A procedure for estimating gestural scores from speech acoustics

Hosung Nam[a)]
Haskins Laboratories, 300 George Street, Suite 900, New Haven, Connecticut 06511

Vikramjit Mitra
Speech Technology and Research Laboratory, SRI International, 333 Ravenswood Avenue, Menlo Park, California 94025

Mark Tiede
Haskins Laboratories, 300 George Street, Suite 900, New Haven, Connecticut 06511

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering, Beckman Institute 2011, University of Illinois, Urbana, Illinois 61801

Carol Espy-Wilson
Speech Communication Laboratory, Institute for Systems Research and Department of Electrical and Computer Engineering, A.V. Williams 2205, University of Maryland, College Park, Maryland 20742

Elliot Saltzman
Department of Physical Therapy and Athletic Training, 635 Commonwealth Avenue, Boston University, Boston, Massachusetts 02215

Louis Goldstein
Department of Linguistics, University of Southern California, Grace Ford Salvatori 301, Los Angeles, California 90089

(Received 8 January 2011; revised 24 September 2012; accepted 28 September 2012)

Speech can be represented as a constellation of constricting vocal tract actions called gestures, whose temporal patterning with respect to one another is expressed in a gestural score. Current speech datasets do not come with gestural annotation and no formal gestural annotation procedure exists at present. This paper describes an iterative analysis-by-synthesis landmark-based time-warping architecture to perform gestural annotation of natural speech. For a given utterance, the Haskins Laboratories Task Dynamics and Application (TADA) model is employed to generate a corresponding prototype gestural score. The gestural score is temporally optimized through an iterative timing-warping process such that the acoustic distance between the original and TADA-synthesized speech is minimized. This paper demonstrates that the proposed iterative approach is superior to conventional acoustically-referenced dynamic timing-warping procedures and provides reliable gestural annotation for speech datasets. © *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4763545]

PACS number(s): 43.72.Ar, 43.72.Ct, 43.70.Bk [DAB] Pages: 3980–3989

I. INTRODUCTION

Several recent studies have suggested that articulatory gestures can be used as an alternative to non-overlapping phone units (e.g., diphones) for more robust automatic speech recognition (ASR), because they can effectively model the effects on coarticulation of factors such as varying prosodic phrasing (e.g., Sun and Deng, 2002; Zhuang *et al.*, 2009; Mitra *et al.*, 2010b; Mitra *et al.*, 2011). Articulatory phonology (Browman and Goldstein, 1992) treats each word as a constellation of vocal-tract constriction actions, called *gestures* (roughly 1 to 3 gestures for each of the phones in a phonetic transcription). Each gesture is viewed as a dynamical system that controls one of the constricting devices (end-effectors) of the vocal tract: LIPS, tongue tip (TT), tongue body (TB), velum (VEL), and glottis (GLO). The gestural goals, or targets, for the constrictions of these end-effectors are defined in the set of tract variables (TVs), and each TV has its own set of associated articulators. Table I presents the constricting organs (end-effectors) and the associated vocal TVs and Fig. 1 shows how the variables are geometrically defined in the vocal tract. The 8 TVs [lip protrusion (LP), lip aperture (LA), tongue tip constriction location (TTCL), tongue tip constriction degree (TTCD), tongue body constriction location (TBCL), tongue body constriction degree (TBCD), VEL, and GLO] in Table I and Fig. 1 are the set of task-specific coordinates for characterizing the shape of the vocal tract tube in terms of constriction degrees and locations along the TV dimensions, and the kinematic trajectories of the TVs are the outcomes of constriction gestural activation.

[a)]Author to whom correspondence should be addressed. Electronic mail: nam@haskins.yale.edu

TABLE I. Constriction organs and their vocal TVs.

Constriction organs	Vocal TVs
Lip	LA
	LP
TT	TTCD
	TTCL
TB	TBCD
	TBCL
VEL	Velic opening degree (VEL)
GLO	Glottal opening degree (GLO)

For example, the /b/ in "tub" corresponds to a constriction gesture in the LA TV. Each gesture is specified for the activation interval, i.e., where in time it is active, and the dynamic parameters of gestural *target* and *stiffness*. The targets are defined in millimeters, degrees, or arbitrary units. The targets of LP, LA, TTCD, and TBCD gestures are defined in millimeters, those of TTCL and TBCL gestures in degrees, and those of VEL and GLO gestures in arbitrary units. The targets of LP are the horizontal location of the lips and those of LA, TTCD, and TBCD are the constriction degree of the constriction organs. The targets of TTCL and TBCL are defined using a polar grid, ranging from 0° to 180° as shown in Fig. 1, in which 0° is in front of the chin, 90° at the center of the hard palate, and 180° is in the center of the pharynx. The stiffness of a gesture determines how fast the specified target is achieved. It is known that consonant gestures achieve their targets more quickly than vowels (Perkell, 1969, Fowler, 1980). Stiffness can distinguish consonants from vowels. The target of the LA gesture for /b/ is a complete constriction of the lips, defined by a −2 mm aperture target, indicating compression of the lips. Available data suggest that a labial stop release gesture requires 50 to 100 msec, thus the stiffness can be parameterized by a resonant frequency of 8 Hz (8 closure-release cycles per second); that of a vowel is set to 4 Hz. Consonants are defined by the lips (LP and LA), tongue tip (TTCL and TTCD), or tongue body (TBCD and TBCL) gestures, unrounded vowels are defined by the tongue body (TBCL and TBCD) gestures only, and rounded vowels are also associated with LA gestures as well as TB gestures. Vowels are distinguished from one another by the targets of TBCL and TBCD gestures. For example, /i/ and /I/ are distinguished by different target values of TBCD (5 mm for /i/ and 8 mm for /I/) although their TBCL values are both 95°. Both /i/ and /I/ are distinguished from /ɑ/ with a TBCL target of 180°.

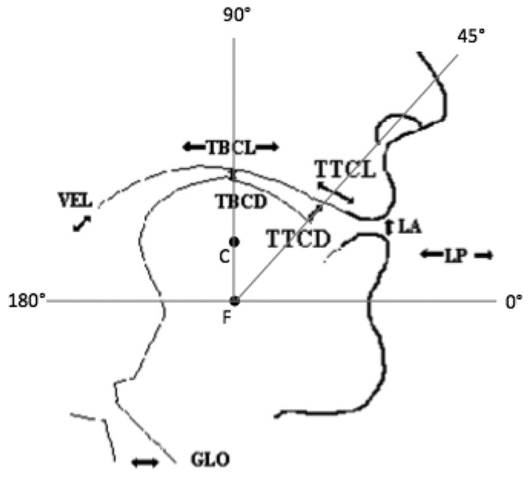

FIG. 1. Vocal TVs at five distinct constriction organs.

Each word is represented as an ensemble of these distinctive gestures coordinated in time with respect to one another in the form of a *gestural score*. Gestural coordination patterns exhibit both temporal overlap and sequential dependence, and gestures can change their relative timing and their magnitudes as a function of factors such as syllable position, lexical stress, prosodic stress, and the strength of prosodic phrasal boundaries (cf., Browman and Goldstein, 1995; Byrd, 1995; Byrd and Saltzman, 1998; Byrd *et al.*, 2009; Cho, 2005; de Jong, 1995; Fougeron and Keating, 1997; Kochetov, 2006; Krakow, 1999; Turk and Sawusch, 1997). Figure 2 shows an example gestural score for the word "span." All the TVs but LP have at least one active gesture. The gestures are coordinated in time appropriately to produce the word. The consonant /s/ involves a pair of TT gestures for TTCL and TTCD, which are temporally coupled. The target of TTCD is critically narrow (1 mm), enough to produce turbulence. A GLO gesture co-occurs with them to make the sound voiceless. The coda consonant /n/ has the same TT gesture pair as /s/ but the target of TTCD is −2 mm which produces a complete closure in TT that is substantially overlapped with a VEL gesture. The vowel /æ/ comprises a pair of TB gestures, TBCL and TBCD, whose targets are specified by numerical values.

Studies have shown that articulatory gestures can be used as an alternative to phones for ASR systems, providing a set of basic units that can effectively model coarticulation in speech (e.g., Sun and Deng, 2002; Zhuang *et al.*, 2009; Mitra *et al.*, 2010b). Unfortunately, because no large natural speech database currently exists that includes transcribed

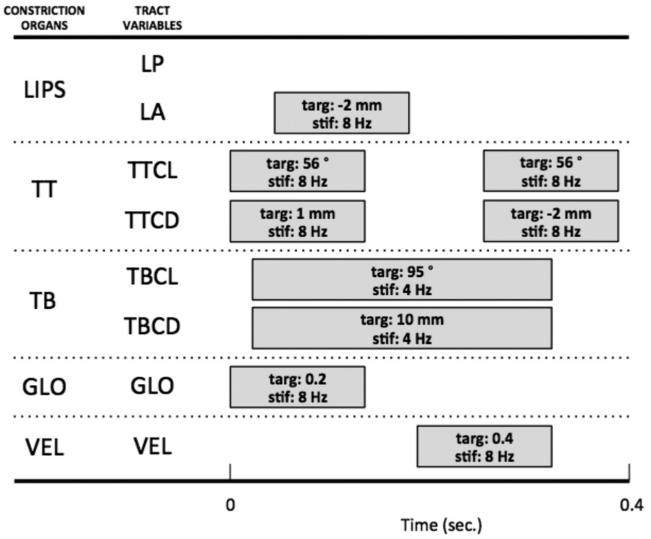

FIG. 2. Gestural score for the word span. Constriction organs and vocal TVs are denoted in the left-most two columns. The gray boxes to the right represent the corresponding gestural activation intervals and parameter values for target and stiffness.

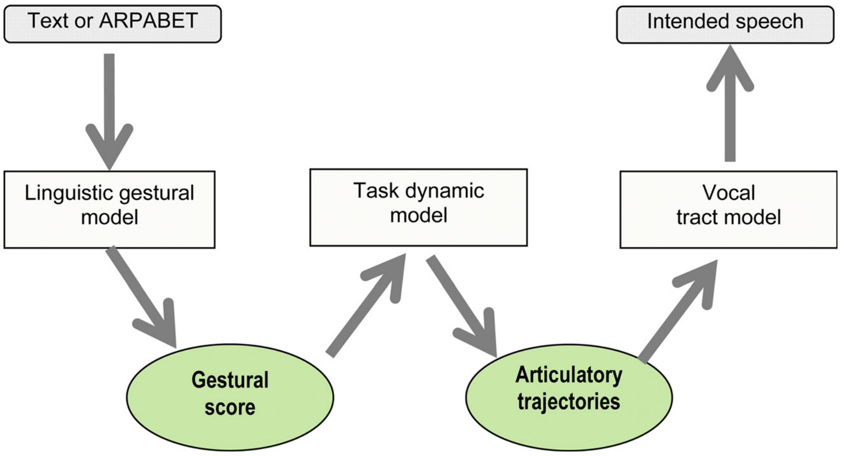

FIG. 3. (Color online) Flow diagram of TADA.

gestural scores, previous studies have been largely limited to using synthetic data for training, namely, that generated by the Haskins Laboratories Task Dynamics and Application model of speech production, also known as TADA (Nam *et al.*, 2004). TADA is a computational implementation of articulatory phonology (Fig. 3) using task dynamics (Saltzman and Munhall, 1989). For a given utterance, the corresponding text (e.g., "bad") or ARPABET string (e.g., BAED), which is a conventional phonetic transcription, is first input to the linguistic gestural model, which determines the gestural score for the utterance, via a segment-to-gesture dictionary and a set of syllable-based inter-gestural coupling (phasing) principles. Given a gestural score, the task-dynamic model computes the TV and articulator kinematics, a vocal tract model (CASY: The Haskins Configurable Articulatory Synthesizer; see Rubin *et al.*, 1996) computes a time-varying area function and formant frequencies. Employing the TV and the formant frequencies, HLsyn, a quasi-articulatory speech synthesizer (Hanson and Stevens, 2002), calculates the corresponding acoustic output. In particular, the constriction degrees near the front of the vocal tract (LA and TTCD) are used for detailed spectra.

Using a synthetic speech corpus with TV and gestural score annotation, we have shown that: (1) Gestures and TVs can be reliably estimated from acoustics (Mitra *et al.*, 2010a,b); (2) estimated gestural scores from TV trajectories produce a word recognition accuracy of around 91% (Hu *et al.*, 2010); and (3) gestures and TVs can potentially improve the noise robustness of ASR systems (Mitra *et al.*, 2010b).

Annotating a large natural speech database with gestural score specifications would have benefits not just for speech technology but also in such related fields as phonological theory, phonetic science, speech pathology, etc. Several efforts have been made to obtain gestural information from the speech signal. A temporal decomposition method was proposed by Atal (1983) for estimating functions similar to gestural activations from the acoustic signal but that method was limited to gestural activation functions, not the associated dynamic parameters. Jung *et al.* (1996) also used temporal decomposition, with which they were able to successfully retrieve gestural parameters such as constriction targets, assuming prior knowledge of articulator records (the time functions of flesh-point pellets). Sun *et al.* (2000) proposed a semi-automatic annotation model of gestural scores that required manual annotation of gestures to train the model; in practice because of the difficulties discussed below, researchers since 2000 have focused on methods that avoid the need for manual gestural annotation of large datasets. Zhuang *et al.* (2009) and Mitra *et al.* (2010b) showed that gestural activation intervals and dynamic parameters such as target and stiffness could be estimated from TVs using a TADA-generated synthetic database. Tepperman *et al.* (2009) used an hidden Markov model (HMM)–based iterative bootstrapping method to estimate gestural scores from acoustics for a small dataset. Despite these efforts, there is as yet no gesturally-labeled speech database sufficiently large for adequate training of a speech recognizer (e.g., nothing comparable in size to the TIMIT phonetically-labeled corpus).

The experiments in this paper demonstrate that it is possible to refine a candidate gestural score by defining a correspondence between consonant closure or consonant release *gestures*, on the one hand, and the *acoustic landmarks* of consonant closure or consonant release, on the other hand. According to Shannon and Bode (1950), the information content of a signal (measured in bits per sample) can be quantified in terms of its innovation, defined to be the difference between the observed value of the signal and its optimal prediction. During quasi-static intervals, the power spectrum of a speech signal is predictable; the bit rate required to encode speech information increases primarily at the consonant release and consonant closure landmarks. It has been shown that human listeners derive more information per unit time from transition regions (consonant-vowel boundaries) than from steady-state regions (Furui, 1986), and early versions of the theory of landmark-based speech recognition focused on the role of acoustic boundaries in speech perception (Stevens, 1985). Although transitions have the highest information density, they are not as perceptually salient or developmentally fundamental as syllable nuclei (Mehler *et al.*, 1988; Jusczyk, 1993); therefore modern theories of landmark-based speech recognition (Stevens *et al.*, 1992) and implemented landmark-based speech recognition systems (Juneja and Espy-Wilson, 2008) generally comprise four types of landmarks: Consonant release landmarks, consonant closure landmarks, vowel landmarks, and inter-syllabic glide landmarks. This paper focuses only on consonant release and consonant closure landmarks because vowel landmarks, though perceptually salient, are difficult to localize precisely in time, e.g., there are contexts in which neither the energy peak nor the first formant peak is a reliable marker of the vowel gesture time alignment (Mermelstein, 1975; Howitt, 2000). It is possible that the results presented in this paper could be improved by the use of a high-accuracy algorithm for detection and alignment of vowel landmarks.

II. DATABASE

Our proposed architecture assumes only that the natural speech database upon which it will be implemented has the phones delimited in advance; using HMM forced alignment (Yuan and Liberman, 2008) it is possible to perform this acoustic segmentation on any speech database. For our study, we chose the University of Wisconsin x-ray microbeam (XRMB) database (Westbury, 1994). In addition to the acoustics, the XRMB database includes the time functions of flesh-point pellets tracked during speech production which allows us to cross-validate our approach by verifying the articulatory as well as the acoustic information synthesized in our analysis-by-synthesis (ABS) loop. The XRMB database includes speech utterances recorded from 47 different American English speakers, 25 of whom are females and 22 who are males. Each speaker produced up to 56 short speech reading examples ("tasks") including a series of digits, sentences from the TIMIT corpus, and entire paragraphs. The sampling rate for the acoustic signals is 21.74 kHz.

III. ARCHITECTURE: GESTURAL ANNOTATION

For our study, the acoustic data for the XRMB utterances were word- and phone-delimited by using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). Consider a natural speech utterance represented by a set of time-indexed acoustic feature vectors, $\boldsymbol{S} = [\vec{s}_1, \ldots, \vec{s}_T]$, and transcribed as a sequence of phones $\Phi = [\phi_1, \ldots, \phi_L]$ whose phone boundary times $\psi = [\psi_1, \ldots, \psi_{L-1}]$ are labeled, such that phone ϕ_l is the label that extends from frame $\psi_{l\text{-}1} + 1$ to frame ψ_l, with utterance start and end times given by $\psi_0 \equiv 0$ and $\psi_0 \equiv T$. A gestural pattern vector (GPV), $\vec{g}_m$, is a list of simultaneously active gestures, specifying target constriction location (CL), constriction degree (CD), and stiffness of one or more TVs. A gestural score can be written as a sequence of abutting GPVs, $\boldsymbol{G} = [\vec{g}_1, \ldots, \vec{g}_M]$, and a set of corresponding boundary times $\boldsymbol{B} = [b_1, \ldots, b_{M-1}]$, such that $\vec{g}_m$ lists all of the gestures active from frame number $b_{m\text{-}1} + 1$ through frame number b_m. Figure 4 illustrates how a GPV is defined in the gestural score for the word span.

The TADA synthesizer is used to produce a synthesized speech signal $\hat{\boldsymbol{S}}$ using the sequence of steps shown in Fig. 3. First, the ARPABET transcription of a phone structure $\{\Phi, \psi\}$ for each word is explicitly parsed into syllables by means of a syllabification algorithm based on the English phonotactics principle of the maximization of syllable onsets (Goldwater and Johnson, 2005). Then, from the syllabified ARPABET transcription, the linguistic gestural model generates a synthesized gestural score, $\{\hat{\boldsymbol{G}}, \hat{\boldsymbol{B}}\}$, specifying the CL and CD targets and stiffnesses for each TV, and the times during which those targets are in force. Central to this model is a gestural dictionary that specifies the ensemble of gestures corresponding to a syllable's phones, and a set of temporal coupling principles (Goldstein *et al.*, 2006; Nam, 2007; Saltzman *et al.*, 2008) that couple the syllable's gestures to one another in time. Gestures at the margins of neighboring syllables and words are also coupled to each other to coordinate the syllables in time.

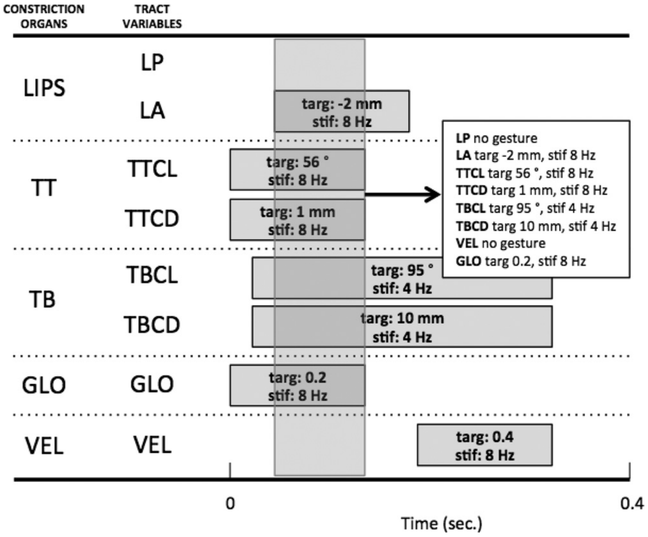

FIG. 4. The definition of a GPV in a gestural score. A GPV is defined by the list of gestures that are continuously active over a given sequence of time-frames of the gestural score. Using the gestural score for the word span from Fig. 2, a GPV is shown as the semi-transparent rectangular box spanning the interval during which gestures are simultaneously active in the LIPS, TT, TB, and GLO; the set of TVs and target and stiffness values associated with the GPV is indicated to the right of the box.

Second, $\{\hat{\boldsymbol{G}}, \hat{\boldsymbol{B}}\}$ is input to a task dynamic model (Saltzman and Munhall, 1989). The model implements the gestures as a set of second-order critically damped systems governing the dynamics of the TVs and generates TV time functions $\hat{\boldsymbol{V}} = [\hat{v}_1, \ldots, \hat{v}_T]$, where $\vec{v}_t$ is an eight-dimensional vector specifying the values, at time t, of the eight variables listed in Table I, and coordinated motion patterns for the system's model articulators (e.g., jaw, upper lip, and lower lip, etc.). The vocal tract model (CASY) and HLsyn then convert the synthesized TVs, $\hat{\boldsymbol{V}}$, into a synthesized acoustic signal, $\hat{\boldsymbol{S}}$.

One of the core hypotheses of articulatory phonology is that each word corresponds to a set of gestures that is invariant within the language community (except for large-scale sociolect and dialect shift phenomena), and that the routine pronunciation variability of spontaneous speech is caused not by changes in gestural composition but by variation in gestural timing (Browman and Goldstein, 1992) that can result in changes in the amount of temporal overlap between pairs of gestures and in spatial reduction due to reduced activation times. The algorithm developed here makes the strong assumption that the dynamical parameters (target and stiffness) of gestures do not change from instance to instance, only the durations and relative timings of gesture activation are allowed to vary. In that sense, the procedure developed here is best thought of as analogous to forced alignment of a phonetic transcription, in which the parameters of the units do not change, only their temporal boundaries. However, because of the nature of the mapping from gestural dynamical parameters to articulatory movement to sound, changing these durations will have complex effects on the acoustics, as discussed further below. The hypothesis that gestural parameters do not change is, of course, overly strong. There is good evidence (for example, Byrd and Saltzman, 2003; Cho, 2006) that dynamical parameters of gestures may be effectively influenced by prosody. However, the goal of the present work is not to develop a complete, optimal gestural model of an utterance but rather a transcription indicating the temporal intervals during which the phonological gesture units are likely to be active. The ability to transcribe a large database in this way constitutes a new source of knowledge about the temporal regularities of gestural structure, and could also lead to an automated investigation of dynamical parameter variation that is not itself being addressed here. For example, the gestural transcription of the microbeam database will allow automatic analysis of the articulatory kinematics of a particular gesture type as a function of various contextual variables.

In order to make automatic gestural annotation computationally tractable, we introduced a further constraint, namely, any gestures that overlap at all in the canonical pronunciation remain overlapped in every reduced pronunciation, and vice versa. The result is that the sequence of distinct GPVs, $\boldsymbol{G} = \hat{\boldsymbol{G}}$, is invariant, and all pronunciation variability must be explained by variation in the gestural boundary times, $\hat{\boldsymbol{B}}$. As the boundary times vary, the amount of times two gestures overlap can vary; however, any changes of boundary times associated with the estimation process are not allowed to remove (or introduce) intergestural overlap intervals that are present (or absent) in the prototype gestural scores.[1] Additionally, gestural activation durations are allowed to vary with the estimation process; significant decreases in activation duration can result in undershoot, even to the point of apparently deleting phones or syllables. The goal of ABS gestural annotation, therefore, is to find the boundary times $\hat{\boldsymbol{B}}$ such that the resulting synthesized speech signal $\hat{\boldsymbol{S}}$ matches the observed signal $\boldsymbol{S}$ with minimum error. We propose to solve this problem by using the linguistic gestural model of TADA to generate an initial set of prototype gestural boundary times, $\hat{\boldsymbol{B}}^{(0)}$ for the corresponding TVs, $\hat{\boldsymbol{V}}^{(0)}$, and acoustic signal $\hat{\boldsymbol{S}}^{(0)}$. The initial boundary times are then iteratively refined, producing synthesized outputs, $\{\hat{\boldsymbol{B}}^{(i)}, \hat{\boldsymbol{V}}^{(i)}, \hat{\boldsymbol{S}}^{(i)}\}$ with successively reduced distances $\mathrm{D}(\boldsymbol{S}, \hat{\boldsymbol{S}}^{(i)})$.

A reasonable baseline may be generated by dynamic time warping (DTW) of $\hat{\boldsymbol{S}}^{(0)} = [\hat{\boldsymbol{s}}_1^{(0)}, \ldots]$, the acoustic signal synthesized by TADA based on $\hat{\boldsymbol{B}}^{(0)} = [\hat{b}_1^{(0)}, \ldots]$. DTW computes a warping function $w(t)$ such that the warped signal $\hat{s}_{w(t)}^{(\mathrm{DTW})} = \hat{s}_t^{(0)}$ minimizes the target distance metric $\boldsymbol{D}(\boldsymbol{S}, \hat{\boldsymbol{S}}^{(\mathrm{DTW})})$ (Sakoe and Chiba, 1978). The final gestural score computed by DTW is then given by the canonical gestural score, $\hat{\boldsymbol{G}} = \boldsymbol{G}$, with its boundary times re-aligned as

$$\hat{b}_m^{(\mathrm{DTW})} = w\left(\hat{b}_m^{(0)}\right). \qquad (1)$$

DTW is limited by two important sources of mismatch between the synthesized signal and the natural speech signal. First, the synthesized signal is generated based on a standard male vocal tract model, regardless of the gender or other characteristics of the speaker who produced the natural speech utterance. Second, the signal $\hat{\boldsymbol{S}}^{(0)}$ is, by default, a careful production with considerably less phoneme reduction than one would expect in a natural utterance. DTW of the speech signal does not produce phoneme reduction. If $\hat{\boldsymbol{S}}^{(0)}$ contains a carefully pronounced stop consonant or fricative that is completely missing in the natural speech utterance, DTW has no way to implement this undershoot (or deletion) process. Therefore, it must find a time alignment that in some way forces an approximate match despite the difference in apparent phonemic content of the two signals.

We used an ABS procedure that eliminates mismatch caused by phoneme undershoot (Fig. 5). Specifically, we adapted a two-stage process in which (a) the first stage provided a coarse-grained warping to put the synthesized acoustics, $\hat{\boldsymbol{S}}^{(1)}$, into the same temporal ball park as the target acoustics, $\boldsymbol{S}$, and (b) the second stage created a series of iteratively refined boundary time vectors, $\hat{\boldsymbol{B}}^{(i)}$, for $i \in \{1, 2, \ldots\}$. Each set of boundary times, $\hat{\boldsymbol{B}}^{(i)}$ is used by TADA to generate a new set of TVs, $\hat{\boldsymbol{V}}^{(i)}$, and from them, a new synthesized speech signal $\hat{\boldsymbol{S}}^{(i)}$. The mapping from $\hat{\boldsymbol{B}}^{(i)}$ to $\hat{\boldsymbol{B}}^{(i+1)}$ is performed in a manner that guarantees $\boldsymbol{D}(\boldsymbol{S}, \hat{\boldsymbol{S}}^{(i+1)}) \le \boldsymbol{D}(\boldsymbol{S}, \hat{\boldsymbol{S}}^{(i)})$.

The first stage of the process, generating $\hat{\boldsymbol{B}}^{(1)}$ from $\hat{\boldsymbol{B}}^{(0)}$, could be performed using DTW between $\hat{\boldsymbol{S}}^{(0)}$ and $\boldsymbol{S}$, and then the resulting time-warping function to transform $\hat{\boldsymbol{B}}^{(0)}$ to $\hat{\boldsymbol{B}}^{(1)}$. However, experiments indicate that greater accuracy can be achieved by taking advantage of the relationships between the natural utterance's phone transcription, $\{\Phi, \psi\}$ and the gestural score. Although there is no one-to-one mapping between the gestural boundary times $\hat{\boldsymbol{B}}^{(i)}$ and any corresponding phone boundary times, in many cases approximate synthetic phone boundary times, $\hat{\psi}$, can be estimated from $\hat{\boldsymbol{B}}$ using a set of simple heuristics that follow from the assumption that the gestural time constants are assumed to remain fixed. For example, the acoustic onset of consonantal phones (or word-initial vocalic phones) is taken to be approximately 60 msec after the onset of the corresponding gesture.

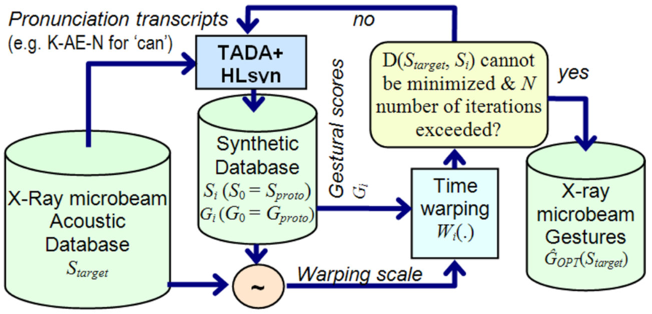

FIG. 5. (Color online) Block diagram of the overall iterative ABS warping architecture for gesture specification.

This lag captures the (gesture-independent) time required for an articulatory movement (which begins slowly) to have a measurable effect on the acoustics. For example, Mooshammer *et al*. (2012) found that in [ə-(C)V] sequences, the onset of the formant transitions between the schwa and the following segment (regardless of the manner class of the segment–stop, fricative, liquid, vowel) coincided approximately with the peak velocity of the constriction gesture associated with the segment, which in turn occurred regularly 60 to 70 msec after the gesture onset. Of course, the onset of formant transitions is not the standard acoustic landmark associated with segments in a forced alignment, and the time of those standard landmarks is expected to be dependent on consonant manner. But 60 msec is the minimum lag, and further manner-related differences are expected to emerge in the course of the iterative procedure. Similarly, the onset of vocalic phones is approximated as 40 msec after the onset of the release gesture for the preceding consonant. Utterance offsets are approximated as 60 msec after the offsets of final gestures. Taking advantage of this approximate meta-information, the approximated phone boundary times $\hat{\psi}^{(0)}$ of the synthetic utterance are warped so that $\hat{\psi}^{(1)} = \psi$. The same time-warping function is then applied to $\hat{\boldsymbol{B}}^{(0)}$ to generate $\hat{\boldsymbol{B}}^{(1)}$.

Due to possible errors in estimating phone boundaries, the boundary times $\hat{\boldsymbol{B}}^{(1)}$ may not be optimal, and therefore it is useful to iteratively refine them in the second stage of our procedure. The phone boundaries for $\hat{\psi}^{(1)}$ are individually changed in one of 5 ways [no change, ±10 msec, ±20 msec] to find an optimal warping scale. All of the 5^{L-1} possible piece-wise warpings of $\hat{\psi}^{(1)}$ are tested (where $\boldsymbol{L}$ is the number of phones), and the one whose corresponding synthesized speech signal minimizes $\boldsymbol{D}(\boldsymbol{S}, \hat{\boldsymbol{S}})$ is retained as $\hat{\psi}^{(2)}$. This procedure (piecewise phone boundary modulation and distance measure) is performed iteratively until $\boldsymbol{D}(\boldsymbol{S}, \hat{\boldsymbol{S}}^{(i)})$ is minimized. At each iteration, the algorithm is allowed to shift each phone boundary by at most 20 msec; in effect, this constraint avoids over-fitting, in much the same way that slope constraints avoid over-fitting in DTW.

Figure 6 compares the XRMB (top panel), prototype TADA (middle panel), and time-warped TADA (bottom panel) utterances for the word "seven" from task003 of XRMB speaker 11, in which each panel shows the corresponding waveform and spectrogram. Figure 6 (middle and bottom panels) also displays the gestural scores for the prototype and time-warped TADA utterances, with lips, TT, and TB gestures as gray blocks overlaid on the spectrogram showing how gestural timing is modulated by the proposed time-warping procedure.

Each iteration of the ABS procedure requires synthesizing and testing 5^{L+1} speech waveforms, where $\boldsymbol{L}$ is the number of phones in the utterance. Large values of $\boldsymbol{L}$ are not practical. In order to control complexity, all experiments in this paper perform time warping on a word-by-word basis. After the last iteration, for the purpose of evaluation to be described in Sec. IV, the obtained word-level gestural scores are concatenated to yield the utterance-level gestural score such that the final phone offset time of one word (as predicted by the estimated phone boundary sequence $\hat{\psi}^{(i)}$) is equal to the initial phone onset time of the following word. Since gesture onsets and offsets extend beyond the corresponding phone boundaries, such cross-word phone boundary concatenation will typically result in gestural overlap at the boundaries. Further, since the activation duration of gestures on either side of the boundary will vary (for example, due to prosodic factors), the overall percentage overlap of gestures across the boundary will also vary. However, the cross-word overlap is not directly optimized, and should be in future work. TADA is finally executed again on the utterance-level gestural scores to generate the final synthesized TVs and synthesized speech signal.

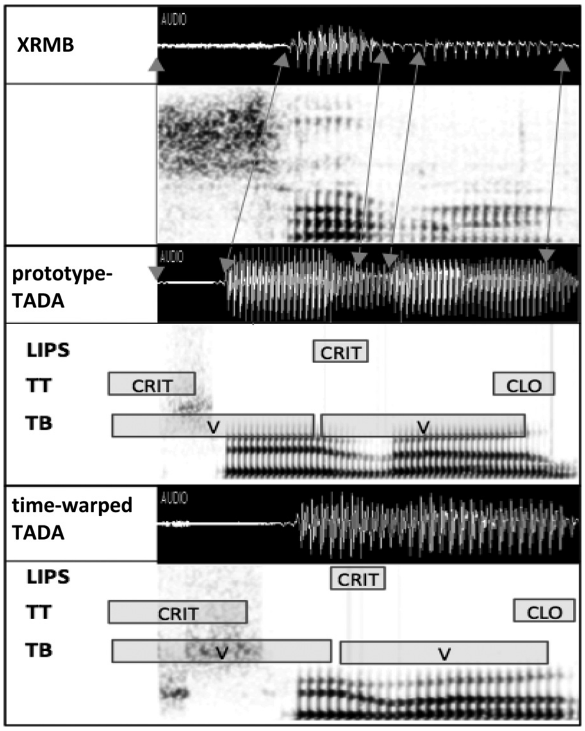

FIG. 6. Waveform and spectrogram of XRMB, prototype TADA, and time-warped TADA speech for "seven." For simplicity's sake, CL, VEL, and GLO gestures are not presented in this figure. CRIT denotes 1 mm constriction degree for the fricatives (/s/ and /v/) and CLO denotes 0 mm constriction degree for the stop /n/.

The proposed approach is independent of any articulatory information from XRMB. It is also independent of differences among talkers. Based on word and phone transcriptions, the architecture generates gestural scores and TV trajectories using the default speaker characteristics predefined in TADA. It is possible to imagine a similar procedure that would iteratively refine the talker characteristics of TADA in order to improve the match between $\boldsymbol{S}$ and $\hat{\boldsymbol{S}}$, but as noted above, the goal is not to produce an optimized gestural model of the utterances but rather an indication of the temporal span of gesture activations. And it is not clear that the map from gesture activation times to the time that consequences are observable will vary due to talker's morphology and/or voice characteristics. Of course, the overall spectral match will be better for some speakers than others. However, as we will see in Sec. IV, gestural scores generated by the proposed algorithm without the need for talker adaptation are validated by the microbeam pellet data.

IV. EVALUATION OF THE GESTURAL ANNOTATION PROCEDURE

We have implemented the proposed landmark-based ABS time-warping architecture for gestural score annotation across all 56 speech tasks from the 47 speakers of the XRMB database (however, some speakers performed only a subset of the 56 tasks). We performed two tests to evaluate our methodology.

First, we compared the proposed time-warping strategy to that of the standard acoustic DTW (Sakoe and Chiba, 1978). We used an acoustic distance measure between the XRMB natural speech, S_{target} and the TADA speech after: (1) DTW only and (2) our iterative landmark-based ABS time-warping method. We tested three types of distance metrics: The standard log-spectral distance ($\boldsymbol{D}_{\text{LSD}}$), the log-spectral distance using linear prediction spectra ($\boldsymbol{D}_{\text{LSD-LP}}$), and the Itakura distance ($\boldsymbol{D}_{\text{ITD}}$). $\boldsymbol{D}_{\text{LSD}}$ is defined in Eq. (2),

$$D_{\text{LSD}}(\boldsymbol{S},\hat{\boldsymbol{S}}) = \sum_{t=1}^{T}\sqrt{\frac{1}{N}\sum_{k=0}^{N-1}\left[10\log_{10}\frac{s_t[k]}{\hat{s}_t[k]}\right]^2}, \qquad (2)$$

where $s_t[k]$ and $\hat{s}_t[k]$ are defined to be the kth frequency bins of the magnitude short-time Fourier transforms computed at time t from the natural and synthesized speech signals, respectively. T is the number of frames for a given utterance. $\boldsymbol{D}_{\text{LSD-LP}}$ is identical to $\boldsymbol{D}_{\text{LSD}}$ except that the feature vectors are linear prediction spectra rather than short-time Fourier spectra. $\boldsymbol{D}_{\text{ITD}}$ can be defined in terms of the linear prediction spectrum, $s_t[k]$ as in Eq. (3)

$$D_{\text{ITD}}(\boldsymbol{S},\hat{\boldsymbol{S}}) = \sum_{t=1}^{T}\left(\ln\left[\frac{1}{N}\sum_{k=0}^{N-1}\frac{s_t[k]}{\hat{s}_t[k]}\right] - \frac{1}{N}\sum_{k=0}^{N-1}\ln\left[\frac{s_t[k]}{\hat{s}_t[k]}\right]\right). \qquad (3)$$

DTW can only manipulate the temporal alignment of S_{target} and the TADA speech, while manipulation of gestural activation durations will produce changes in the spectral content, as a function of changes in overlap and undershoot of gestures. If these changes in gestural score are capturing real information about the temporal patterning of gestures in these utterances, the resulting spectral changes should produce an improved match of the TADA model to S_{target} over what can be achieved by purely temporal modulation of the spectral pattern.

Twelve different speech tasks (available from all speakers) were selected randomly from the XRMB database to obtain the distance measured between the natural and synthetic speech. The distances between DTW and the iterative time-warping approach were compared over entire utterances, and the results are presented in Table II. The results confirm the hypothesis that the spectral changes induced by changes in the patterning of gestural activation is capturing significant information about the natural utterance's temporal structure. Because the spectral changes due to overlap and undershoot might be hypothesized to influence unstressed syllables more than stressed ones, and vowels more than consonants, the distance metrics were calculated separately for vowel and consonant regions under primary stress, secondary stress, and unstressed conditions. Results are presented in Tables III and IV. Consideration of the spectral distance measures ($\boldsymbol{D}_{\text{LSD}}$ and $\boldsymbol{D}_{\text{LSD-LP}}$) reveals a substantial improvement using the iterative warping method over DTW. For obstruent intervals, improvement is quite robust, roughly a factor of 2; more in some stress contexts, less in others. This is not too surprising since the clear acoustic landmarks provided by obstruent closures and releases are directly employed by the iterative warping algorithm. Perhaps more revealing is the large improvement (also approximately 2:1) shown by the stressed vowel intervals. Since vowel *gesture* onsets and offsets do not directly produce acoustic landmarks, this suggests that the TADA model is capturing some vowel-consonant gesture coordination regularities in a useful way. This is further supported by the weaker improvement shown for secondary-stressed and particularly reduced vowel intervals. The coordination generalizations captured by the TaDA model are those for stressed syllables, and it would be expected that the unstressed syllables show considerable variability in that coordination. Less improvement is found during sonorant consonant intervals but this appears to be because of how well DTW does during these intervals, rather than poor performance of iterative warping. DTW may do well because of the smooth change from vowel to consonant exhibited by sonorants.

Secondly, we evaluated how similar the TV trajectories generated from our proposed approach are compared to those derived from the recorded flesh-point (pellet) measurements available in the XRMB database. The TVs were estimated from the pellet information as follows. LA can be

TABLE II. Distance measures between the warped signal and the XRMB signal from using (1) DTW and (2) proposed landmark-based iterative ABS time-warping strategy.

	D_{LSD}	$D_{\text{LSD-LP}}$	D_{ITD}
DTW	3.112	2.797	4.213
Iterative warping	2.281	2.003	3.834

TABLE III. Distance measures between the warped signal and the XRMB signal from using (1) DTW and (2) proposed landmark-based iterative ABS time-warping strategy at vowel regions.

	Primary stress vowels			Secondary stress vowels			Unstressed vowels		
	D_{LSD}	$D_{\text{LSD-LP}}$	D_{ITD}	D_{LSD}	$D_{\text{LSD-LP}}$	D_{ITD}	D_{LSD}	$D_{\text{LSD-LP}}$	D_{ITD}
DTW	2.86	2.53	4.11	2.67	2.39	3.98	2.29	1.97	3.95
Iterative warping	1.54	1.28	4.01	1.90	1.69	3.99	2.24	1.93	3.58

TABLE IV. Distance measures between the warped signal and the XRMB signal from using (1) DTW and (2) proposed landmark-based iterative ABS time-warping strategy at consonant regions.

	Primary stress sonorants			Secondary stress sonorants			Unstressed sonorants		
	D_{LSD}	$D_{LSD\text{-}LP}$	D_{ITD}	D_{LSD}	$D_{LSD\text{-}LP}$	D_{ITD}	D_{LSD}	$D_{LSD\text{-}LP}$	D_{ITD}
DTW	1.82	1.49	3.53	1.66	1.29	3.69	1.74	1.47	3.33
Iterative warping	1.42	1.15	3.45	1.54	1.27	3.70	1.33	1.04	3.11
	Primary stress obstruents			Secondary stress obstruents			Unstressed obstruents		
DTW	3.80	3.59	4.25	3.53	3.32	4.10	3.21	3.01	3.70
Iterative warping	1.37	1.21	4.01	2.22	2.05	3.71	1.27	1.08	3.57

readily estimated as a Euclidean distance between the upper and lower lip pellets in the XRMB data. However, the tongue-associated TVs (TBCL, TBCD, TTCL, and TTCD) involve more complex procedures for estimation from pellets. We derive TV trajectories based on a polar coordinate system with reference to the origin labeled F in Fig. 1. TTCL is an angular measure of T1 with respect to the coordinate origin of the polar grid reference system, F, and TTCD is the minimal distance from T1 to the palate outline, which was traced for every talker in the database. For TBCL and TBCD, a circle was estimated for the TB such that it passed through T3 and T4 with a fixed radius.[2] TBCL was estimated as the angle of a line connecting the TB circle's center (C in Fig. 1) and the coordinate origin. To measure TBCD, it was necessary to estimate the dorsal vocal tract outline—the palate trace and the pharyngeal wall are available in the XRMB but there is a gap between the two. The palate trace was extended rearwards by obtaining the convex hull of all tongue pellet data for that subject; the remaining gap to the pharyngeal wall was then linearly interpolated. TBCD was estimated as the shortest distance from the TB circle to the dorsal outline. Note that GLO and VEL were excluded from the evaluation because XRMB does not contain any corresponding flesh-point data for these TVs.

Once the TV trajectories were derived from the XRMB recordings, their correlation with the TVs TADA-generated from the final estimated gestural scores was computed. We have used the Pearson product moment correlation (PPMC) which indicates the strength of a linear relationship between the TADA-generated and the XRMB-derived TV trajectories and is defined as

$$r_{\mathrm{PPMC}}=\frac{N\sum_{i=1}^{N}\hat{\tau}_i\tau_i-\left[\sum_{i=1}^{N}\hat{\tau}_i\right]\left[\sum_{i=1}^{N}\tau_i\right]}{\sqrt{N\sum_{i=1}^{N}\hat{\tau}_i^2-\left(\sum_{i=1}^{N}\hat{\tau}_i\right)^2}\sqrt{N\sum_{i=1}^{N}\tau_i^2-\left(\sum_{i=1}^{N}\tau_i\right)^2}}, \tag{4}$$

where τ and $\hat{\tau}$ represent the TADA-generated and XRMB-derived TV vector, respectively, and N represents their length. Each phone is associated with a set of gestures controlling the corresponding TVs. The correlation measure was performed during the activation interval of each phone's primary gesture(s); e.g., TT gesture for /t/ and /s/, lip gesture for /p/ and /f/, etc. Before the correlation is computed, the XRMB-derived TVs (145.6452 Hz sampling rate) were up-sampled to the sampling rate (200 Hz) of the TADA-generated TVs by linear interpolation. The final overall PPMC for a phone was calculated by averaging all the individual PPMCs for that given phone. Table V shows the correlations[3] obtained between the TADA-generated TVs and those derived from XRMB flesh-point data. The correlation analysis for vowel phones was only performed for TBCL and TBCD for unrounded vowels, as they are produced in the model with no active lip control, and also for LA for rounded vowels. The PPMC for the constriction location variables (TTCL and TBCL) were lower than the degree variables (TTCD, TBCD, and LA) because the location variables capture more speaker specific information (e.g., the tongue ball radius, the hard palate contour, etc.). However, the TVs recovered from acoustics by our TADA-based method were speaker invariant, whereas the TVs approximated from XRMB movement data were speaker specific; hence, the correlation results for location were not as high as those obtained for the speaker independent constriction degree variables. This is particularly an issue for the TBCL for vowels, which has the lowest correlation overall. The CL for low back vowels is in the pharynx but the tongue pellets are only on the front part of the tongue. So determining the appropriate CL using the procedure adopted here would depend on having the correct radius for the tongue ball, and this radius is not being adjusted at all. Future work in this area could remedy this by optimizing the radius for a given speaker separately on each optimization cycle. TTCL is also poor, presumably because of differences in the overall size of the talkers' heads, which caused the center of the grid system to be effectively misplaced.

TABLE V. Correlation between the annotated TVs and the TVs derived from the measured flesh- point information of XRMB database.

	Correlation (r)	
TVs	Consonants	Vowels
LA	0.715	0.686
TTCL	0.291	—
TTCD	0.596	—
TBCL	0.510	0.391
TBCD	0.579	0.587
Avg	0.538	0.555

An alternative approach could leverage methods developed by Carreira-Perpiñán and colleagues for estimating the complete tongue contour from limited landmarks like those available in the XRMB (Qin and Carreira-Perpiñán, 2010). This method is particularly effective when a complete contour is available as training data, such as might be obtained from the x-ray raster scans available for some XRMB speakers that were used to locate the initial positions of the pellets. Even without such training data the estimations resulting from this machine-learning technique have been shown to outperform spline interpolants, particularly when extrapolating beyond the pellets to the tongue root or tip locations. By deriving TBCL from an improved estimate of the mid-sagittal tongue surface obtained by this method (particularly extrapolated pharyngeal tongue position) we could expect an improvement in the speaker-specific correlation results.

V. CONCLUSION AND FUTURE DIRECTIONS

We have proposed a landmark based iterative ABS time-warping architecture that can potentially provide an articulatory gesture annotation for any speech database containing time-aligned word and phone transcriptions. The proposed method is robust to speaker and contextual variability, and generates a summary of the acoustic signal that is more useful for ASR applications than a phone-based transcription. We are currently in the process of generating a set of gestural annotations for a large vocabulary speech database and aim to extend our automated gestural annotation approach to other speech recognition databases as well.

ACKNOWLEDGMENTS

This research was supported by NSF Grant Nos. IIS0703859, IIS0703048, and IIS0703782, and NIH-HIDCD Grant No. DC-02717. We gratefully acknowledge the assistance of Dr. Jiahong Yuan (Department of Linguistics, University of Pennsylvania) for providing us with the forced-aligned phones for the Wisconsin University XRMB database. H.N. and V.M. contributed equally to this study.

[1] Some phonological processes such as extreme reductions and assimilations in casual speech and stop epenthesis in words like "prince" have been analyzed as involving changes in the pattern of which gestures overlap (Browman and Goldstein, 1990). If this analysis is correct, then the warping scheme outlined here would fail to produce acoustics that match well in such cases, and thus local increases in the distance metric might be able to be used to identify instances of this type of process, as opposed to ones that involve only changes in the temporal extent of overlap, which should be well modeled.

[2] We used a TB circle of 20 mm radius, which is for a default speaker in TADA.

[3] Note that LP is not included in the correlation result because LP is not used as the primary articulation distinguishing consonantal gestures.

Atal, B. S. (**1983**). "Efficient coding of LPC parameters by temporal decomposition," in *Proceedings of ICASSP*, Boston, MA, pp. 81–84.

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

Browman, C. P., and Goldstein, L. (**1990**). "Gestural specification using dynamically-defined articulatory structures," J. Phonetics **18**(3), 299–320.

Browman, C. P., and Goldstein, L. (**1995**). "Gestural syllable position effects in American English," in *Producing Speech: Contemporary Issues (for Katherine Safford Harris)*, edited by F. Bell-Berti and L. J. Raphael (AIP Press, Woodbury, NY), pp. 19–33.

Byrd, D. (**1995**). "C-centers revisited," Phonetica **52**, 285–306.

Byrd, D., and Saltzman, E. (**1998**). "Intragestural dynamics of multiple phrasal boundaries," J. Phonetics **26**, 173–199.

Byrd, D., and Saltzman, E. (**2003**). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," J. Phonetics **31**(2), 149–180.

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

Cho, T. (**2005**). "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English," J. Acoust. Soc. Am. **117**(6), 3867–3878.

Cho, T. (**2006**). "Manifestation of prosodic structure in articulatory variation: Evidence from lip movement kinematics in English," in *Laboratory Phonology 8 (Phonology and Phonetics)*, edited by L. Goldstein, D. H. Whalen, and C. Best (Walter de Gruyter, Berlin), pp. 519–548.

de Jong, K. J. (**1995**). "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," J. Acoust. Soc. Am. **97**(1), 491–504.

Fougeron, C., and Keating, P. A. (**1997**). "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.

Fowler, C. (**1980**). "Coarticulation and theories of extrinsic timing," J. Phonetics **8**, 113–133.

Furui, S. (**1986**). "On the role of speech transition for speech perception," J. Acoust. Soc. Am. **80**(4), 1016–1025.

Goldstein, L., Byrd, D., and Saltzman, E. (**2006**). "The role of vocal tract gestural action units in understanding the evolution of phonology," in *From Action to Language: The Mirror Neuron System*, edited by M. Arbib (Cambridge University, Cambridge), pp. 215–249.

Goldwater, S., and Johnson, M. (**2005**). "Representational bias in unsupervised learning of syllable structure," in *Proceedings of Computational Natural Language Learning*, pp. 112–119.

Hanson, H. M., and Stevens, K. N. (**2002**). "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," J. Acoust. Soc. Am. **112**(3), 1158–1182.

Howitt, A. W. (**2000**). "Vowel landmark detection," in *Proceedings of Interspeech, International Speech Communication Association*, Beijing, China, pp. 628–631.

Hu, C., Zhuang, X., and Hasegawa-Johnson, M. (**2010**). "FSM-based pronunciation modeling using articulatory phonological code," in *Proceedings of Interspeech*, pp. 2274–2277.

Juneja, A., and Espy-Wilson, C. (**2008**). "Probabilistic landmark detection for automatic speech recognition using acoustic-phonetic information," J. Acoust. Soc. Am. **123**(2), 1154–1168.

Jung, T. P., Krishnamurthy, A. K., Ahalt, S. C., Beckman, M. E., and Lee, S. H. (**1996**). "Deriving gestural scores from articulator-movement records using weighted temporal decomposition," IEEE Trans. Speech Audio Process. **4**(1), 2–18.

Jusczyk, P. (**1993**). "From general to language-specific capacities: The WRAPSA model of how speech perception develops," J. Phonetics **21**, 3–28.

Kochetov, A. (**2006**). "Syllable position effects and gestural organization: Articulatory evidence from Russian," in *Papers in Laboratory Phonology 8*, edited by L. Goldstein, D. Whalen, and C. Best (Mouton deGruyter, Berlin), pp. 565–588.

Krakow, R. A. (**1999**). "Physiological organization of syllables: A review," J. Phonetics **27**, 23–54.

Mehler, J., Jusczyk, P. W., Lambertz, G., Halstead, N., Bertoncini, J., and Amiel-Tison, C. (**1988**). "A precursor of language acquisition in young infants," Cognition **29**, 143–178.

Mermelstein, P. (**1975**). "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am. **58**, 880–883.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2010a**). "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," IEEE J. Sel. Top. Signal Process. **4**(6), 1027–1045.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2010b**). "Robust word recognition using articulatory trajectories and gestures," in *Proceedings of Interspeech*, Makuhari, Japan, pp. 2038–2041.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2011**). "Gesture-based dynamic Bayesian network for noise robust speech recognition," in *Proceedings of ICASSP*, Prague, Czech Republic, pp. 5172–5175.

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E., and Tiede, M. (**2012**). "Temporal planning in speech: Syllable structure as coupling graph," J. Phonetics **40**(3), 374–389.

Nam, H. (**2007**). "Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure," in *Laboratory Phonology 9 (Phonology and Phonetics)*, edited by J. Cole and J. I. Hualde (Walter de Gruyter, Berlin), pp. 483–506.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (**2004**). "TADA: An enhanced, portable task dynamics model in Matlab," J. Acoust. Soc. Am. **115**(5), 2430.

Perkell, J. (**1969**). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (Maple Press, York, PA), 104 pp.

Qin, C., and Carreira-Perpiñán, M. (**2010**). "Reconstructing the full tongue contour from EMA/X-Ray microbeam," in *Proceedings of ICASSP*, pp. 4190–4193.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, M., Tiede, M., and Browman, C. (**1996**). "CASY and extensions to the task-dynamic model," in *Proceedings of the First ESCA ETRW on Speech Production Modeling—4th Speech Production Seminar: Models and Data*, Autrans, France, pp. 125–128.

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **26**(1), 43–49.

Saltzman, E., and Munhall, K. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological Psychol. **1**(4), 332–382.

Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L. (**2008**). "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of Speech Prosody*, Campinas, Brazil, pp. 175–184.

Shannon, C. E., and Bode, H. (**1950**). "A simplified derivation of linear least square smoothing and prediction theory," Proc. IRE **38**, 417–425.

Stevens, K. N. (**1985**). "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic Press, Orlando, FL), pp. 243–255.

Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., and Liu, S. (**1992**). "Implementation of a model for lexical access based on features," in *International Conference on Spoken Language Processing*, International Speech Communication Association, Banff, Alberta, pp. 499–502.

Sun, J. P., and Deng, L. (**2002**). "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," J. Acoust. Soc. Am. **111**(2), 1086–1101.

Sun, J. P., Jing, X., and Deng, L. (**2000**). "Annotation and use of speech production corpus for building language universal speech recognizers," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, Beijing, China, Vol. 3, pp. 31–34.

Tepperman, J., Goldstein, L., Lee, S., and Narayanan, S. (**2009**). "Automatically rating pronunciation through articulatory phonology," *Proceedings of Interspeech*, pp. 2771–2774.

Turk, A. E., and Sawusch, J. R. (**1997**). "The domain of accentual lengthening in American English," J. Phonetics **25**, 25–41.

Westbury, J. (**1994**). *X-ray Microbeam Speech Production Database User's Handbook* (University of Wisconsin, Madison, WI), pp. 1–135.

Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," J. Acoust. Soc. Am. **123**(5), 3878.

Zhuang, X., Nam, H., Hasegawa-Johnson, M., Goldstein, L., and Saltzman, E. (**2009**). "Articulatory phonological code for word classification," in *Proceedings of Interspeech*, Brighton, UK, pp. 2763–2766.