Recognizing articulatory gestures from speech for robust speech recognition

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

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

Carol Espy-Wilson
Speech Communication Laboratory, Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742

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

Louis Goldstein
Department of Linguistics, University of Southern California, Los Angeles, California 90089

(Received 16 March 2011; revised 5 January 2012; accepted 7 January 2012)

Studies have shown that supplementary articulatory information can help to improve the recognition rate of automatic speech recognition systems. Unfortunately, articulatory information is not directly observable, necessitating its estimation from the speech signal. This study describes a system that recognizes articulatory gestures from speech, and uses the recognized gestures in a speech recognition system. Recognizing gestures for a given utterance involves recovering the set of underlying gestural activations and their associated dynamic parameters. This paper proposes a neural network architecture for recognizing articulatory gestures from speech and presents ways to incorporate articulatory gestures for a digit recognition task. The lack of natural speech database containing gestural information prompted us to use three stages of evaluation. First, the proposed gestural annotation architecture was tested on a synthetic speech dataset, which showed that the use of estimated tract-variable-time-functions improved gesture recognition performance. In the second stage, gesture-recognition models were applied to natural speech waveforms and word recognition experiments revealed that the recognized gestures can improve the noise-robustness of a word recognition system. In the final stage, a gesture-based Dynamic Bayesian Network was trained and the results indicate that incorporating gestural information can improve word recognition performance compared to acoustic-only systems.
[DOI: 10.1121/1.3682038]

PACS number(s): 43.72.Ar, 43.72.Bs, 43.70.Bk, 43.72.Ne [ADP] Pages: 2270–2287

I. INTRODUCTION

Current state-of-the-art automatic speech recognition (ASR) systems represent speech as a sequence of non-overlapping phone units. Although such systems perform fairly well for clearly articulated speech under "controlled" conditions, their performance degrades for spontaneous speech, which contains acoustic variations shaped by linguistic and speaker-specific properties (segmental and prosodic structure, speaking style, speaker-identity, etc.) and by physical properties of the environment (noise, channel differences etc.). Studies have shown that human speech recognition performance is considerably more robust against such contextual variation (Lippman, 1997) and environmental noise (Cooke *et al.*, 2006). These differences indicate the need for more robustness in ASR systems.

Performance degradation of current phone-based ASR systems for spontaneous speech can partly be attributed to the non-overlapping phone-based modeling assumption (Ostendorf, 1999), that limits the acoustic model's ability to properly learn the underlying variations in natural speech. A part of this variation is due to coarticulation that arises due to the overlapping, asynchronous nature of speech articulator movements. Current ASR systems model coarticulation using tri- or quin-phone based acoustic models, where different models are created for each phone in all possible phone-contexts. Unfortunately, such tri- or quin-phone models limit the contextual influence only to immediately close neighbors which may fail to account for the full extent of coarticulation effects [e.g., syllable deletions (Jurafsky *et al.*, 2001)] and can suffer from data-sparsity due to the relative rarity of some tri- or quin-phone units.

To address the problem posed by variability in speech for ASR, Stevens (1960) suggested incorporating speech production knowledge into ASR architectures. Incorporating such knowledge is challenging, since acoustic waveforms

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

 0001-4966/2012/131(3)/2270/18/$30.00

are usually the only inputs to ASR systems and no speech production related data (such as vocal tract shapes, articulatory configurations, their trajectories over time, etc.) are used. Hence, the first logical step in order to introduce speech production knowledge into ASR is to estimate such information from the acoustic signal.

Several approaches have been used to incorporate different types of speech production knowledge into ASR systems; a detailed description of such approaches is given in King *et al.* (2007) and we briefly discuss some of those approaches below.

A. Articulatory features (AF) in ASR

Distinctive features were first developed and further elaborated by linguists to distinguish and classify segments or sounds (Jakobson *et al.*, 1952; Chomsky and Halle 1968). Features are discrete or categorical (typically binary), and may be defined either acoustically or articulatorily. ASR studies have incorporated *articulator-bound* features (AFs) of the acoustic signal that reflect actions of particular articulators (e.g., +/− voicing for larynx; +/− rounding for lips). ASR studies using AFs fall into two broad categories according to the roles played by the AFs in the recognition model. One type of ASR system uses local classifiers (e.g., artificial neural networks [ANNs]) to identify features which are then treated as observations to be processed by, for example, a Hidden Markov Model (HMM); the second, more recent type of system treats AFs as hidden variables within an HMM or Dynamic Bayesian Network (DBN; e.g., Bilmes and Zweig, 2002). An example of the former type of model was presented by Schmidbauer (1989), who used a probabilistic model to obtain 19 AFs (describing the manner and place of articulation) from speech, and showed that the resulting AF-HMM system provided an improvement of 4% over a baseline HMM phoneme recognizer using Mel-frequency cepstral coefficients (MFCCs) as acoustic observations for a small German speech database. Relatedly, Deng and Sun (1994) used 5 multi-valued features representing the lips, tongue blade, tongue dorsum, velum and larynx, and mapped the overlapping patterns of features to HMM state transition graphs. Their method resulted in a relative phone recognition improvement of 9% over a baseline MFCC-HMM for the TIMIT database. Finally, Kirchhoff (1999) used quantized (rather than binary) AFs that were estimated using a multi-layered perceptron (MLP), and showed that using AFs in addition to MFCCs for ASR helped to improve word error rate (WER) and increased recognition robustness against background noise.[1]

The more recent use of DBNs in ASR systems allows the AFs to be treated as hidden variables while explicitly taking into account inter-dependencies among these variables during word recognition. For example, Frankel *et al.* (2004) showed that the average AF recognition accuracy can be improved from 80.8% to 81.5% by using a DBN to model inter-feature dependencies. In a different study Frankel and King (2005) described a hybrid artificial neural network (ANN) - DBN architecture that combined the discriminative training capability of ANNs with the inter-feature dependency modeling capability of DBNs and reported a feature recognition accuracy of 87.8% for the OGI Number corpus. Richardson *et al.* (2003) proposed the Hidden Articulatory Markov Model (HAMM), in which each HMM state represents an articulatory configuration for each di-phone context, allowing asynchrony among the articulatory features. When used in tandem with a traditional HMM (4-state MFCC-HMM speech recognizer), HAMM helped to reduce the absolute WER by an average of approximately 1.07% compared to the MFCC-HMM system.

B. Continuous articulatory trajectories in ASR

A different line of research has focused on testing the degree to which the use of (either directly measured or estimated) continuous articulatory trajectory information can improve the erformance of ASR systems. Typically, articulatory trajectories are measured from the positions of transducers (or pellets) attached to flesh-points on the different articulators in the mid-sagittal plane of the vocal tract (Wrench and Hardcastle, 2000). In a typical ASR scenario, however, such trajectory information is not available and must be estimated from acoustic observations using an inverse mapping procedure, commonly known as "speech inversion" or "acoustic-to-articulatory inversion." Since the reliability of speech inversion is known to suffer from the inherent non-linearity and non-uniqueness of the acoustic-to-articulatory map (Richmond, 2001), there are few ASR results in the literature using estimated articulatory trajectories. Rather, several studies have focused on the utility for ASR of using veridical, measured articulatory trajectories as inputs to ASR models. Frankel and King (2001) built a phone recognition system that used a combination of acoustic features (MFCCs) and articulatory data (flesh-point articulatory information modeled as linear dynamic model (LDM) parameters at each phone-segment) as inputs, and demonstrated a 9% improvement in phone recognition over a system using only the MFCCs. However the phone recognition accuracies based on estimated articulatory data in conjunction with the MFCCs did not show any improvement over the system using only the MFCCs, indicating the need for more accurate and reliable speech inversion architectures. Markov *et al.* (2006) described a hybrid HMM and Bayesian network (BN) model, where the BN described the dependence of acoustics on quantized EMA articulations, and of quantized EMA articulations on phones; and the HMM modeled phone transitions. Probabilistic dependencies between EMA and acoustics were learned during training, but during recognition, only the acoustic observations were input and the articulatory variables were hidden states. They reported that the HMM-BN trained using articulatory information always performed better than the baseline HMM system trained only with the acoustic features.

C. Articulatory gestures in ASR

Acoustic variations in the speech signal due to coarticulation can be simply and elegantly described with reference to the spatiotemporal behavior of discrete constricting actions in the vocal tract called *gestures*. In Articulatory

TABLE I. Constrictors and their vocal tract variables.

Constrictors	Vocal tract variables
Lip	Lip Aperture (LA)
	Lip Protrusion (LP)
Tongue Tip	Tongue tip constriction degree (TTCD)
	Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD)
	Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

Phonology (Browman and Goldstein, 1989, 1992), a phonological theory that views an utterance as a constellation of gestures that may overlap in time, gestures are defined as discrete action units whose activation results in constriction formation or release by five distinct constrictors (lips, tongue tip, tongue body, velum and glottis) along the vocal tract. The kinematic state of each constrictor is defined by its corresponding constriction degree and location coordinates, which are called vocal tract constriction variables (henceforth, tract-variables or TVs) (see Table I and Fig. 1; note that a constriction-location TV is not needed for the glottis and velum constrictors since they are fixed in location). Table II presents the dynamic range and the measuring units for each TV. All constriction-degree TVs are defined in mm. For the constriction location TVs, TBCL and TTCL are polar angular distance measures for the tongue body and tongue tip constrictors with respect to a reference line originating at the floor of the mouth (F, in Fig. 1), and LP is a measure of horizontal protrusion of the lip constrictor relative to a reference position. Figure 1 shows a vocal tract configuration with a TBCL of 90° and TTCL of 45°.

Each gesture is associated with a given constrictor and is specified by an activation onset and offset time and by a set of dynamic parameters (target, stiffness, and damping); when a given gesture is activated, its parameters are inserted into the associated constrictor's TV equations of motion. These equations are defined as critically damped second order systems (Saltzman and Munhall, 1989), as shown in (1):

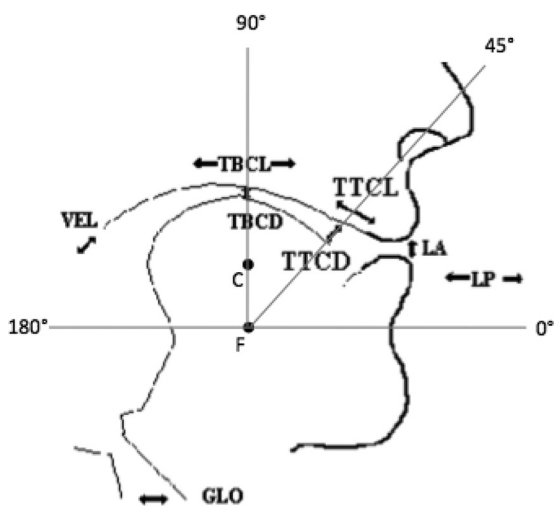

FIG. 1. Vocal tract variables at 5 distinct constriction organs.

TABLE II. Units of measurement and dynamic range of each TV.

		Dynamic range	
TVs	Unit	Max	Min
GLO	–	0.74	0.00
VEL	–	0.20	−0.20
LP	mm	12.00	8.08
LA	mm	27.00	−4.00
TTCD	mm	31.07	−4.00
TBCD	mm	12.50	−2.00
TTCL	degree	80.00	0.00
TBCL	degree	180.00	87.00

$$M\ddot{z} + B\dot{z} + K(z - z_0) = 0, \quad (1)$$

where M, B, and K are the mass, damping, and stiffness parameters of each TV (represented by z), and z_0 is the TV's target position; every parameter except M is a time-varying function of the corresponding parameters of the currently active set of gestures; and, due to the assumption of constant mass and critical damping, the damping coefficients are constrained to be simple functions of the ongoing stiffness values. The gestural structure of an utterance is represented by its *gestural score,* which includes the set of gestural activation intervals for the utterance, the pattern of relative timing among these intervals, and the associated sets of gestural dynamic parameters. The gestural score and set of tract variable trajectories or time functions (TV_ts) for an arbitrary utterance are generated using the Haskins Laboratories Task Dynamics Application [TaDA, (Nam *et al.,* 2004)]. In this model, gestural scores are generated from orthographic or ARPABET transcription inputs, according to the principles of Browman and Goldstein's Articulatory Phonology; TV_ts and articulator trajectories are computed using Saltzman and Munhall's (1989) Task Dynamic model of gestural pattern dynamics; and vocal tract shapes, area functions, and formants are calculated by an articulatory synthesis model (Rubin *et al.,* 1981). Finally, the outputs of TaDA are used in conjunction with HLsyn [a parametric quasi-articulatory synthesizer developed by Sensimetrics Inc. (Hanson and Stevens, 2002)] to generate the resulting audio signal. Figure 2 shows a gestural score for the utterance "miss you," and the corresponding TV_ts, acoustic signal, and formant structure. Note that gestural onsets and offsets are not always aligned to acoustic landmarks, e.g., the beginning of the frication for /s/ is delayed with respect to the onset of the tongue tip constriction gesture (TTCD) for /s/, due to the time it takes for the tongue tip to attain a position close enough to the palate to generate turbulence; such asynchronies highlight the distinction between gestures and AFs.

However, like other articulatory information (e.g., articulator flesh-point trajectories), gestures are not readily obtainable in a typical ASR situation; rather, they need to be estimated from the speech signal. Sun and Deng (2002) described an automated gestural score annotation model that was trained with manually annotated overlapping gestures, and reported an improvement in ASR performance. Gestural activation recovery from multi-channel articulatory recordings was performed by Jung *et al.* (1996) using a temporal

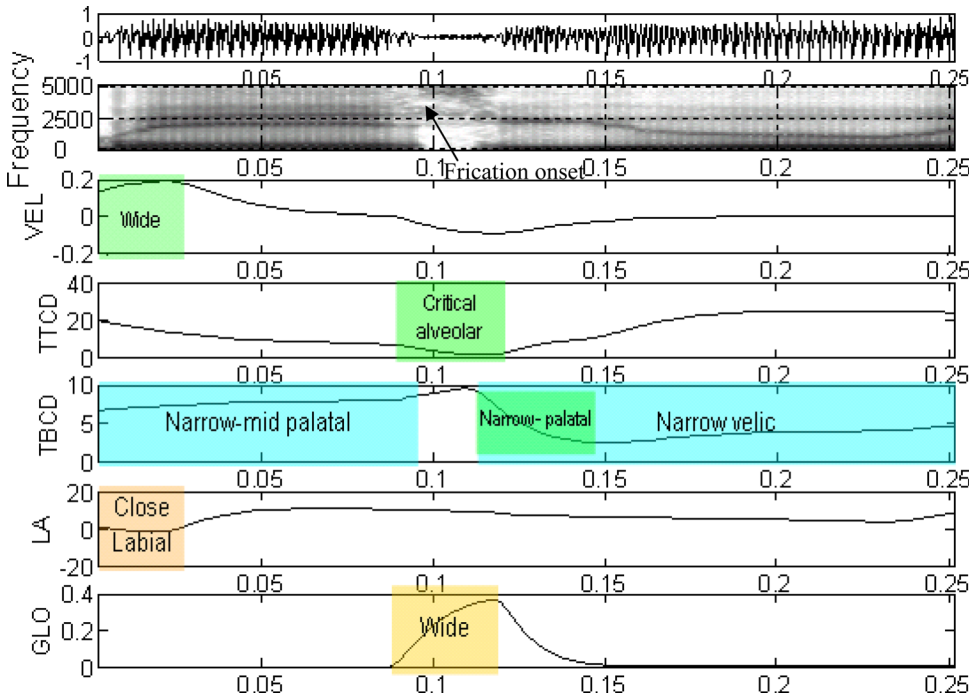

FIG. 2. (Color online) Gestural score for the utterance "miss you." Active gesture regions are marked by rectangular solid blocks. Smooth curves in the background represent the corresponding TVs.

decomposition (TD) method (Atal, 1983) that was applied to various CVC syllables embedded in frame sentences. Although gestural activations were estimated in their work, the associated sets of gestural dynamic parameters such as stiffnesses and targets were not recovered. Such parameters are crucial, however, to distinguish utterances in a gesture-based lexicon (Browman and Goldstein, 1992). The stiffness helps to distinguish consonants from vowels: the motion for consonants, which is parameterized as a gesture with higher stiffness, is faster than that of vowels. Similarly, gestural targets provide spatial information about the location and degree of a constriction. For example, in the case of/s/as in "miss" (shown in Fig. 2), the tongue-tip gesture will have a "critical" (very narrow) constriction degree target (TTCD) whose location target (TTCL) is the alveolar ridge. Hence, estimating only gestural activations is not sufficient for lexical access.

Ghosh *et al.* (2009) trained a system using dynamic programming to estimate gestural targets (but not activations or stiffnesses) for a corpus of 213 natural and phonetically balanced sentences from the Harvard IEEE Corpus. They obtained word identification accuracies of 66.67% for recognition from these gesture target vectors. Zhuang *et al.* (2009) proposed the use of gestural pattern vectors (GPVs) (which represent gestural activations and their corresponding dynamic parameters for each time frame) as recognition units. They proposed a tandem ANN-GMM model that predicts the GPVs from *a priori* knowledge of TV_ts (groundtruth TV_ts[2]) using a synthetically generated speech corpus which contained 380 unique GPVs. They performed recognition based on 181 out of the 380 GPVs, because the remaining 199 GPVs were rare in terms of frequency and, hence, the model could not learn them sufficiently. The GPVs were correctly recognized 80% of the time, which was later improved to 90% by pronunciation modeling with finite state machines (FSM) (Hu *et al.*, 2010). However, since the number of possible GPVs is potentially huge when there is large variability in gestural overlap in natural, spontaneous speech, using GPVs as hidden variables in a full-blown ASR system might introduce data-sparsity issues similar to those encountered with triphone models. In addition, the GPV recognizer in Zhuang *et al.* (2009) employed the groundtruth TV_ts, which are generally not available true in typical ASR situations.

D. Overview of the present study

The goal of this study is to develop a methodology to recognize Articulatory Phonology-based gestures from the acoustic waveform. We first show that gestures can be accurately inferred from synthetic, TaDA-generated speech for which the underlying gestures (activations and dynamic parameters) and TV_ts are known *a priori*. Gesture recognition is performed using a two-stage cascaded architecture that (a) recognizes gestural activation intervals in the first stage and (b) estimates the dynamic parameters associated with these activation intervals in the second stage. We then demonstrate that the model, trained to infer gestures using synthetic speech, can be successfully applied to real-speech inputs, and that doing so improves the noise-robustness of word recognition, consistent with the demonstration of Kirchhoff *et al.* (1999) that articulatory information is especially useful for ASR in noise. We demonstrate additionally that the TV_ts, when combined with acoustic information, provide a particularly rich source of constraint for the gestural recovery process during ASR.

The outline of the paper is as follows. First, we describe in detail the construction of two synthetic databases (XRMB-SYN and AUR-SYN) that were used for training and testing our gesture-recognition models. Second, we describe the structural optimization, training and evaluation of the TV estimator and gesture recognition models using these databases, and also describe an alternative gesture-

based Dynamic Bayesian Network (DBN) ASR architecture that treats articulatory gestures as hidden variables. Third, we present the results of experiments using the (clear and noisy) real speech of the Aurora-2 corpus: (a) TV_ts and gestures are estimated using the synthetically trained models, and these are added as additional inputs to an HMM- based word recognition system; (b) estimated TVs are added as inputs to the gesture-based DBN ASR architecture. Finally we conclude with a discussion of the implications of our findings for guiding future research within this framework.

II. DATABASES AND METHODOLOGIES

A. Databases

1. Synthetic databases

To obtain gestural score and TV_t specifications for a large set of training utterances, we used TaDA (Nam *et al.*, 2004); corresponding acoustic signals were synthesized using HLsyn (Hanson and Stevens, 2002) in conjunction with TaDA (both models were described previously in section I C.). Two types of synthetic dataset were prepared using TaDA-HLsyn, where the sampling rate for the acoustic signal was 10 kHz and that for TV_ts and gestural scores was 200 Hz. First, we created a synthetic speech corpus containing gestures and TV_ts to develop and obtain the model parameters for the gesture-recognizer. The dataset consists of 420^3 distinct words found in the X-Ray MicroBeam (XRMB) database (Westbury, 1994) and we refer to it as XRMB-SYN. Seventy five percent of the XRMB-SYN data was used for training, and 10% for validation and the rest was used for testing. Second, to ensure that the acoustic portion of the synthetic corpus used for training was phonetically similar to the natural dataset to be used for the ASR experiments, we created another TaDA-generated dataset using 960 utterances randomly selected from the Aurora-2 (Pearce and Hirsch, 2000) clean training corpus. For each utterance, their ARPABET specification, mean pitch and gender information were input to TaDA-HLsyn and the resulting corpus was named as AUR-SYN. The generation and function of these two synthetic datasets is illustrated in Fig. 3.

2. Natural speech database

The digit-corpus of Aurora-2 was selected to perform gesture and word recognition on natural speech. Aurora-2 is created from the TIdigits corpus and consists of connected digits spoken by American English speakers. The speech data in Aurora-2 are sampled at 8 kHz and have three test sections: A, B and C. Set A and B each has four subparts representing four different noise types; hence A and B altogether contain eight different noise types. Section C involves channel-effects, which we did not use in our experiments. All of our experiments involved training with clean and testing with noisy data.

B. Acoustic parameterization

The acoustic features used in both the TV_t-estimator and the gesture recognizer were Mel-frequency cepstral coefficients (MFCCs) and Acoustic-Phonetic parameters (APs) (Juneja, 2004). The APs are measures that meant to capture the acoustic correlates of acoustic-phonetic features. Examples include periodic and aperiodic energy in different frequency bands which are relevant for the phonetic feature *voiced/unvoiced* and normalized energy-based measures which are relevant for the phonetic feature *syllabic/non-syllabic*. Finally, for the word recognition experiments discussed in Sec. III below, we also used the RASTA-PLPs (Hermansky and Morgan, 1994) since they have been shown to be robust to noise. For all parameterizations, we used an analysis window of 10 ms with a frame advance of 5 ms. The dimensionality of the APs was higher than that of the MFCCs and the RASTA-PLPs; 40 different APs were used for the TV_t estimation process and their selection is explained in Mitra *et al.* (2009). All of the acoustic features and the target groundtruth TV_ts were mean subtracted, variance-normalized (with std. dev. $= 0.25$) and scaled such that their dynamic range was confined within $[-0.95, +0.95]$.

The acoustic features (MFCCs and APs) were temporally contextualized by stacking multiple frames before being sent to the TV_t-estimator and gesture recognizer. The contextualized features for a given frame at t ms, are created from a context-window of duration d ms ($d > 10$ ms), where every second feature vector (time-shift of 10 ms) evaluated

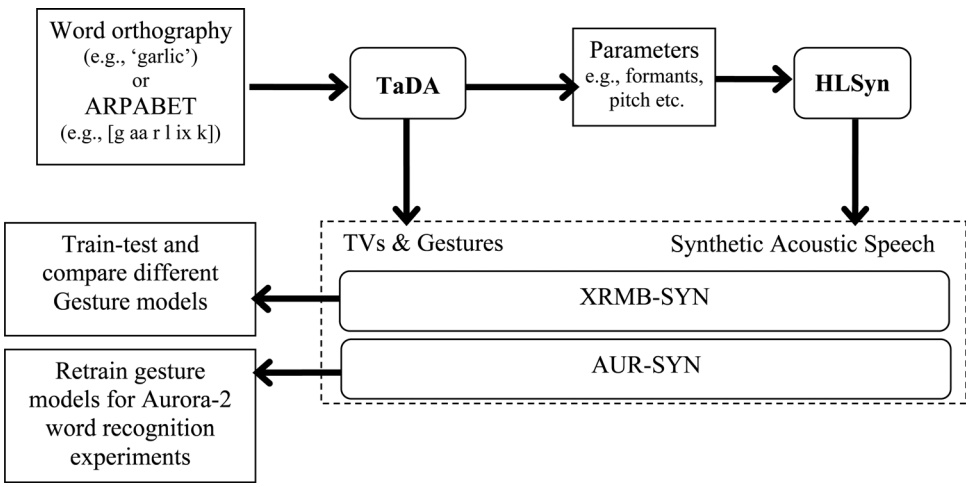

FIG. 3. Flow diagram for generating synthetic speech and the associated articulatory information using TaDA and HLsyn.

in the range $[t - d/2, t + d/2]$ ms is stacked with the feature vector at t ms to form a contextualized super-vector. From our prior research (Mitra *et al.*, 2010) we know that the optimal[4] TV_t estimation context windows for MFCCs and APs are 170 ms and 190 ms, respectively, and their dimension after contextualization is 221 and 760, respectively.

C. Gestural scores and their use in ASR

1. The TV_t estimator

During speech production, the articulators in the human vocal tract shape the acoustic resonator, resulting in an acoustic signal y that is a nonlinear function f of articulator configuration x:

$$y = f(x). \tag{2}$$

In recognition tasks, although the speech acoustic signal is available directly, articulatory data is typically available only indirectly from the speech signal via an estimation process called speech inversion that seeks to find a function g that provides an optimal estimate of x (according to some quantitative quality metric):

$$\hat{x} = g(y). \tag{3}$$

Typically speech inversion problems suffer from non-linearity and non-uniqueness (Richmond, 2001). However, recent separate studies by Qin *et al.* (2007) and Neiberg *et al.* (2008) showed that much of normal speech is produced with unique vocal tract shapes, and that non-unique instances are few. Their findings also suggested that non-linearity is more critical than non-uniqueness for speech-inversion.

In our past work (Mitra *et al.*, 2010, 2011) and current work on gestural estimation, we first estimate TV_ts from the speech signal and use this articulatory information to guide and constrain subsequent stages of gestural recovery. McGowan (1994) pointed out that TV_ts specify the salient features of the vocal tract area functions more directly than the absolute spatial trajectories of articulatory flesh-points. Similarly, we have shown (Mitra *et al.*, 2011) that speech inversion using TV_ts suffers less from non-uniqueness problems compared to approaches that rely on traditional flesh-point (pellet) trajectory information.

From our prior analysis (Mitra *et al.*, 2010) we observed that a 3-hidden layer feed- forward (FF) ANN offers reasonably accurate TV_t estimates compared to other machine-learning approaches [Support Vector Regression (Toutios and Margaritis, 2005), Trajectory Mixture Density Networks (Richmond, 2007), Distal Supervised Learning (Jordan and Rumelhart, 1992) etc.] that have either been successfully used for flesh-point based speech-inversion or are well known for inverse problems. This section reviews the FF-ANN based TV_t-estimator model trained with XRMB-SYN, which estimates TV_ts given acoustic features (MFCCs or APs) as input. In ANN- based speech-inversion models, instantaneous non-uniqueness has been addressed by using a temporal context window (Richmond, 2001) to exploit dynamic information in the input space, as specified in Sec. II B.

FF-ANNs with a minimum of 2 hidden layers (Lapedes and Farber, 1988) have the capability to learn arbitrary complex non-linear mappings of an M-dimensional input space (R^M) to an N-dimensional output space (R^N). Such ANNs are implicitly capable of exploiting any existing cross-correlations between components of the output (e.g., TV_t information in our case) (Mitra *et al.*, 2010). A single 3-hidden layer FF-ANN with eight output nodes (one for each TV_t) and tan-sigmoid activation function was trained (using the scaled conjugate gradient [SCG] algorithm), for each of the input acoustic feature sets AP and MFCC, respectively. The number of neurons in each hidden layer was optimized by analyzing the root mean squared error (RMSE) from the validation set. During this optimization stage we observed that the performance of the TV_t estimation improved as the number of hidden layers was increased. It may be the case that additional hidden layers incorporated additional non-linear activation functions into the system, which in turn increased the architecture's potential to cope with the high non-linearity inherent in a speech-inversion process [a detailed discussion on this is provided by Bengio and Le Cun (2007)]. However, the number of hidden layer was confined to three because (a) the error surface becomes more complex (with many spurious minima) as the number of hidden layer was increased, which increases the probability that the optimization process finds a local minimum and (b) increasing the number of hidden layers increases the training time and the network complexity. The optimal architectures for our networks using MFCC and AP inputs were found to be (221)-150-100-150 -(8) and (760)-250-300-250 -(8), where the first value within parenthesis represent the dimensionality of the contextualized input feature vector, the last value in parenthesis denotes the total number of output TV_ts and the three values in between represent the number of neurons in each of the three hidden layers.

The estimated TV_ts from the FF-ANN models were found to be noisy due to estimation error. TV_ts by definition are smooth[5] trajectories due to neuro-biomechanical constraints; hence to ensure smoothness of the estimated TV_ts a Kalman-smoother was used as a post-processor. The performance of the TV_t estimator used in our experiments reported in this paper has already been reported in Mitra *et al.* (2010) and, hence, we will not be presenting those results here.

2. The gesture recognizer

Recognizing an utterance's gestural structure entails recovering the utterance's gestural score, i.e., gestural activation intervals, their pattern of relative timing, and the associated sets of gestural dynamic parameters (targets and stiffnesses). Note that the activation value of a gesture for a given TV at a given instant of time is treated as a discrete binary variable: activation values are 1 for active gestures and 0 for non-active gestures. If a gesture is active, its parameter set will influence the ongoing spatiotemporal shape of the associated TV. Due to the assumption of critical damping, note that the damping parameter, B, in (1) is constrained to be a simple function of stiffness and does not need to be estimated independently.[6]

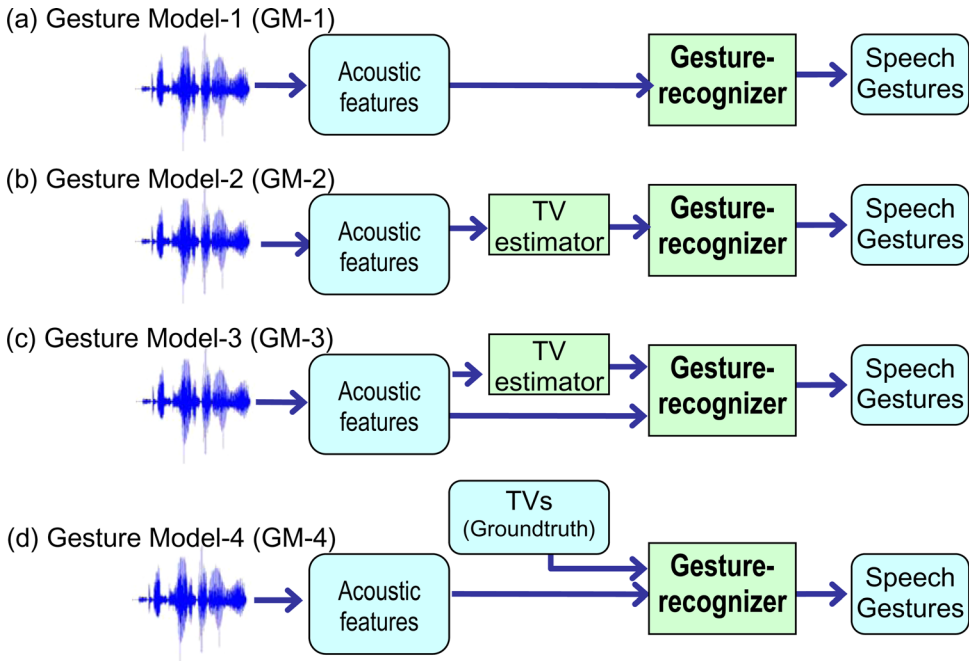

FIG. 4. (Color online) The Four approaches for Gesture recognition.

We compared the performance of four types of gesture recognition models that differed in terms of the types of inputs used (see Fig. 4). Gesture model 1 (GM-1) used the acoustic features only (i.e., the MFCCs or the APs); Gesture model 2 (GM-2) used only the TV_ts estimated from the acoustic features (using the model presented in last section); Gesture models 3 (GM-3) and 4 (GM-4) both used TV_ts along with the acoustic features, with the former using estimated TV_ts and the latter using groundtruth TV_ts. A 2-stage cascade ANN architecture (shown in Fig. 5) was adopted for all four models, in which gestural activation (onset and offset) information was obtained in the first stage using a nonlinear autoregressive (AR) ANN, and gestural parameter estimation (target and stiffness values) was performed in the second stage using an FF-ANN.

We considered 10 different TVs for these models: LP, LA, TTCL, TTCD, TBCL_C, TBCL_V, TBCD_C, TBCD_V, VEL and GLO. Note that, since tongue body gestures are shared by velar consonants and vowels with distinct timescales (fast for consonants and slow for vowels), the original TBCL and TBCD TVs used in TaDA were differentiated into consonant (TBCL_C and TBCD_C) and vowel (TBCL_V and TBCD_V) sub-TVs. Separate cascaded gesture-recognition models were trained for each TV and sub-TV using each of the four input combinations shown in Fig. 4; thus, 4 cascade models were trained for each TV/sub-TV, except for GLO and VEL.[7] The best architecture would probably have gestures for TVs recognized conjointly based on all TVs, because shared articulators between TVs can cause passive movement in a TV that is not being controlled by an active gesture. For example, a TT constriction gesture will typically engage jaw-raising, which will, everything else being equal, cause a passive decrease in LA. That LA change should not count as evidence of a LA gesture, but when gestures are recognized independently, based on their respective TVs, a LA gesture could incorrectly be recognized. However, we chose to perform gesture recognition separately for each TV in order to keep the model simple and easier to train.

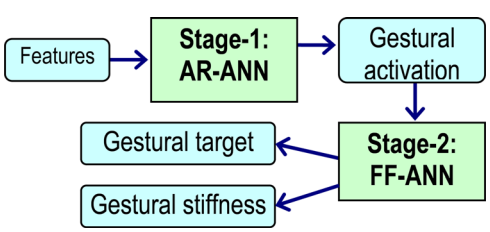

FIG. 5. (Color online) The 2-stage cascaded ANN architecture for gesture recognition.

In the gestural estimation process, gestural activation is treated as a discrete binary random variable that, at any given instant of time i, can only be in one of the two possible states: $S_i \in \{0,1\}$, with $S_i = 1$ when active, or $S_i = 0$ when inactive. Once a gesture is active or inactive it maintains that state for a certain interval of time (from 50 ms to 300 ms). We model this state duration property by incorporating *memory* into the gestural activation detection process, using the recurrent feedback loop of an AR-ANN (Demuth *et al.*, 2008). Memory is used to remember the sequence of prior activation states (S_{t-1}, S_{t-2},…, $s_{t-\Delta}$) and that information along with the current acoustic observation $u(t)$ is used to predict the activation state S_t for the tth time instant. As shown by Eq. (4)

$$S_t = f_{AR-ANN}(S_{t-1}, S_{t-2}, \ldots, S_{t-\Delta}, u(t)), \quad (4)$$

where $f_{AR\text{-}ANN}$ represents the nonlinear AR-ANN network. Note that the autoregressive memory serves to effectively prevent instantaneous switching between the binary states.

The second stage of the gesture recognition model uses an FF-ANN to estimate gestural dynamic parameter values (targets and stiffnesses) during active gestural intervals. Obtaining gestural dynamic parameters is essentially a

function estimation problem where the parameters target and stiffness can theoretically have any real value and FF-ANNs can be trained to approximate any function (with a finite number of discontinuities) (Lapedes and Farber, 1988).

The acoustic features (MFCCs or APs) used as inputs to the cascaded ANNs were temporally contextualized in a similar manner as was done for TV_t estimation (see Secs. II B and II C 1). The optimal context windows for each stage were found to vary for different TVs, and the optimal values are reported along with the results in Sec. III A.

D. Gesture-based Dynamic Bayesian Network (G-DBN) for speech recognition

Using the observations gleaned from the ANN gestural models discussed above, we present a fully deployable Dynamic Bayesian Network (DBN) based ASR architecture, where the gestural states are modeled as discrete random variables. Instead of explicitly recognizing the gestures as done by our ANN-based models, the DBN treats gestures as discrete random variables (RVs) that are observed during training and hidden during testing.

A DBN architecture can be seen as a generalization of the HMM architecture (Ghahramani, 1998). We chose to use a DBN instead of a HMM for two reasons. First, DBNs have the flexibility to realize multiple hidden variables at a given time. As a result, a DBN can model articulatory gestures as individual state variables, one for each articulatory gesture. Second, a DBN can explicitly model the interdependencies amongst the gestures and can simultaneously perform gesture recognition and word recognition, eliminating the need to perform gesture recognition as a prior separate step before word recognition. For our DBN implementation we used the Graphical Models Tool-Kit (GMTK) (Bilmes and Zweig, 2002), where conditional probability tables (CPT) are used to describe the probability distributions of the discrete RVs given their parents, and GMMs are used to define the probability distributions of the continuous RVs.

In a typical HMM based ASR setup, word recognition is performed using maximum *a posteriori* probability

$$w - \text{argmax}_i P(wi|o) - \text{argmax}_{wi} \frac{P(w_i)P(o|w_i)}{P(o)}, \tag{5}$$

where o is the observation variable and $P(w_i)$ is the language model that can be ignored for an isolated word recognition task where all of the words w are equally probable. Hence we are left with $P(o/w_i)$ which is given as

$$\begin{aligned} P(o|w) &= \sum_s P(s,o|w) = \sum_s P(s|w)P(o|s,w) \\ &\approx \sum_s P(s_1|w)P(o_1|s_1,w)\Pi_{i=2}^{n} P(s_i|s_{i-1},w)P(o_i|s_i,w), \end{aligned} \tag{6}$$

where s is the hidden state in the model. In this setup the likelihood of the acoustic observation given the model is calculated in terms of the emission probabilities $P(o_i/s_i)$ and the transition probabilities $P(s_i|s_{i-1})$. Use of articulatory information introduces another RV a, which alters (6) as

$$\begin{aligned} P(o|w) \approx \sum_s P(s_1|w)P(o_1|s_1,a_1,w) \prod_{i=2}^{n} P(s_i|s_{i-1},w) \\ \times P(a_i|a_{i-1},s_i)P(o_i|s_i,a_i,w). \end{aligned} \tag{7}$$

DBNs can model both (a) the causal relationship between the articulators and the acoustic observations $P(o/s,a,w)$ and (b) the dependency of articulators on the current phonetic state and previous articulators $P(a_i|a_{i-1},s_i)$.

Tying each individual gestural state with a word state can potentially result in large CPTs, which can significantly slow down the DBN to the extent of not being able to test it. To address this, we slightly modified Eq. (7) as follows

$$\begin{aligned} P(o|w) = \sum_s P(s_1|w)P(a_1|w)P(o_{1,1}|s_1,w)P(o_{2,1}|a_1,w) \\ \times \left[\prod_{i=2}^{n} P(s_i|s_{i-1},w)P(a_i|a_{i-1},w)P(o_{1,i}|s_i,w) \right. \\ \left. \times P(o_{2,i}|a_i,w) \right]. \end{aligned} \tag{8}$$

Equation (8) is based upon the assumption that gestural states and word states are individual entities tied directly to the word RVs. This is represented by the graphical model shown in Fig. 6.

1. Model architecture

In the G-DBN model of Fig. 6, square/circular nodes represent discrete/continuous RVs, and shaded/unshaded nodes represent observed/hidden RVs. As can be seen, the DBN consists of four discrete hidden RVs (W, P, S and T), two continuous observable RVs (O_1 and O_2) and N partly observable and partly hidden gesture RVs (A_1 to A_N). The prologue and the epilogue in Fig. 6 denote the initial and the final frame(s) (where the frame specifications are same as the observation) and the center represents the intermediate frames, which are unrolled in time to match the duration of a given utterance. For a word model, S represents the word state, W represents the word RV, P represents the word position RV (word position at a given time instant gives the position in the whole word state for a given word at that time instant) and T represents the word transition RV. As in our explicit gesture recognition models (described in the previous section), we incorporated memory into the gesture RVs (although for simplicity we have incorporated only a single-step memory) where the gestural states at the current time instant is tied to the gestural state at the previous time instance (see Fig. 6). Note that there are no cross- gesture dependencies among the gesture RVs, just as in the explicit gesture recognizer. Thus, despite the potential noted above that the DBN offers for modeling gestural dependencies, it was not employee here, for reasons of model tractability.

We modeled six articulatory gesture activations (GLO, VEL, LA, LP, TT and TB) as hidden RVs in the G-DBN architecture, so N (the subscript of A) in Fig. 6 is six. The gestural activations for TTCL and TTCD are identical. As a result, they were replaced by a single RV, TT tongue tip) and the same was true for TBCL and TBCD, which were

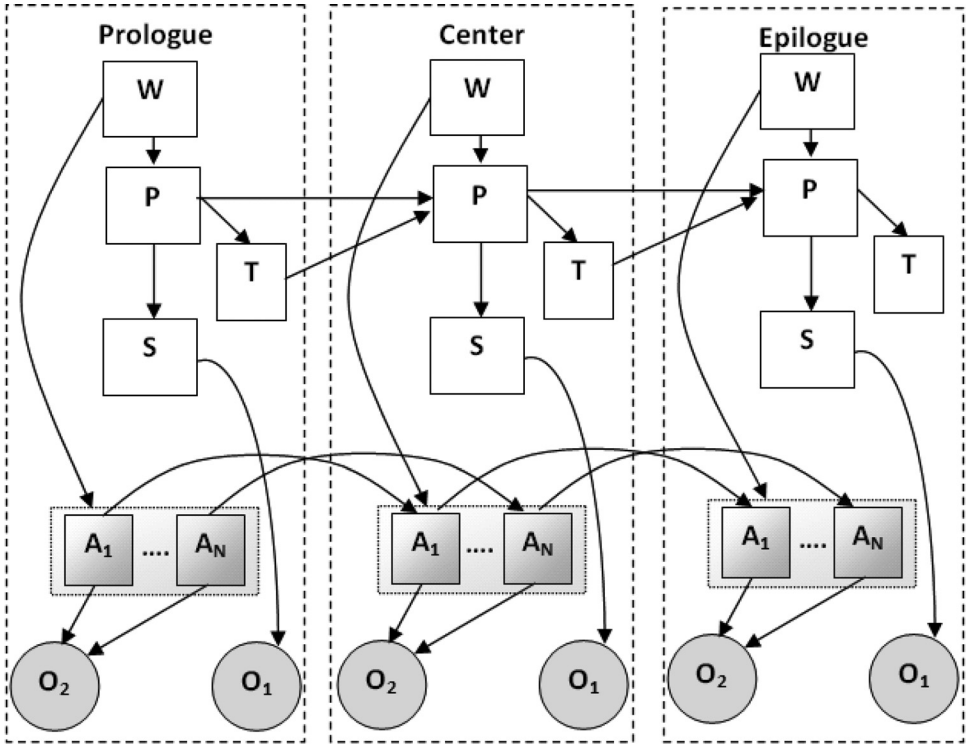

FIG. 6. G-DBN graphical model for a word (square/circular nodes represent discrete/continuous RVs, and shaded/unshaded nodes represent observed/hidden RVs).

replaced by TB (tongue body). Also note that the gesture RVs represent the gestural activations; i.e., whether the gesture is active or not, and hence are binary RVs. The gesture RVs do not have degree/location of constriction information, which was done deliberately to reduce the cardinality of the RVs in the DBN in order to prevent very large CPTs. Models with large CPTs were found to be intractable and very slow to train. Even if such models were trained after getting a good triangulation, they failed to generate any hypothesis during the test runs due to the model complexity.

The G-DBN word model has 16 states/word, which (like the default HMM based ASR models distributed with Aurora-2) has 11 whole word models (zero to nine and oh). There are 2 additional models for "sil" and "sp"; where "sil" had 3 states and "sp" had one state, respectively. The maximum number of mixtures allowed per state was four with vanishing of mixture-coefficients allowed for weak mixtures.

2. Observation RVs

As will be discussed in Sec. III A, experiments with the explicit gesture recognizer showed that (a) temporal contextualization of acoustic features and (b) use of estimated TV_ts, helped to improve the gesture recognition performance, we used contextualized acoustic features and TV_ts as the observation set O_2, which is used as the input to the gesture RVs. Although results to be presented in Sec. III A showed that the contextual window varied depending upon the TV, for the sake of simplicity we have used a common temporal context window of 190 ms for the observations tied to the gesture RVs in G-DBN. (In essence, the observations from the explicit gesture recognition experiment were used to help create the overall G-DBN architecture.) The remainder of the O_2 vector consists of the 13 cepstral coefficients in the form of MFCCs of with time contextualization. The 13D cepstral coefficients were mean subtracted and variance normalized and concatenated with 8D estimated TV_ts, then contextualized (covering 190 ms of speech data) by stacking cepstral coefficients from nine frames (selecting every 4th frame, 20 ms hop) where the 5th frame is centered at the current time instant. The resulting contextualized feature vector had a dimensionality of 189 ($= 9 \times 21$) which constitutes the second observation set O_2. Note that unlike the explicit gesture recognizer, all of the TV values are input to every gesture RV.

The continuous observed RV O_1 is input to the acoustic state random variable, S, and consists of acoustic observation in the form of MFCCs, (39D: 13 cepstral coefficients and their Δs and Δ^2s). In all the experiments reported here, the MFCCs were computed using an analysis window of 10 ms and a frame rate of 5 ms.

3. Training

Note that for the word recognition results reported in Sec. III C, the Aurora-2 clean training dataset was annotated with gestural scores using an iterative analysis-by-synthesis time-warping procedure presented by Nam *et al.* (2010). In this procedure, TaDA was used to create a prototype gestural score given an utterance, which was then aligned to the target utterance using a phone-landmark based iterative time-warping procedure. Note that the gestural annotation was performed only for the training set of Aurora-2, which in turn was used to train the G-DBN model discussed in this section.

III. EXPERIMENTS AND RESULTS

The experiments performed and the results obtained are reported in three sections. In Sec. A, we compare the performances of the four types of gesture recognition models

described in Sec. II C 2 (see Fig. 4), using the XRMB-SYN database; we also compare these gesture recognition performances to that obtained by Zhuang *et al*. (2009). In Sec. III B, we present the word recognition results for Aurora-2, obtained from using estimated TV_ts, recognized gestures (from the proposed ANN models), and acoustic features in an HMM-based word recognition system. Finally, in Sec. III C we present the results of word recognition using the gesture-based DBN architecture and compare its performance with published state-of-the-art results.

A. The gesture recognizer

The XRMB-SYN data was used to train-test the gesture models in Fig. 4, where 75% of the data was used for training, 10% for validation and the rest for testing. The network configurations (i.e., input contextual information, number of neurons and the delay chain in the feedback path of the AR-ANN) were optimized separately for each TV using the development set of XRMB-SYN. The networks in both stages contained a single hidden layer with tan- sigmoid activation functions, and were trained using the SCG algorithm up to a maximum epoch of 2500 iterations. The performance of the gesture recognizers was evaluated by first quantizing the gestural parameters obtained from the second stage based on a quantization code[8] constructed from the training set, and then computing a frame-wise gesture recognition accuracy score using Eq. (9) as specified below:

$$\text{Rec.Acc.} = \frac{N - S}{N} \times 100, \tag{9}$$

where N is the total number of frames in all the utterances and S is the number of frames having at least one of the three gestural parameters (activation, target and stiffness) wrongly recognized. Figure 7 presents the overall gesture recognition accuracy (averaged across the eight different gestures ignoring GLO and VEL) obtained from the four approaches using MFCCs and APs as the acoustic features.

Several observations can be made from the results presented in Fig. 7:

(1) GM-4 offers the best recognition accuracy for both MFCCs and APs. This is expected as it uses the ground-truth or actual TV_ts. In practice we cannot assume *a priori* knowledge of the actual TV_ts, which renders GM-4 infeasible for ASR applications. Nevertheless, GM-4 provides the accuracy ceiling that could be expected in case of an absolutely accurate TV_t-estimator in GM-3.
(2) For GM-4, using the APs as the acoustic feature gives higher recognition accuracy [at 1% significance level using the significance-testing procedure described by Gillick and Cox (1989)] than using MFCCs, which may indicate that APs provide a better acoustic parameterization than the MFCCs for gesture recognition.
(3) GM-1 uses only the APs or MFCCs for gesture recognition, and the APs show overall higher recognition accuracy (at 5% significance level) than the MFCCs, confirming the statement made in (2).
(4) GM-2 uses only the estimated TV_ts and the results show that the MFCCs offer better recognition accuracy than the APs (at 5% significance level).
(5) For GM-3, the AP and the MFCC based system gave recognition accuracies that are not significantly different from one another. From above observations we can see that APs do better for GM-1 (which requires only acoustic observation but no TV_t estimation for gesture recognition), while MFCCs perform better for GM-2 (which uses only estimated TV_ts but no acoustic observation for gesture estimation) and based on our prior observations (Mitra *et al*., 2010) we know that MFCCs perform better TV_t estimation compared to the APs. GM-3 uses both acoustic observations and TV_t estimation, hence the relative advantage of MFCCs and APs over one another in each of these two conditions are compensated, resulting in similar accuracies in this setup. Finally, no TV_t

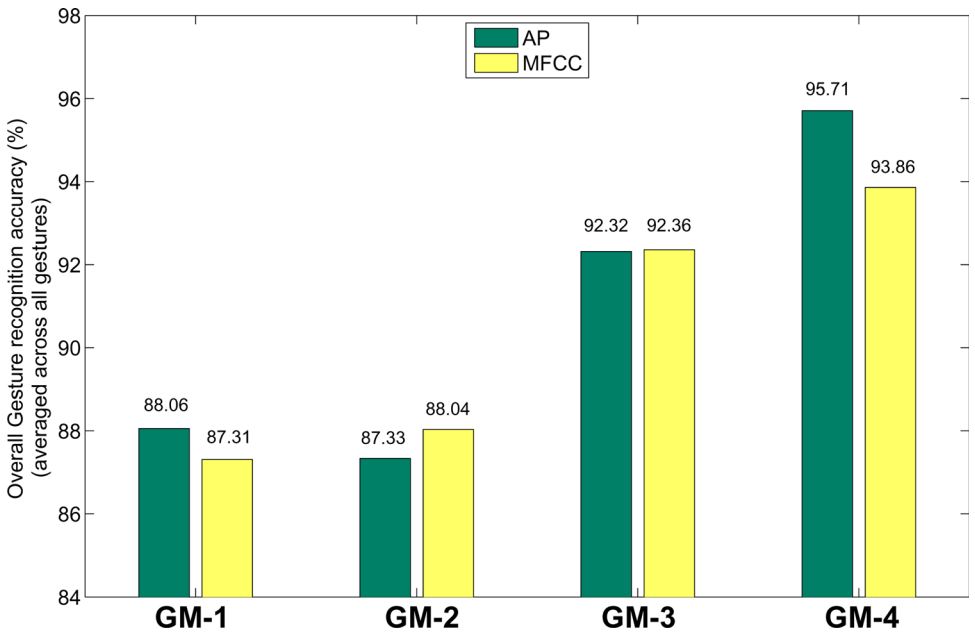

FIG. 7. (Color online) Average gesture recognition accuracy (%) obtained from the four gesture models (GM- 1 to GM-4) using AP and MFCC as acoustic feature.

estimation is required for GM-4, which is why the APs do better again.

(6) GMs 1, 2, and 3 are more realistic gesture-recognition architectures for ASR applications, as only the acoustic features are considered as the observable and the TV_ts in GMs 2 and 3 are estimated from the acoustic features. Amongst these three approaches, GM-3 offered the best recognition accuracy indicating that estimating TV_ts for gesture recognition is indeed beneficial. GM-3 is analogous to the use of tandem features used in ASR (Hermansky *et al.*, 2000) where an ANN is used to perform a nonlinear transform of the acoustic parameters to yield the estimated TV_ts, which in turn helps to improve the recognition of gestural scores when used in conjunction with the acoustic parameters. Note that the improvement resulting from the TV_ts cannot be just due to the increased number of input parameters. If that was the case, then the APs would be far superior to the MFCCs in GM-1.

Given these observations, we can state that the cascaded neural network gesture recognizer using acoustic features and estimated TV_ts as input will recognize gestures relatively more accurately than when only the acoustic features or the estimated TV_ts are used as the input.

Figure 8 presents the recognition accuracies individually for all gesture types, where GM-1 is only used for GLO and VEL and GM-3 is used for all of the remaining gestures. Figure 8 shows that using GM-1 the GLO and VEL gestures were recognized quite well (accuracy $> 98\%$). This observation is encouraging as it indicates that it is relatively simple to estimate parameters for these gestures from synthetic speech. The APs offered better recognition accuracy for the GLO, VEL, TBCL-V, TBCD-V and TBCD-C gestures; this was expected as the APs have specific features for capturing voicing [the periodic and aperiodic information using the approach specified in Deshmukh *et al.* (2005)] and nasalization information [using APs proposed in Pruthi (2007)], whereas the MFCCs have none. However, some APs rely on formant information and formant tracking for noisy speech is prone to errors, rendering the AP-based models to be unreliable for recognizing gestures from noisy speech. Thus, in our ASR experiment presented in Sec. III B we have selected the MFCC-based model, as our aim was to obtain gestures for performing word recognition experiments on natural utterances, both in clean and noisy conditions.

Table III presents the optimal configuration for the 2-stage cascaded gesture recognition model for each gestural type. Note that in the two stages of the cascaded model, different optimal context window lengths were found for gestural activation and parameter detection. The $\mathbf{\Delta}$ in Table III represents the order of the delay chain in the feedback path of the AR-ANN architecture used for gestural activation detection.

Note that for a given gesture, the optimal input feature context window for activation detection (i.e., for AR-ANN) is smaller compared to that for gestural parameter estimation (i.e., for FF-ANN). This difference might be because the recognizer could not effectively recognize a gesture's specified target until the corresponding TV_t reaches its target (requiring a larger window of observation) whereas activation can be recognized by simply detecting a constricting motion of a TV_t (requiring a smaller observation window). Also, the acoustic feature context windows for gesture-recognition are different than those used for TV_t estimation, where the optimal context window for the MFCCs and the APs was found to be 170 ms and 190 ms, respectively (Mitra *et al.*, 2010). Hence, there are three factors that may have contributed to the superior performance of GM-3 relative to GMs 1 and 2:

(1) GM-3 has the benefit of using three context windows (one each for TV_t estimation, activation detection and parameter estimation), and the associated power of the multi-resolution analysis they provide.

(2) GM-3 uses two streams of input information, the acoustic features and the estimated TV_ts, whereas GMs 1 and 2 uses only one of those two.

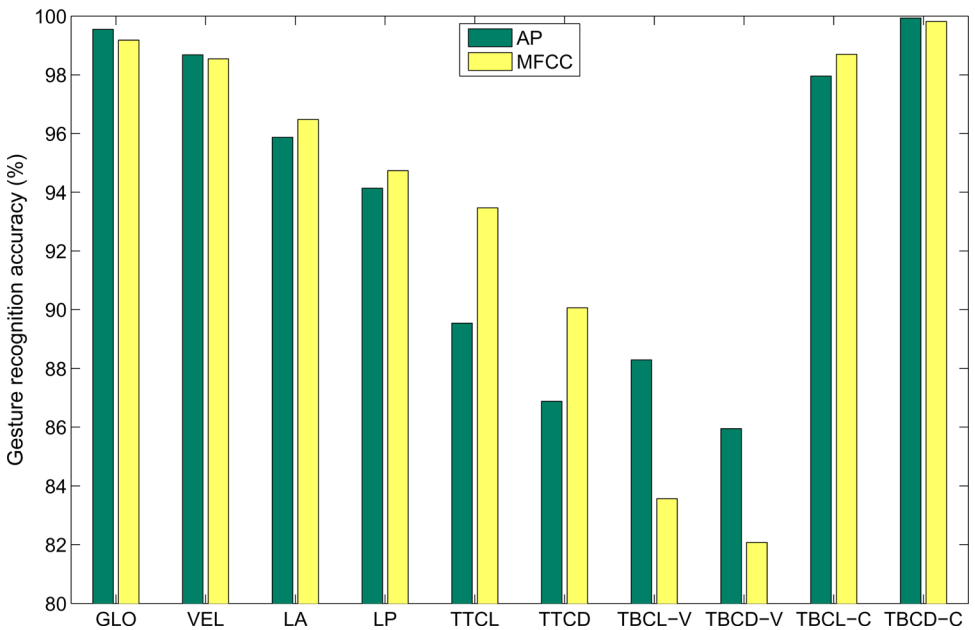

FIG. 8. (Color online) Gesture recognition accuracy (%) obtained for the individual gesture types using the cascaded ANN architecture, where the inputs for GLO and VEL were acoustic features only (i.e., AP or MFCC) while for the remainder, the input was defined by the concatenation of estimated TVs and acoustic features.

TABLE III. Optimal configuration for gesture recognition (activation and parameter) using approach-1 for GLO and VEL and approach-3 for the rest.

	AP			MFCC		
	Activation detection		Parameter estimation	Activation detection		Parameter estimation
Gesture	Δ	Context (ms)	Context (ms)	Δ	Context (ms)	Context (ms)
GLO	4	170	210	5	190	210
VEL	4	150	210	4	130	210
LA	3	90	210	10	90	210
LP	4	90	290	9	90	290
TTCL	4	90	210	4	90	210
TTCD	7	190	210	4	190	230
TBCLV	4	130	290	4	170	290
TBCDV	9	150	290	7	190	290
TBCLC	4	150	210	10	190	210
TBCDC	4	150	210	4	170	210

(3) Finally, as stated before, acoustic signals have higher bandwidth whereas speech gestures are quasi-stationary discrete units, having bandwidths close to zero. Hence trying to create a direct mapping between them will be prone to errors. TV_ts are smoothly varying trajectories (with bandwidth lower than the acoustic waveform but higher than gestures) that are not only coupled strongly with gestures but are also coupled well with the acoustic signal; hence using them as an intermediate representation turns out to be a better strategy.

To compare the performance of our system with an existing gesture recognition model published in the literature, we selected the system reported by Zhuang *et al.* (2009). Their system was selected for two reasons. First, it is the only system reported in the literature that obtains the fine grained gesture-level information that we are seeking (that is, representing gestures not only by their activation functions, but also with their target and stiffness parameters). Second, their system uses the same XRMB-SYN data that we have used in this study, except their train-test file sets are different than ours. The train and test lists in Zhuang *et al.* (2009) consists of 277 words for training and 139 words for testing from the XRMB-SYN dataset, without any word identity overlapping. Our gesture recognition models were retrained using the same 277 word training set. The gestural scores were transformed to an instantaneous gestural pattern vector (GPV) as discussed in Zhuang *et al.* (2009) and illustrated in Fig. 9. In Table IV, we present the F-score of the recovered gestural parameters (constriction targets and stiffness) from our models and compared that with respect to the reported results of Zhuang *et al.* Note that we have used MFCC as the acoustic observations for the four different approaches shown in Table IV below.

Table IV shows that the F-scores (%) reported by Zhuang *et al.* (2009) for the same task matches closely with that obtained from the gesture recognition models presented in this study. Zhuang *et al.* (2009) used groundtruth TV_ts which is also the case for our GM-4 model. However, the GM 1-3 models are more practical because they used acoustic parameters derived from the speech signal as input (GM 2 and 3 estimated TV_ts from the acoustic parameters). It is noteworthy that our GM-3 model, using estimated TVs performed better than the Zhuang model for several gesture types, despite using GM-3 using estimated TVs, compared to the use of ground truth TV's of Zhuang *et al.* Overall, Table IV reiterates the observations in Fig. 7 and confirms that use of TV_ts with acoustic observations result in better gesture recognition performance than using either of them alone. Since for both GLO and VEL, only the GM-1 model was trained (i.e., no TV_t information was used). Thus, their entries for GM-2, GM-3 and GM-4 are empty. Finally, note that the combined target and stiffness F-scores (first two rows in Table IV) are lower than the individual entries, this happened as the results from our models (GM-1 to GM-4) showed lowering of precision and recall values when

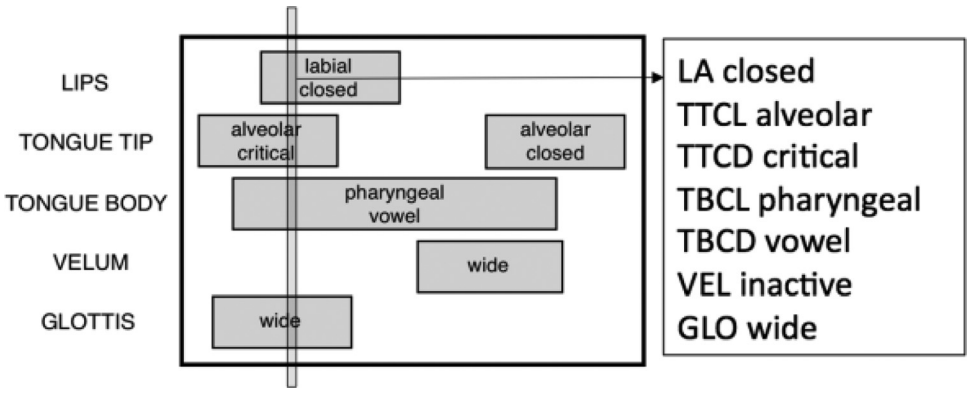

FIG. 9. Gestural score for the word "span." Constriction organs are denoted on the left and the gray boxes at the center represent corresponding gestural activation intervals. A GPV is defined to be the set of gestures active at any given point in time, as shown by the vertical slice in the figure.

TABLE IV. F-scores (%) of the recovered discretized gestural parameters as reported by Zhuang *et al*. (2009) and obtained from the different gesture recognition approaches presented in this study.

		Zhuang *et al*. (2009)		Models reported in this article			
		Uniform ergodic	GPB-bigram	GM-4	GM-3	GM-2	GM-1
Targ.		73.51	79.07	78.08	75.12	63.6	54.86
Stif.		80.79	84.50	85.98	81.95	71.59	64.38
Target	GLO	62.80	72.34	–	–	–	99.17
	VEL	64.79	75.21	–	–	–	98.53
	LA	69.44	77.28	98.90	96.54	95.01	94.75
	LP	78.03	84.83	99.68	94.82	93.53	85.96
	TTCL	64.14	69.14	97.93	94.57	92.76	91.36
	TTCD	63.05	68.44	96.19	90.11	82.73	87.66
	TBCL	78.56	82.90	84.22	82.71	72.46	72.98
	TBCD	83.16	86.01	80.34	81.93	71.10	68.14
Stiffness	LA	69.99	77.36	98.90	97.14	96.10	94.96
	LP	78.46	85.24	99.68	94.82	93.53	85.96
	TBCL	83.41	85.90	90.46	89.22	81.20	82.77
	TBCD	83.43	85.91	89.70	91.14	85.30	84.99

combined (to generate the combined target and stiffness values), hence resulting in lowering of their F-scores.

For the word recognition experiments presented in the next sub-section we used MFCCs as the acoustic feature, the GM-1 gesture recognition models for GLO and VEL, and the GM-3 models for the remaining eight gesture types.

B. Word recognition experiments using HMMs with estimated TVs and recognized gestures

The TV_t-estimator and gesture-recognition models were re-initialized and retrained with the synthetic speech database AUR-SYN. The trained TV_t-estimator and gesture-recognizer were applied to the training and testing sets (A and B) of Aurora-2 to obtain the corresponding set of estimated TV_ts and recognized gestures. We then performed word recognition experiments on the Aurora-2 corpus using the estimated TV_ts and recognized gestures along with the acoustic features. The HTK-based speech recognizer distributed with the Aurora-2 corpus (Pearce and Hirsch, 2000) was employed for the experiment. Training was performed on clean data and testing on noisy utterances. The raw recognized gestural scores were converted to vectors (GPVs) before being fed to the word recognizer; unlike the GPVs discussed above, these were not quantized and hence they represent the raw ANN posteriors (one for each of the three parameters [activation, target and stiffness] at a given tract variable site) from the gesture models. The acoustic features used as input to the Aurora-2 whole-word models were parameterized as MFCC or RASTA-PLP (Hermansky and Morgan, 1994), and were obtained using a 25 ms window with a 10 ms frame-advance. The estimated TV_ts and GPVs (sampled at 5 ms) were resampled for seamless concatenation with the acoustic features. The 39-dimensional acoustic features consisted of 13 feature coefficients, 13 velocity coefficients and 13 acceleration coefficients. We tested different combinations of GPVs, TV_ts and acoustic features (MFCC or RASTA-PLP), and also each of them singly as possible inputs to the word recognition system. The HMM used eleven whole word models ("zero" to "nine" and "oh") and two silence/pause models "sil" and "sp," each with 16 states. The number of Gaussian mixtures was optimized for each input feature set using a development set consisting of 200 files selected randomly from each noise type at clean condition and removed from the test set of Aurora-2. The remaining 801 files were used for testing. (Note: The Aurora-2 test set for each noise type at each SNR contains 1001 files). It was observed that for the case when input features were concatenations of acoustic features with the TV_ts and GPVs (i.e., MFCC+TV_t+GPV or RASTA-PLP+TV_t+GPV), the optimal number was five mixtures for the word models, and eight mixtures for the "silence/speech-pause" models. For all other input scenarios, the optimal number of Gaussians/mixture was three for word models and six for "silence/speech-pause" models.

Table V presents word recognition accuracies obtained using MFCCs and RASTA-PLPs with and without the TV_ts and the GPVs as inputs to the word recognizer. The two rows above the last one show the recognition accuracy when only TV_ts or GPVs were used as the input to the word recognizer. The estimated TV_ts and GPVs helped to improve the noise robustness of the word recognition

TABLE V. Overall Word Recognition accuracy.

	Clean	0–20 dB	−5 dB
MFCC	99.12	58.02	6.99
MFCC+TV	98.82	70.37	10.82
MFCC+TV+GPV	98.56	73.49	16.36
RASTA-PLP	99.01	63.03	10.21
RASTA-PLP+TV	98.96	68.21	12.56
RASTA-PLP+TV+GPV	98.66	75.47	19.88
TV	72.47	42.07	10.06
GPV	82.80	47.50	9.48
ETSI-AFE (European Telecommunications Standards Institute, 2007)	99.09	86.13	27.68

system when used in addition to the acoustic features (MFCC or RASTA-PLP). However, the estimated TV_ts and GPVs by themselves were not sufficient for word recognition, which indicate that the acoustic features and the articulatory parameters (TV_ts and GPVs) are providing complementary information; hence neither of them alone offers results as good as when used together. Note also that recognition accuracies of the GPVs were better than that of the TV_ts, implying that the GPVs are more discriminative than the TV_ts. The main factor behind the GPVs' failure to perform as well as the acoustic features for the clean condition is most likely the inaccuracy of the gesture-recognizers and TV_t estimator. These models were trained with only 960 synthetic utterances (AUR-SYN) which are roughly 11% of the size of the Aurora-2 training set (8440 utterances). Also note that the models were trained on synthetic speech and deployed on natural speech, hence the recognized gestures and the estimated TV_ts both suffer from acoustic mismatch. However, the results in Table V are encouraging in the sense that even with such inherent inaccuracies, the estimated TV_ts and the GPVs, when used with the acoustic features, provided improvement in word recognition performance. The last row in Table V shows the results from using the ETSI-AFE (European Telecommunications Standards Institute–Advanced Front End, 2007), which is amongst the state-of-the-art results reported in the literature for the Aurora-2 noisy digit recognition task. The ETSI-AFE has been specifically designed for improving noise-robustness of speech recognition systems and incorporates noise reduction in its front-end processing (Flynn and Jones, 2008). In the TV_t estimation or the gesture recognition steps, we have not incorporated any noise reduction or speech signal enhancement procedure. With the design of the TV_t estimator or gesture recognizer models that use noise robust acoustic features as input, we can expect to have performance better than the ETSI-AFE for spontaneous speech.

Figure 10 presents the overall word recognition accuracy (averaged across all noise types at all SNRs) when the MFCCs or RASTA-PLPs are used with and without TV_ts and the GPVs.

Figure 11 breaks down the word recognition accuracy (averaged across all noise types) for six different SNRs using the MFCCs and RASTA-PLPs as the acoustic features with and without the estimated TV_ts and GPVs. We have added here the word recognition accuracy obtained from using generalized spectral subtraction (GSS) speech enhancement (Virag, 1999), which shows better accuracy than that of using MFCCs only. Use of estimated TV_ts and GPVs in addition to the acoustic features (without any speech enhancement) provided higher recognition accuracy than that obtained from using GSS speech enhancement.

C. Word recognition experiments using the gesture-based Dynamic Bayesian Network

The G-DBN architecture was tested on the Aurora-2 database, using MFCC features and cepstral features from the ETSI-AFE. The results are shown in Table VI, compared to some state-of-the art results on Aurora-2 reported in the literature.

The MFCC+TV_t+GPV HMM system in Table VI is the result from Sec. III B, which used a left-to-right HMM word recognizer. Note that the gesture recognition models used in Sec. III B were trained using a small synthetic speech corpus. Consequently, the gesture recognition models may have inaccuracies when deployed on natural speech. This mismatch in conditions may be one of the reasons why the G-DBNs result in better performance compared to the MFCC+TV_t+GPV HMM system. Alternatively, SME and SME+MVN results are borrowed from Xiao *et al.* (2010). The MVA frontend processing was performed using the approach laid out in Chen and Bilmes (2007) (with ARMA order of 3) and the ETSI-AFE was obtained from the ETSI portal (European Telecommunications Standards Institute–Advanced Front End, 2007).

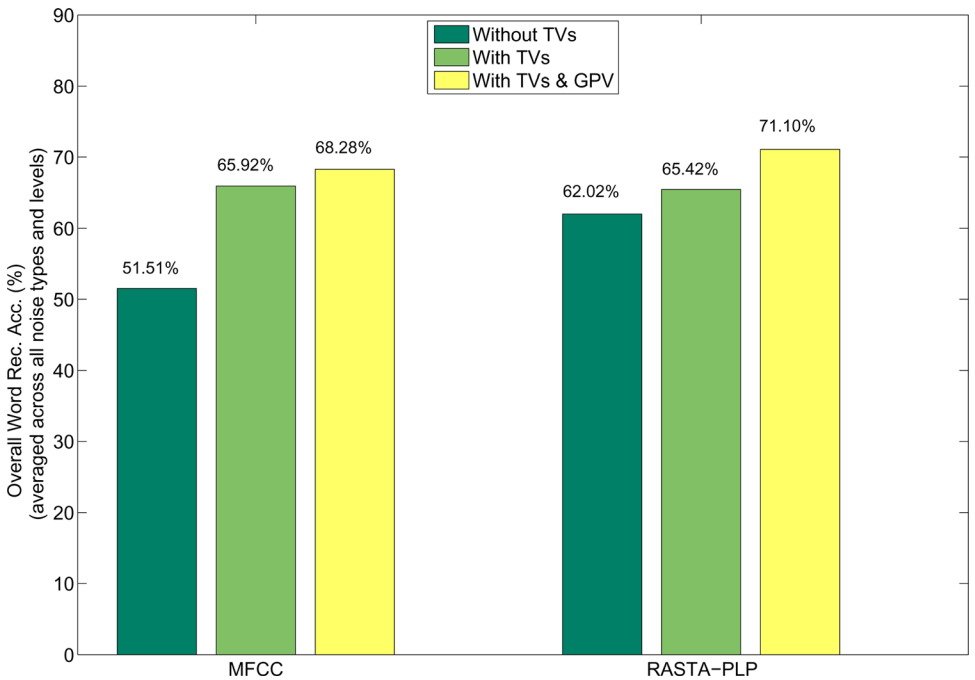

FIG. 10. (Color online) Overall word recognition accuracy using MFCC and RASTA-PLP with and without the estimated TVs and gestures.

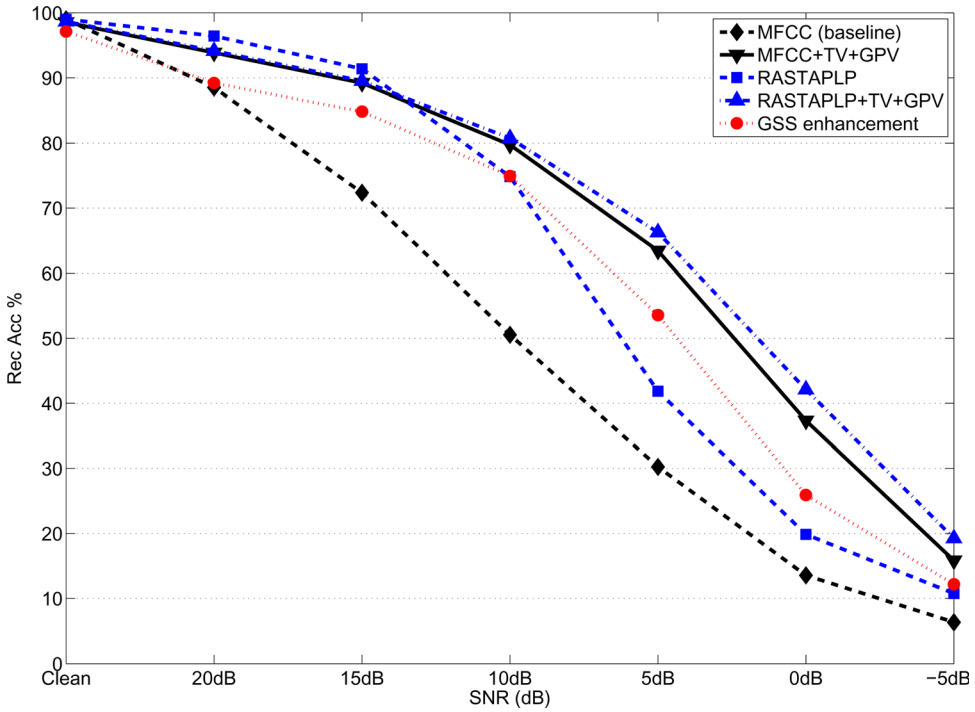

FIG. 11. (Color online) Word recognition accuracy (averaged across all noise types) at various SNR in using (a) the baseline MFCC, (b) MFCC+TV+GPV, (c) RASTA-PLP, (d) RASTA-PLP+TV+GPV, and (e) MFCCs after GSS based speech enhancement of the noisy speech.

For both of them, the word recognition experiments were carried out by us in house. The results from a maximum likelihood linear regression (MLLR1 and MLLR2) and feature compensation (FC) are borrowed from Cui and Alwan (2005). Note that the work by Cui and Alwan (2005) does not report results at −5 dB SNR. Table VI shows that for both noisy conditions, the combination of ETSI-AFE with the GDBN front end provided the best word recognition accuracy of all the systems tested. The table also shows that using the G-DBN improved performance over HMM for both the MFCCs and the ETSI-AFE inputs, showed a relative performance improvement of 37.76% and 0.63%, respectively, over the ETSI-AFE using HMM acoustic model (in the two noise conditions combined) So even with sophisticated enhancement, the gestural information provided by the G- DBN was able to show an improvement.

TABLE VI. Aurora-2 word recognition accuracies from G-DBN and other state-of-the art systems reported in the literature.

		Clean	0–20 dB	−5 dB
HMM	MFCCs	99.12	58.02	6.99
	MFCCs+TV_ts+GPVs	98.56	73.49	16.36
	Soft Margin Estimation (SME) (Xiao *et al.*, 2010)	99.64	67.44	11.70
	SME + Mean and Variance Normalization (MVN) (Xiao *et al.*, 2010)	99.68	86.01	24.9
	Mean, Variance Normalization and ARMA filtering (MVA) (Chen and Bilmes, 2007)	99.18	83.75	24.72
	MLLR1 (Cui and Alwan, 2005)	97.35	77.95	–
	MLLR2 (Cui and Alwan, 2005)	98.95	76.76	–
	Feature Compensation (FC) (Cui and Alwan, 2005)	99.00	83.50	–
	ETSI-AFE	99.09	86.13	27.68
G-DBN	MFCC	99.27	84.00	26.56
	ETSI-AFE	99.19	86.41	29.64

IV. DISCUSSION AND CONCLUSION

Representing the phonological structure of speech as a pattern of discrete, temporally overlapping gestures is an alternative to traditional phone-based representations (Browman and Goldstein, 1992). Here we tested the utility of using gestures as intermediate representations between the acoustic signal and lexical items in a word recognition system. To do so, it was necessary to develop a system to automatically extract discrete gestural representations from the acoustic signal, and we presented results from two different approaches for doing so.

In the first approach, we developed a cascaded neural network architecture for recognizing gestures from the acoustic waveform, trained on only synthetic speech generated by the TaDA model. Evaluation of the network's performance on gesture recognition using synthetic test utterances revealed successful recognition of gestures' activations and dynamical control parameters. This architecture was then used to recognize gestures for the natural speech of Aurora-2 corpus, and the recognized gestures were used in tandem with the original acoustics to perform word recognition experiments. Our results showed that adding the gestural representations to the baseline MFCC or RASTA-PLP features improved the recognition rates for noisy speech.

Such improvements in recognition accuracy under noise can be interpreted in the following way. The gestural representations can be thought of as lower dimensional projections of the acoustic signal. They can be computed from the acoustic signal but contain less information than a (rich) acoustic signal. For example, the acoustic signal contains information about talker identity, emotional state, physical environment, etc., none of which is expressed in the gestural representations. But the lower dimensional gesture representations are

designed (Browman and Goldstein, 1992) precisely to represent how words are distinguished from one another, so that in a situation in which the total amount of information is sharply reduced (noisy conditions), projecting available information into a form optimized for word discrimination is highly advantageous. The experiments showed that this gestural information was robust enough to provide a significant improvement to the recognition of natural speech data, despite its limitations: it was trained only on synthetic data and gestures were recognized independently of one another. When these limitations are removed, the contribution of gestures to recognition would be expected to increase.

In the second approach, we designed a gesture-based DBN word recognition system, which was trained with automatic gesture annotations of the Aurora-2 clean training corpus. This architecture (G-DBN) models the gestural activations as random variables with causal links to the acoustic observations that are learned during training. During recognition, the gestural activations are hidden random variables, so gestures do not need to be explicitly recognized, and multiple gestural possibilities at any time are maintained with their own probabilities. The G- DBN uses acoustic observations in the form of MFCCs or the ETSI-AFE, as well as the estimated TV_ts. The results show that the proposed architecture significantly improves the recognition performance of the connected digits in noise, both over a traditional MFCC-HMM system, and also over our first approach that required discrete gesture decisions before word recognition. Overall, G-DBN using ETSI-AFE showed the best recognition accuracy of all the systems we tested and its performance was found to be the best among some of the state-of-the- art noise-robust techniques, indicating that the G-DBN architecture has the potential to push state-of-the art noise-robustness beyond what has been previously published.

The results showing that gestural representations can aid word recognition in noise echo some recent findings with human listeners on the role of motor areas of the cortex in speech perception. Meister *et al.* (2007) showed that subjects were impaired in discriminating stop consonants in noise when transcranial magnetic stimulation (TMS) was applied to the premotor cortex, but were not impaired in a control task that was matched in difficulty level and response characteristics. D'Ausilio *et al.* (2009) found evidence for more specific motor cortex effects on perception in noise. Low levels of TMS applied to areas of the motor cortex controlling lips or tongue had selective effects during speech perception in noise: stimulating in lip areas enhanced recognition of labial stops while stimulating in tongue areas enhanced recognition of coronals. Thus, the kind of models developed here can be seen as having additional rationale—they are mimicking some aspects of how the human listener (who excels at this task) performs recognition in noisy environments.

One major challenge to the gesture-recognition models was that they were trained with limited "clean synthetic" speech and executed on "clean and noisy natural" speech from different speakers, which probably introduced severe acoustic mismatch to the models. They also suffered from limited amounts of training data. Future model experiments that are designed without these limitations can expect to achieve even better results. Also, the most powerful advantages of gestural representations can be expected in more natural connected speech materials with unstressed syllables, in addition to the stressed ones in digit strings. In such utterances, it has been shown that the same set of gestures are in play in reduced forms as in more careful forms, but with increased amounts of overlap, and decease in magnitude in space and time (Browman and Goldstein, 1989, 1992).

Gestures, if properly recognized, can not only benefit speech recognition tasks but can also have impacts on other speech-technology areas, e.g., visual speech, speech activity detection, speech enhancement, etc. Gestures along with TV_ts can be used to obtain the dynamics of the vocal tract shape and the movements of the different articulators, which in turn can help to create a visual representation of speech articulation. Such visual speech can have applications such as assistance for the hearing impaired, speech based animations, second language teaching, correcting speech dysfluencies etc. Gestures during pauses, for example, have been shown to differ systematically, depending on whether the pauses are grammatical or dysfluent (Ramanarayanan *et al.*, 2009). Finally gestures and TV_ts can be used in speech enhancement algorithms, where it is increasingly difficult to separate consonants overlapping with background noise. Since gestures and TV_ts specify constriction locations and degrees which indicate the consonantal place and manner information, it may be possible to employ them to extract the voiceless consonants from background noise.

ACKNOWLEDGMENTS

This work was supported by NSF Grants No. IIS0703859, No. IIS-0703048, and No. IIS0703782. The first two authors contributed equally to this study.

[1]Kirchhoff (1999) used pink noise at four different SNRs: 30 dB, 20 dB, 10 dB and 0 dB along with the clean speech.

[2]From now onwards, "groundtruth TV" denotes TVs generated by the TAsk Dynamic and Applications (TaDA) model of speech production.

[3]First we obtained all the unique words present in the XRMB corpus, purging all interjections and fillers and then performed stemming to get rid of all common endings. This resulted in a residual corpus of 420 distinct words.

[4]The optimality of the context window here refers to the window that provided the best TV estimation accuracy for a given feature set as observed from our prior experiments.

[5]The low-pass nature of the articulators was observed by Hogden *et al.* (1998), who stated that the articulatory trajectory usually has a smoother path and is defined by one that does not have any Fourier components over the cut- off frequency of 15 Hz.

[6]In the TaDA model, each TV is associated with a set of model articulators. For a given TV gesture, the articulator weight parameter specifies the different relative contributions of the associated articulators. The current study does not include the articulator weight parameters in the gestural estimation due to the greater complexity that would be entailed in the estimation process.

[7]We observed that GLO and VEL gestures can be recognized with an accuracy of around 99% using only acoustic features. This is because their targets are binary (either "open" or "closed"), as opposed to the larger number of possible target values for other TV gestures. Hence for GLO and VEL, only cascade model GM-1 was implemented.

[8]The number of quantization levels used to perform quantization of the gestures GLO, VEL, LA, LP, TTCL, TTCD, TBCLV, TBCDV, TBCLC and TBCDC are 6, 4, 8, 10, 14, 16, 10, 10, 4 and 4, respectively.

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

Bengio, Y., and Le Cun, Y. (**2007**), "Scaling learning algorithms toward AI," in *Large Scale Kernel Machines*, edited by L. Bottou, O. Chapelle, D. De-Coste, and J. Weston (MIT Press, Cambridge, MA), pp. 321–360.
Bilmes, J., and Zweig, G. (**2002**), "The graphical models Toolkit: An open source software system for speech and time-series processing," *Proceedings of ICASSP*, Orlando, FL, Vol. 4, pp. 3916–3919.
Browman, C., and Goldstein, L. (**1989**), "Articulatory gestures as phonological units," Phonology **6**, 201–251.
Browman, C., and Goldstein, L. (**1992**), "Articulatory phonology: An overview," Phonetica **49**, 155–180.
Chen, C., and Bilmes, J. (**2007**), "MVA processing of speech features," IEEE Trans. Audio Speech Lang. Processing **15**(1), 257–270.
Chomsky, N., and Halle, M. (**1968**), *The Sound Pattern of English* (MIT Press, Cambridge, MA), 484 pp.
Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**), "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**, 2421–2424.
Cui, X., and Alwan, A. (**2005**), "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," IEEE Trans. Speech Audio Processing **13**(6), 1161–1172.
D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, C. (**2009**), "The motor somatotopy of speech perception," Curr. Biol. **19**, 381–385.
Demuth, H., Beale, M., and Hagan, M. (**2008**), "Neural Network ToolboxTM**6**, User's Guide," The MathWorks Inc., Natick, MA. www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf (Last viewed June 28, 2010).
Deng, L., and Sun, D. (**1994**), "A statistical approach to automatic speech recognition using atomic units constructed from overlapping articulatory features," J. Acoust. Soc. Am. **95**(5), 2702–2719.
Deshmukh, O., Espy-Wilson, C., Salomon, A., and Singh, J. (**2005**), "Use of temporal information: Detection of the periodicity and aperiodicity profile of speech," IEEE Trans. Speech Audio Process. **13**(5), 776–786.
European Telecommunications Standards Institute–Advanced Front End (**2007**), "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Adv. frontend feature extraction algorithm; Compression algorithms," ES 202 050 Ver. 1.1.5,
Flynn, R., and Jones, E. (**2008**), "Combined speech enhancement and auditory modelling for robust distributed speech recognition," Speech Comm. **50**, 797–809.
Frankel, J., and King, S. (**2001**), "ASR—Articulatory speech recognition," *Proceedings of Eurospeech*, Aalborg, Denmark, pp. 599–602.
Frankel, J., and King, S. (**2005**), "A hybrid ANN/DBN approach to articulatory feature recognition," *Proc. of Eurospeech, Interspeech*, Lisbon, Portugal, pp. 3045–3048.
Frankel, J., Wester, M., and King, S. (**2004**), "Articulatory feature recognition using dynamic Bayesian networks," *Proc. of ICSLP*, Jeju, Korea, pp. 1202–1205.
Ghahramani, Z. (**1998**), "Learning dynamic Bayesian networks," in *Adaptive Processing of Temporal Information*, edited by C. L. Giles and M. Gori (Springer-Verlag, Berlin), pp. 168–197.
Ghosh, P., Narayanan, S., Divenyi, P., Goldstein, L., and Saltzman, E. (**2009**), "Estimation of articulatory gesture patterns from speech acoustics," *Proceedings of Interspeech*, Brighton, UK, pp. 2803–2806.
Gillick, L., and Cox, S. J. (**1989**), "Some statistical issues in the comparison of speech recognition algorithms," *Proceedings of ICASSP*, pp. 532–535.
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.
Hermansky, H., and Morgan, N. (**1994**), "RASTA processing of speech," IEEE Trans. Speech Audio Process., **2**, 578–589.
Hermansky, H., Ellis, D., and Sharma, S. (**2000**), "Tandem connectionist feature stream extraction for conventional HMM systems," *Proceedings of ICASSP*, Istanbul, Turkey, pp. 1635–1638.
Hogden, J., Nix, D., and Valdez, P. (**1998**), "An articulatorily constrained, maximum likelihood approach to speech recognition," Tech. Report, LA-UR–96-3945 (Los Alamos National Laboratory, Los Alamos, NM).
Hu, C., Zhuang, X., and Hasegawa-Johnson, M. (**2010**), "FSM-based pronunciation modeling using articulatory phonological code," *Proceedings of Interspeech*, pp. 2274–2277.
Jakobson, R., Fant, C. G. M., and Halle, M. (**1952**), "Preliminaries to speech analysis: The distinctive features and their correlates," MIT Acoustics Laboratory Technical Report 13 (MIT Press, Cambridge, MA).
Jordan, M. I., and Rumelhart, D. E. (**1992**), "Forward models-supervised learning with a distal teacher," Cogn. Sci. **16**, 307–354.
Juneja, A., (**2004**), "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. thesis, University of Maryland, College Park, MD.
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.
Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., and Sen, Z. (**2001**), "What kind of pronunciation variation is hard for triphones to model?," *Proceedings of ICASSP*, Utah, Vol. 1, pp. 577–580.
King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M. (**2007**), "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am. **121**(2), 723–742.
Kirchhoff, K. (**1999**), "Robust speech recognition using articulatory information," Ph.D. Thesis, University of Bielefeld, Germany.
Lapedes, A., and Farber, R. (**1988**), "How neural networks work," Technical Report LA-UR-88-418, Los Alamos National Library, Los Alamos, NM.
Lippmann, R. (**1997**), "Speech recognition by machines and humans," Speech Comm. **22**, 1–15.
Markov, K., Dang, J., and Nakamura, S. (**2006**), "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," Speech Comm. **48**, 161–175.
McGowan, R. S. (**1994**), "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Comm. **14**(1), 19–48.
Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (**2007**), "The essential role of premotor cortex in speech perception," Curr. Biol. **17**, 1692–1696.
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), pp. 2430.
Nam, H., Mitra, V., Tiede, M., Saltzman, E., Goldstein, L., Espy-Wilson, C., and Hasegawa- Johnson, M. (**2010**), "A procedure for estimating gestural scores from natural speech," *Proceedings of Interspeech*, Makuhari, Japan, pp. 30–33.
Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2010**), "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," IEEE J. Selected Topics Signal Process. **4**, 1027–1045.
Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2011**), "Speech inversion: Benefits of tract variables over pellet trajectories," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, pp. 5188–5191.
Mitra, V., Özbek, I., Nam, H., Zhou, X., and Espy-Wilson, C. (**2009**), "From acoustics to vocal tract time functions," *Proceedings of ICASSP*, pp. 4497–4500.
Neiberg, D., Ananthakrishnan, G., and Engwall, O. (**2008**), "The acoustic to articulation mapping: Non-linear or non-unique?," *Proceedings of Interspeech*, Brisbane, Australia, pp. 1485–1488.
Ostendorf, M. (**1999**), "Moving beyond the 'beads-on-a-string' model of speech," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Colorado, Vol. 1, pp. 79–83.
Pearce, D., and Hirsch, H. G. (**2000**), "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proceedings of ICSLP*, *ASR*, Beijing, China, pp. 181–188.
Pruthi, T. (**2007**), "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. Thesis, University of Maryland, College Park, MD.
Qin, C., and Carreira-Perpiñán, M. Á. (**2007**), "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," *Proceedings of Interspeech*, Antwerp, Belgium, pp. 74–77.
Ramanarayanan, V., Bresch, E., Byrd, D., Goldstein, L., and Narayanan, S. (**2009**), "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation," J. Acoust. Soc. Am. **126**(5), EL160–EL165.
Richardson, M., Bilmes, J., and Diorio, C. (**2003**), "Hidden-articulator Markov models for speech recognition," Speech Comm. **41**(2–3), 511–529.
Richmond, K. (**2001**), "Estimating articulatory parameters from the acoustic speech signal," Ph.D. Thesis, University of Edinburgh, UK.
Richmond, K. (**2007**), "Trajectory mixture density network with multiple mixtures for acoustic-articulatory inversion," *ITRW on Non-Linear Speech Processing*, *NOLISP-07*, Paris, France, pp. 67–70.

Rubin, P. E., Baer, T., and Mermelstein, P. (**1981**), "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 321–328.
Saltzman, E., and Munhall, K. (**1989**), "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**(4), 332–382.
Schmidbauer, O. (**1989**), "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations," *Proceedings of ICASSP*, pp. 616–619.
Stevens, K. (**1960**), "Toward a model for speech recognition," J. Acoust. Soc. Am. **32**, 47–55.
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.
Toutios, A., and Margaritis, K. (**2005**), "A support vector approach to the acoustic-to- articulatory mapping," *Proceedings of Interspeech*, Lisbon, Portugal, pp. 3221–3224.
Virag, N. (**1999**), "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Process. **7**(2), 126–137.
Westbury, J. (**1994**), "X-ray microbeam speech production database user's handbook," University of Wisconsin, Madison.
Wrench, A. A., and Hardcastle, W. J. (**2000**), "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar on Speech Production: Models and Data*, Bavaria, Germany, pp. 305–308.
Xiao, X., Li, J., Chng, E. S., Li, H., and Lee, C. (**2010**), "A study on the generalization capability of acoustic models for robust speech recognition," IEEE Trans. Audio Speech Lang. Process. **18**(6), 1158–1169.
Zhuang, X., Nam, H., Hasegawa-Johnson, M., Goldstein, L., and Saltzman, E. (**2009**), "Articulatory phonological code for word classification," *Proceedings of Interspeech*, Brighton, UK, pp. 2763–2766.