supporting the creation of novel screening and diagnostic measures. The next section highlights some of these possibilities using a case study of ASD.

## The Role of Speech Engineering Techniques and Tools for Autism Research and Its Clinical Applications

ASD is a neurodevelopmental disorder typically characterized by difficulties with social communication, repetitive behaviors, and restricted interests. ASD is among the most highly heritable psychiatric disorders (Bargmann & Gilliam, 2013; Tick, Bolton, Happé, Rutter, & Rijsdijk, 2016), but the pattern of inheritance is complex, and the complex etiology remains incompletely understood, with only 10%–20% of clinical cases having a known genetic cause (Geschwind, 2011). A consequence of this uncertain etiology is that the diagnosis of ASD is based not on disease pathology but rather on expert assessment of child behavior and cognition.

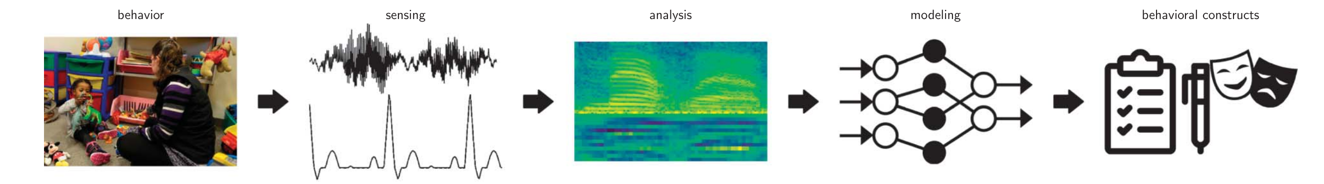
The assessment of behavior is challenging due to variability in behavior and subjectivity in its interpretation. Variability in child behavior reflects a diverse set of factors, including comorbid psychiatric disorders, instantaneous physiological state, attention, emotion, and motivation. In addition to this variability in the generation of behavior, the observer interprets the behavior differently depending on the nature of the interaction: A physician screening the child for developmental disorders in a health care setting may assign clinical codes to the behavior, whereas a parent playing with the child may attribute to the behavior such qualities as engagement, enjoyment, and responsiveness.

Behavioral signal processing (BSP) is a framework for sensing the signals that arise from behavior, analyzing behavior from these signals, and modeling the behavioral constructs that observers abstract from the signals (Narayanan & Georgiou, 2013; see Figure 5). For instance, a BSP system may use a microphone and an electrocardiograph to sense a child's speech and heart rate during a dyadic interaction with a clinical psychologist, recognize speech intonation and autonomic arousal, and model the behavioral construct of social anxiety (Bone, Mertens, et al., 2017). The purpose of a BSP model in the domain of ASD is not to replace but rather to complement expert assessment of behavior and cognition (Bone et al., 2016; Bone, Goodwin, et al., 2015; Bone, Mertens, et al., 2017). This section describes state-of-the-art engineering tools for enhancing the health care of individuals with ASD and accelerating scientific discovery in the disorder.

### *Speech Prosody in ASD*

Speech processing methods have led to scientific discovery in ASD in the area of speech prosody. Prosody captures emphasis, emotion, and grammatical distinctions, such

**Figure 5.** Behavioral signal processing. Human behavior generates signals that are transduced by sensors such as a microphone or an electrocardiograph. Features (e.g., pitch) are extracted from the behavioral signals. Statistical models and machine learning predict behavioral constructs (e.g., social anxiety, autism severity) from the features.




as between statements, questions, and commands. Prosody is oftentimes disrupted in ASD (Hubbard & Trauner, 2007; McCann & Peppé, 2003); children with ASD frequently exhibit "monotonic" speech, variable volume, atypical voice quality, and slow rate of speech. Engineers have begun to use speech processing to identify the basis of this perception in the speech signal.

ASD may be reflected in pitch patterns. Children with more severe ASD tend to have a greater fall in pitch at the end of a speaking turn than children with less severe ASD (Bone, Lee, Black, et al., 2014). In American English, falling pitch generally indicates that the utterance is a statement, but in the speech of children with ASD, listeners may perceive a persistently falling pitch slope as an indicator of monotone speech.

Aperiodicity in the speech signal, reflected in the features of jitter and shimmer, has been linked to perceptions of breathiness, hoarseness, and roughness (McAllister, Sundberg, & Hibi, 1998). ASD severity can be indexed using these measures, with high values and high variability of jitter being associated with more severe ASD (Bone, Lee, Black, et al., 2014). Signal periodicity can be measured by cepstral peak prominence and harmonics-to-noise ratio. These features have been linked to perceptions of vocal breathiness (Hillenbrand, Cleveland, & Erickson, 1994) and harshness (Halberstam, 2004). Children with ASD tend to exhibit low and variable harmonics-to-noise ratio (Bone et al., 2014). Given that intrarater reliability of voice quality is low (Gelfer, 1988; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993), the identification of features such as jitter, shimmer, cepstral peak prominence, and harmonics-to-noise ratio that quantitatively and objectively characterize voice quality takes on particular importance.

The engineering tools of speech processing surveyed above provide the means to analyze facets of prosody such as pitch, volume, rate, and voice quality. Although these speech features are objective descriptors of prosody, speech features alone do not characterize clinically and ecologically important behavioral constructs such as "awkwardness" and "expressivity." The promise of applying BSP to characterize prosody in ASD lies in the potential for discovering and modeling the signal properties that produce the perception of expressivity or awkwardness. BSP studies reveal that speech rate, speech rhythm, and the way in which lexical content and prosody are combined all contribute to the perception of awkwardness, whereas the dynamics of pitch and intensity contribute to the perception of expressivity (Bone, Black, Ramakrishna, Grossman, & Narayanan, 2015). Reliable modeling of how these factors interact to give rise to speech that is perceived as awkward or expressive has the potential to contribute to the formulation of therapeutic biofeedback tools. These tools may help individuals with ASD modulate the relevant aspects of the speech signal based on targets derived from quantified measures to produce speech that is perceived as more typical. Finally, BSP allows for jointly modeling the interplay between the prosodic patterns of the interlocutors such as a child and a clinician interacting in a diagnostic setting to offer further insights about coordination and interaction synchrony not easily available by human observation (Bone, Lee, Black, et al., 2014).

### *Nonverbal Communication in Autism*

Speech communication is inherently multimodal. Verbalizations are augmented by nonverbal cues, such as facial expressions, head movements, and eye gazes. Engineering methods have also led to a scientific discovery in the nonverbal components of communication as affected by ASD. Such nonverbal components can be extremely subtle, signaled by minor adjustments in body posture and small deformations of the face. The complexity of the behavior furnishes the challenge of identifying meaningful data modalities and feature sets for quantifying nonverbal communication, as well as the challenge of determining the dimensions along which nonverbal communication components differ between typically developing children and children with ASD.

Facial expression unfolds over time on different regions of the face and is perceived highly subjectively. This makes deficits in facial expression difficult to characterize by visual inspection alone. To this end, studies have adopted motion capture and video of the face to characterize facial expression. Motion capture is a technology that uses a camera array to track the position of facial markers in space with high precision and accuracy (e.g., Fernández-Baena, Susín, & Lligadas, 2012; Wagner, Malisz, & Kopp, 2014; Windolf, Götzen, & Morlock, 2008; Supplemental Material S5). Computational modeling of video data allows for quantifying facial expressions and head movement by using the two-dimensional video to reconstruct a three-dimensional point mesh that represents the surface of the face (Jeni, Cohn,

& Kanade, 2015). Studies using these methods show that children with ASD produce facial expressions with different spatiotemporal dynamics compared with typically developing children. Facial expression in ASD has reduced complexity, primarily due to the nature of movement in the region of the eyes (Guha, Yang, Grossman, & Narayanan, 2018). Additionally, children with ASD tend to display asynchrony and asymmetry between expressions of the left and right face and asynchrony between upper and lower face movement (Guha et al., 2015; Metallinou, Grossman, & Narayanan, 2013).

Differences in the nonverbal communication component between children with ASD and typically developing children depend on the social context in which the children generate the behavior. For instance, computational analysis of the head movement video shows that children with ASD display greater rigid head motion than typically developing children when presented with social stimuli but show no differences when presented with nonsocial stimuli (Martin et al., 2018). Similarly, infants who are later diagnosed with ASD scan the face differently from typically developing infants when the face is producing speech but not when the face is not producing speech (Shic, Macari, & Chawarska, 2014). This parallels findings that show that children with ASD display atypical attention to biological and social stimuli during development and show atypical visual attention to faces compared with typically developing children (Jones, Carr, & Klin, 2008; Klin, Jones, Schultz, Volkmar, & Cohen, 2002). Children with ASD preferentially attend to the mouth and body, whereas typically developing children prefer the eyes. Moreover, unlike typically developing children, 2-year-olds with ASD do not preferentially attend to biological motion over nonbiological motion (Klin, Lin, Gorrindo, Ramsay, & Jones, 2009). The dependence of behavior in ASD on the social environment of the child offers insight into the experience of children with ASD that may underlie experience-dependent changes in behavior. Atypical attention to biological and social stimuli limits the input available for learning nonverbal communication and may accompany the child failing to learn to generate typical facial expressions.

### *Social Interaction in Autism*

A core diagnostic criterion for ASD is persistent deficits in social communication and social interaction. These deficits affect both verbal and nonverbal components of speech communication. Depending on the severity of the disorder, this deficit ranges from poor integration of verbal and nonverbal communication to abnormalities in eye contact and body language or deficits in understanding and use of gestures and to a total lack of facial expressions and nonverbal communication (American Psychiatric Association, 2013). Although clinicians excel at identifying and rating communication deficiencies, automatically quantifying aspects of social communication remains difficult due to the holistic nature of social interaction. For this reason, modeling social communication between children with ASD and peers, parents, and health care professionals takes on importance as a promising application of BSP and a unique opportunity to quantify a core deficit of the disorder.

Psychologist–child interaction has been studied in a health care setting in which the psychologist is diagnosing the child with ASD (Bone, Lee, Black, et al., 2014) or screening the child for developmental disorders (Gupta, Bone, Lee, & Narayanan, 2016). As is typical of dyadic interactions, the vocal behaviors of the child and psychologist are correlated with each other during these interactions (Bone, Lee, Black, et al., 2014). Specifically, vocal intensity and voice quality are correlated between the child and the psychologist. Thus, psychologist behavior also reflects the ASD severity of the child. The psychologist exhibits differences in variability in pitch contours, variability in vocal intensity, and voice quality, depending on the ASD severity of the child whom they are diagnosing. In fact, Bone, Lee, Black, et al. (2014) find that three of four features identified as significant predictors of ASD severity were prosodic features of the psychologist's speech, whereas only one of four was a prosodic feature of the child's speech, indicating a profound difference in psychologist behavior depending on the child's ASD severity.

The application of speech technologies such as automatic voice activity detection ("When is speech present?") and diarization ("Who is speaking when?") enables the automated determination of speaking times in child–psychologist interactions. This approach has revealed that the percentage of the session that the psychologist spends speaking tends to increase with increasing ASD severity of the child (Bone, Lee, Black, et al., 2014). This likely reflects some combination of low engagement and social communication deficits in severe ASD.

The tools of speech processing surveyed above provide the means to analyze the covariance of prosody (pitch, volume, rate, and voice quality) and turn-taking behavior in social interactions. Although these speech features describe the social interaction as a whole, patterns of covariance alone do not characterize the temporal dynamics of the features and corresponding behavioral constructs such as engagement or emotion. The promise of applying BSP to characterize social interactions in ASD (or indeed in conversations among typical speakers) lies in the potential for modeling how interlocutors adaptively modulate the temporal dynamics of prosody and turn-taking behaviors over the course of an interaction and how these modulations influence the behavioral constructs that observers abstract from the behavior. The development of BSP models has enabled the estimation of behavioral constructs such as a child's level of engagement as it varies over time (Gupta et al., 2016), affective synchrony between child and parent (Hammal, Cohn, & Messinger, 2015), and affective synchrony between child and psychologist (Bone, Lee, Potamianos, & Narayanan, 2014). These models provide the potential to estimate abstract behavioral constructs that provide reliable and clinically meaningful measures of social interaction.

## Conclusions

The collaboration of engineers, speech scientists, and clinicians has resulted in (a) the development of biologically inspired technology that has been proven useful for both small- and large-scale analyses, (b) a deepened practical and theoretical understanding of both typical and impaired speech production, and (c) the establishment and enhancement of diagnostic and therapeutic tools. Although important successes have occurred for these outcomes, collaborative work of this nature inevitably presents challenges. The technological advancements made through this collaborative work may not always be adapted in the field, due to limitations of the technology, itself, or the populations being investigated. Further, questionable validity of experimental findings based on speech technology, due to the use of relatively small sample sizes and the inherently high degree of variability in speech production, particularly among clinical populations, is of concern to engineers, speech scientists, and clinicians alike. Similarly, the feasibility with which developments in speech technology can be practically implemented by clinicians and patients for diagnostic and therapeutic purposes must be seriously considered during creation and refinement of these tools. Collaborative work to resolve these issues has resulted in technologies allowing for high-speed, automatic analysis of large data sets (e.g., Eyben, Wöllmer, & Schuller, 2010) and relatively low-cost computer and mobile device–based applications for clinical interventions that can be practically implemented (e.g., Lingwaves Theravox, 2014; McAllister Byun et al., 2017; Mehta, Zañartu, Feng, Cheyne, & Hillman, 2012). The importance of successfully addressing these and other challenges serves as a call for further collaboration of engineers, speech scientists, and clinicians in the continued development of speech technology and its application to speech data science, with the potential to have far-reaching interdisciplinary significance.

## Acknowledgments

Research reported in this publication was supported in part by the National Institute on Deafness and Other Communication Disorders R01 DC007124-01, by grant 151454 from the National Science Foundation and by the Simons Foundation (Shrikanth Narayanan, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the other granting agencies. The authors thank the Speech Production and Articulation kNowledge group at the University of Southern California for their involvement in this work and the Special Interest Group 19 of the American Speech-Language-Hearing Association for their invitation. This article is an outgrowth of the Willard Zemlin Lecture presented at the 2017 Annual Convention of the American Speech-Language-Hearing Association.

## References

**American Psychiatric Association.** (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

**Bargmann, C. I., & Gilliam, C. T.** (2013). Genes and behavior. In E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, & A. J. Hudspeth (Eds.), *Principles of neural science* (5th ed., pp. 39–65). New York, NY: McGraw-Hill.

**Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S.** (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *The Journal of Child Psychology and Psychiatry, 57*(8), 927–937. https://doi.org/10.1111/jcpp.12559

**Bone, D., Black, M. P., Ramakrishna, A., Grossman, R., & Narayanan, S.** (2015). Acoustic–prosodic correlates of "awkward" prosody in story retellings from adolescents with autism. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015, Dresden, Germany,* 1616–1620. https://www.isca-speech.org/archive/interspeech_2015/papers/i15_1616.pdf

**Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S.** (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders, 45*(5), 1121–1136. https://doi.org/10.1007/s10803-014-2268-6

**Bone, D., Lee, C.-C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S.** (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research, 57*(4), 1162–1177. https://doi.org/10.1044/2014_JSLHR-S-13-0062

**Bone, D., Lee, C.-C., Chaspari, T., Gibson, J., & Narayanan, S.** (2017). Signal processing and machine learning for mental health research and clinical applications [Perspectives]. *The Institute of Electrical and Electronics Engineers Signal Processing Magazine, 34*(5), 195–196. https://doi.org/10.1109/MSP.2017.2718581

**Bone, D., Lee, C.-C., Potamianos, A., & Narayanan, S.** (2014, September). An investigation of vocal arousal dynamics in child–psychologist interactions using synchrony measures and a conversation-based model. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Singapore,* 218–222. https://www.isca-speech.org/archive/interspeech_2014/i14_0218.html

**Bone, D., Mertens, J., Zane, E., Lee, S., Narayanan, S. S., & Grossman, R.** (2017). Acoustic–prosodic and physiological response to stressful interactions in children with autism spectrum disorder. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden,* 147–151. https://doi.org/10.21437/Interspeech.2017-179

**Bresch, E., Kim, Y. C., Nayak, K., Byrd, D., & Narayanan, S.** (2008). Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *The Institute of Electrical and Electronics Engineers Signal Processing Magazine, 25*(3), 123–132.

**Bresch, E., & Narayanan, S. S.** (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *The Institute of Electrical and Electronics Engineers Transactions on Medical Imaging, 28*(3), 323–338.

**Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S.** (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America, 120*(4), 1791–1794.

**Byrd, D., Tobin, S., Bresch, E., & Narayanan, S.** (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics, 37*(1), 97–110.

**Carignan, C., Shosted, R., Fu, M., Liang, Z., & Sutton, B.** (2015). A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *Journal of Phonetics, 50,* 34–51.

**Eyben, F., Wöllmer, M., & Schuller, B.** (2010). OpenSMILE—The Munich versatile and fast open-source audio feature extractor. *Proceedings ACM Multimedia (MM), ACM* (pp. 1459–1462). Florence, Italy: Associated Colleges of the Midwest.

**Feng, X., Blemker, S., Inouye, J., Pelland, C., Zhao, L., & Meyer, C.** (2018). Assessment of velopharyngeal function with dual-planar high-resolution real-time spiral dynamic MRI. *Magnetic Resonance in Medicine, 80*(4), 1467–1474.

**Fernández-Baena, A., Susín, A., & Lligadas, X.** (2012). Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments. *Proceedings of the 2012 4th International Conference on Intelligent Networking and Collaborative Systems, INCoS* (pp. 656–661). https://doi.org/10.1109/iNCoS.2012.66

**Gelfer, M. P.** (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice, 2*(4), 320–326. https://doi.org/10.1016/S0892-1997(88)80024-9

**Geschwind, D. H.** (2011). Genetics of autism spectrum disorders. *Trends in Cognitive Sciences, 15*(9), 409–416. https://doi.org/10.1016/j.tics.2011.07.003

**Guha, T., Yang, Z., Grossman, R. B., & Narayanan, S. S.** (2018). A computational study of expressive facial dynamics in children with autism. *The Institute of Electrical and Electronics Engineers Transactions on Affective Computing, 9*(1), 14–20. https://doi.org/10.1109/TAFFC.2016.2578316

**Guha, T., Yang, Z., Ramakrishna, A., Grossman, R. B., Darren, H., Lee, S., & Narayanan, S. S.** (2015). On quantifying facial expression-related atypicality of children with autism spectrum disorder. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (p. 803). https://doi.org/10.1109/ICASSP.2015.7178080

**Gupta, R., Bone, D., Lee, S., & Narayanan, S.** (2016). Analysis of engagement behavior in children during dyadic interactions using prosodic cues. *Computer Speech and Language, 37,* 47–66. https://doi.org/10.1016/j.csl.2015.09.003

**Hagedorn, C., Kim, K., Zu, Y., Goldstein, L., Sinha, U., & Narayanan, S.** (2017). *Quantifying lingual flexibility in post-glossectomy speech*. American Speech-Language-Hearing Association Convention, Los Angeles, CA.

**Hagedorn, C., Lammert, A., Bassily, M., Zu, Y., Sinha, U., Goldstein, L., & Narayanan, S.** (2014). *Characterizing post-glossectomy speech using real-time MRI*. Paper presented at the International Seminar on Speech Production, Cologne, Germany.

**Hagedorn, C., Proctor, M., Goldstein, L., Wilson, S., Miller, B., Gorno-Tempini, M. L., & Narayanan, S. S.** (2017). Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research, 60*(4), 877–891.

**Halberstam, B.** (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *Journal for Oto-Rhino-Laryngology, 66*(2), 70–73. https://doi.org/10.1159/000077798

**Hammal, Z., Cohn, J. F., & Messinger, D. S.** (2015). Head movement dynamics during play and perturbed mother-infant interaction. *The Institute of Electrical and Electronics Engineers Transactions on Affective Computing, 6*(4), 361–370. https://doi.org/10.1109/TAFFC.2015.2422702

**Hillenbrand, J., Cleveland, R. A., & Erickson, R. L.** (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769

**Hubbard, K., & Trauner, D. A.** (2007). Intonation and emotion in autistic spectrum disorders. *Journal of Psycholinguistic Research, 36*(2), 159–173.

**Jeni, L. A., Cohn, J. F., & Kanade, T.** (2015). *Dense 3D face alignment from 2D videos in real-time*. In 2015 11th Institute of Electrical and Electronics Engineers International Conference and Workshops on Automatic Face and Gesture Recognition (FG). https://doi.org/10.1109/FG.2015.7163142

**Jones, W., Carr, K., & Klin, A.** (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry, 65*(8), 946–954.

**Kim, J., Kumar, N., Lee, S., & Narayanan, S. S.** (2014). *Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data*. Paper presented at the International Seminar on Speech Production (ISSP), Cologne, Germany.

**Kim, Y., Narayanan, S. S., & Nayak, K.** (2009). Accelerated 3D upper airway MRI using compressed sensing. *Magnetic Resonance in Medicine, 61*(6), 1434–1440.

**Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D.** (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*(9), 809–816.

**Klin, A., Lin, D. J., Gorrindo, P., Ramsay, G., & Jones, W.** (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature, 459*(7244), 257–261.

**Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S.** (1993). Perceptual evaluation of voice quality. *Journal of Speech and Hearing Research, 36*(1), 21–40. https://doi.org/10.1044/jshr.3601.21

**Lammert, A., Goldstein, L., Ramanarayanan, V., & Narayanan, S. S.** (2015). Gestural control in the English past-tense suffix: An articulatory study using real-time MRI. *Phonetica, 71*(4), 229–248.

**Lander-Portnoy, M., Goldstein, L., & Narayanan, S.** (2017). Using real time magnetic resonance imaging to measure changes in articulatory behavior due to partial glossectomy. *The Journal of the Acoustical Society of America, 142*(4), 2641–2642.

**Lee, S., Potamianos, A., & Narayanan, S.** (2014). Developmental acoustic study of American English diphthongs. *The Journal of the Acoustical Society of America, 136*(4), 1880–1894.

**Lee, Y., Goldstein, L., & Narayanan, S. S.** (2015). Systematic variation in the articulation of the Korean liquid across prosodic positions. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2015), Glasgow, UK.*

**Lingala, S., Zhu, Y., Kim, Y., Toutios, A., Narayanan, S. S., & Nayak, K.** (2017). A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magnetic Resonance in Medicine, 77*(1), 112–125.

**Lingwaves Theravox.** (2014). WEBOSYS [Computer tool]. Forchheim, Germany: WEBOSYS.

**Martin, K. B., Hammal, Z., Ren, G., Cohn, J. F., Cassell, J., Ogihara, M., . . . Messinger, D. S.** (2018). Objective measurement of head movement differences in children with and without autism spectrum disorder. *Molecular Autism, 9*(1), 14. https://doi.org/10.1186/s13229-018-0198-4

**McAllister, A., Sundberg, J., & Hibi, S. R.** (1998). Acoustic measurements and perceptual evaluation of hoarseness in children's voices. *Logopedics Phoniatrics Vocology, 23*(1), 27–38. https://doi.org/10.1080/140154398434310-1

**McAllister Byun, T., Campbell, H., Carey, H., Liang, W., Park, T. H., & Svirsky, M.** (2017). Enhancing intervention for residual rhotic errors via app-delivered biofeedback: A case study. *Journal of Speech, Language, and Hearing Research, 60*(6S), 1810–1817.

**McCann, J., & Peppé, S.** (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders, 38*(4), 325–350.

**McMicken, B., Salles, F., Von Berg, S., Vento-Wilson, M., Rogers, K., Toutios, A., & Narayanan, S. S.** (2017). Bilabial substitution patterns during consonant production in a case of congenital aglossia. *Journal of Communication Disorders, Deaf Studies and Hearing Aids, 5*(2), 1–6.

**Mehta, D., Zañartu, M., Feng, S., Cheyne, H., II, & Hillman, R.** (2012). Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *The Institute of Electrical and Electronics Engineers Transactions on Biomedical Engineering, 59*(11), 3090–3096.

**Metallinou, A., Grossman, R. B., & Narayanan, S.** (2013). Quantifying atypicality in affective facial expressions of children with autism spectrum disorders. *Proceedings of the IEEE International Conference on Multimedia and Expo, San Jose, CA, USA, Sept. 26, 2013* (pp. 1–6). https://doi.org/10.1109/ICME.2013.6607640

**Narayanan, S. S., Alwan, A. A., & Haker, K.** (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 98*(3), 1325–1347.

**Narayanan, S. S., & Georgiou, P. G.** (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the Institute of Electrical and Electronics Engineers, 101*(5), 1203–1233. https://doi.org/10.1109/JPROC.2012.2236291

**Narayanan, S. S., Nayak, K., Lee, S., Sethy, A., & Byrd, D.** (2004). An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America, 115*(4), 1771–1776.

**Narayanan, S. S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., . . . Proctor, M.** (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America, 136*(3), 1307–1311.

**Nayak, K., & Hu, B.** (2005). The future of real-time cardiac magnetic resonance imaging. *Current Cardiology Reports, 7,* 45–51.

**Perry, J., Kuehn, D., Sutton, B., & Fang, X.** (2017). Velopharyngeal structural and functional assessment of speech in young children using dynamic magnetic resonance imaging. *The Cleft Palate–Craniofacial Journal, 54*(4), 408–422.

**Proctor, M., Bone, D., Katsamanis, A., & Narayanan, S. S.** (2010). *Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis*. Paper presented at the INTERSPEECH 2010 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan.

**Proctor, M., Lo, C., & Narayanan, S. S.** (2015). Articulation of English vowels in running speech: A real-time MRI study. *Proceedings of International Congress of Phonetic Sciences (ICPhS 2015), Glasgow, UK.*

**Shic, F., Macari, S., & Chawarska, K.** (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry, 75*(3), 231–237. https://doi.org/10.1016/j.biopsych.2013.07.009

**Stone, M., Davis, E. P., Douglas, A. S., NessAiver, M., Gullapalli, R., Levine, W. S., & Lundberg, A.** (2001). Modeling the motion of the internal tongue from tagged cine-MRI images. *The Journal of the Acoustical Society of America, 109*(6), 2974–2982.

**Story, B. H., Titze, I. R., & Hoffman, E. A.** (1996). Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America, 100*(1), 537–554.

**Takemoto, H., Honda, K., Masaki, S., Shimada, Y., & Fujimoto, I.** (2006). Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *The Journal of the Acoustical Society of America, 119*(2), 1037–1049.

**Tick, B., Bolton, P., Happé, F., Rutter, M., & Rijsdijk, F.** (2016). Heritability of autism spectrum disorders: A meta-analysis of twin studies. *The Journal of Child Psychology and Psychiatry, 57*(5), 585–595. https://doi.org/10.1111/jcpp.12499

**Toutios, A., Byrd, D., Goldstein, L., & Narayanan, S.** (2017). Articulatory compensation strategies employed by an aglossic speaker. *The Journal of the Acoustical Society of America, 142*(4), 2639.

**Toutios, A., Lingala, S., Vaz, C., Kim, J., Esling, J., Keating, P., . . . Narayanan, S.** (2016). Illustrating the production of the International Phonetic Alphabet sounds using fast real-time magnetic resonance imaging. *Proceedings of Interspeech 2016,* 2428–2432.

**Vaz, C., Ramanarayanan, V., & Narayanan, S.** (2018). Acoustic denoising using dictionary learning with spectral and temporal regularization. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 26*(5), 967–680.

**Wagner, P., Malisz, Z., & Kopp, S.** (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57,* 209–232.

**Windolf, M., Götzen, N., & Morlock, M.** (2008). Systematic accuracy and precision analysis of video motion capturing systems-exemplified on the Vicon-460 system. *Journal of Biomechanics, 41*(12), 2776–2780. https://doi.org/10.1016/j.jbiomech.2008.06.024

**Zhang, Z., Gersdorff, N., & Frahm, J.** (2011). Real-time magnetic resonance imaging of temporomandibular joint dynamics. *The Open Medical Imaging Journal, 5,* 1–7.