Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework

Vikramjit Mitra[1], Hosung Nam[2], Carol Y. Espy-Wilson[3]

[1]*Speech Technology and Research Laboratory, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, USA*
vmitra@speech.sri.com

[2]*Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511, USA*
nam@haskins.yale.edu

[3]*Institute for Systems Research & Department of ECE, University of Maryland, College Park, MD*
espy@umd.edu

***Abstract*— Articulatory Phonology models speech as spatio-temporal constellation of constricting events (e.g. raising tongue tip, narrowing lips etc.), known as articulatory gestures. These gestures are associated with distinct organs (lips, tongue tip, tongue body, velum and glottis) along the vocal tract. In this paper we present a Dynamic Bayesian Network based speech recognition architecture that models the articulatory gestures as hidden variables and uses them for speech recognition. Using the proposed architecture we performed: (a) word recognition experiments on the noisy data of Aurora-2 and (b) phone recognition experiments on the University of Wisconsin X-ray microbeam database. Our results indicate that the use of gestural information helps to improve the performance of the recognition system compared to the system using acoustic information only.**

I. INTRODUCTION

Conventional automatic speech recognition (ASR) systems suffer from acoustic variabilities in speech. Such variabilities can be due to background noises, speaker differences, differences in recording devices etc. Studies [1-11] have shown that articulatory information can potentially improve the performance of a speech recognition system and increase its robustness against noise contamination and speaker variation.

The motivation behind the use of articulatory information in speech recognition is to model coarticulation and reduction in a more systematic way. Coarticulation has been described in several ways, including the spreading of features from one segment to another [12], influence on one phone by its neighbouring phones, overlapping of gestures and so on [13, 14]. Articulatory Phonology argues that human speech can be decomposed into a constellation of vocal-tract constriction gestures. Gestures are defined at discrete constriction organs (lips, tongue tip, tongue body, velum and glottis) as discrete invariant action units. We intend to use the articulatory gestures to model speech for ASR.

Constriction at each organ is represented as its corresponding geometric features of the shape of the vocal tract tube, i.e., tract variables: the constriction degree and location (Table 1), of which geometric definition is described in Fig 1. Note that velum and glottis are defined by the constriction degree alone since their locations are fixed. TTCD and TBCD define the degree of constriction for tongue tip and tongue body and are measured in millimeters representing the aperture created for such constriction. TBCL and TTCL specify the location of the tongue tip and tongue body with respect to a given reference (F in Figure 1) and are measured in degrees. LP and LA are the protrusion and the aperture of the lips and are measured in millimeters. GLO and VEL are abstract measures that specify whether the glottis and velum are open/close, hence distinguishing for voiceless/voiced and nasal/oral sounds.

TABLE I
CONSTRICTION ORGAN, VOCAL TRACT VARIABLES

Constriction organ	Tract variables
Lips	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)

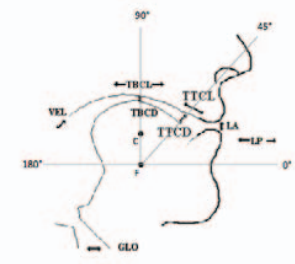

Fig. 1. Vocal tract variables at 5 distinct constriction organs, tongue ball center (C), and floor (F) [16, 17]

Fig 2 displays a set of gestures for the utterance "miss you" (named as a gestural score), which are represented as colored boxes. Their corresponding time-varying physical realizations are shown as smooth curves, called vocal tract constriction time functions or TVs. Note that a TV (more details in [15]) does not stand for a tract variable itself but its time function output, i.e. its temporal trajectory.

To be able to use the gestures as ASR units, they somehow need to be recognized from the speech signal. Additionally, we have observed that the accuracy of gesture recognition improves with prior knowledge of TVs [18], indicating the necessity of estimating TVs from the acoustic signal before performing the gesture based ASR.

In [18] we presented a Hidden Markov Model (HMM) based ASR architecture with Mel-Frequency Cepstral coefficients (MFCCs), estimated TVs and recognized gestures

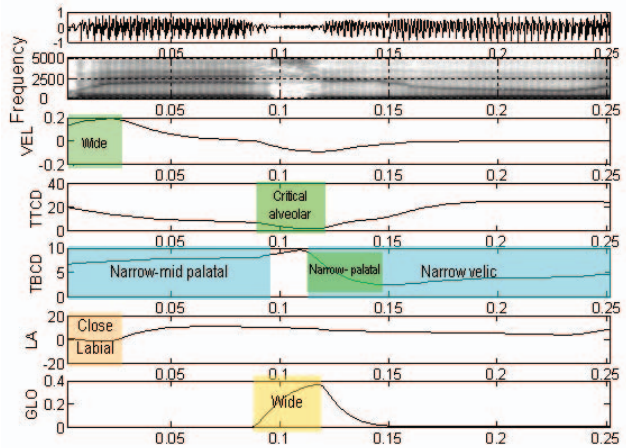

Fig. 2. Gestural activations for the utterance "miss you". Active gesture regions are marked by rectangular solid (colored) blocks. Smooth curves represent the corresponding TVs.

as input and obtained a 27.98% relative improvement over the MFCC-HMM baseline system for the noisy digit recognition task (clean condition training) of Aurora-2 [19]. Such an architecture required explicit recognition of articulatory gestures from the speech signal.

The choice of Dynamic Bayesian Network (DBN) as a backend in this paper was motivated by the fact that (a) gestures can be treated as hidden variables, eliminating the necessity of explicit recognition of articulatory gestures as a separate prior step and (b) they can model simultaneous time evolution and their temporal inter-dependency of multiple random variables. We proposed a gesture-based Dynamic Bayesian Network (G-DBN) architecture in [20], which employed MFCCs and estimated TVs as observations and obtained a 28.62% relative improvement over the MFCC-HMM baseline system for Aurora-2 word recognition task. However, in [20] the number of states/word had to be restricted to eight due to the complexity of the conditional probability tables (CPTs) tying the gesture variables with the word states. In this paper we present a reformulation of the DBN architecture, where the words (or phones) instead of the word states (or phone states) are functionally tied to the gesture variables. With the new DBN architecture (we name it as G-DBN II), we were able to train and evaluate 16 states/word models and hence compare it with some state-of-the-art systems.

To ascertain the generalizability of the G-DBN II architecture, we built context independent monophone models based on the architecture, using a part of the University of Wisconsin X-ray microbeam database (XRMB) [22] for training. We observed consistent improvement in phone error rates in both clean and noise-added conditions, which indicates that the articulatory representation not only improves noise robustness in ASR systems but also provides additional details to distinguish one phone from another compared to the acoustic-alone systems.

The organization of the paper is as follows: Section 2 provides a brief description of the data used in this paper, section 3 describes the gesture based DBN architecture proposed in this paper, followed by experiments and results in Section 4 and conclusions with future directions in Section 5.

II. The Data

The noisy word recognition tests were performed on the Aurora-2 database, which consists of connected digits spoken by 55 male and 55 female American English speakers, sampled at 8 kHz. The training set consists of 8440 clean utterances. Test sets A and B were used in the experiments reported here, where each of those sets have four subparts representing four different real-world noises (section A: subway, babble, car and exhibition; section B: restaurant, street, airport and train-station) at seven different signal-to-noise ratios (SNRs): clean, 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. Training in clean and testing in a noisy condition is used in all the experiments reported in here.

The XRMB speech production database [22] used in this study contains naturally spoken utterances both as isolated sentences and short paragraphs. The speech data were recorded from 47 different American English speakers (22 females and 25 males), where each speaker completed at most 56 tasks, each of which can either be read speech containing a series of digits, TIMIT sentences, or even as large as reading of an entire paragraph from a book. The sampling rate for the acoustic signals is 21.74 kHz. The data comes in three forms: text data consisting of the orthographic transcripts of the spoken utterances, digitized waveforms of the recorded speech and simultaneous X-ray trajectory data of articulator movements obtained from transducers (pellets) placed on the articulators. The trajectory data were recorded for the individual articulators, upper lip, lower lip, tongue tip, tongue blade, tongue dorsum, tongue root, lower front tooth (mandible incisor) and lower back tooth (mandible molar). The XRMB dataset was split into (a) 2000 utterances for training, containing speakers 11 to 54 and (b) 467 utterances for testing containing speakers 55 to 63. Note that speakers 1 to 11, 17, 22, 23 38, 47 and 50 did not exist in the XRMB database that was used in our experiment. The training and testing data split in XRMB are shown in Table II. The test set was contaminated with five different additive noises (subway, car, babble, restaurant noise borrowed from Aurora-2 and speech-shaped noise borrowed from speech-separation challenge 2006 dataset [32]). Noises were added at 4 different SNR levels, 20dB, 10dB, 5dB and 0dB. The addition of noise to the test set was done intentionally to perform phone recognition experiments with noisy acoustic data using acoustic models trained on clean speech.

TABLE II
DETAILS OF THE TRAIN & TEST DATA OF XRMB

	Train	Test
Number of utterances	2000	467
Number of speakers	38	9
Total number of words	59222	14026
Number of unique words	468	372
Number of hours	7.17	1.62

Aurora-2 and XRMB do not come with gestural annotation. The annotation was performed by (a) time aligning the phones using the Penn Phonetics Lab Forced Aligner [23] followed by (b) an iterative analysis-by-synthesis time-warping procedure proposed in [24]. In the analysis-by-synthesis time-warping procedure, Haskins Laboratories Task Dynamics and Application (TADA) model [25] is used to create a prototype gestural score given an utterance, which in turn is adapted to the target utterance using a phone-landmark based iterative time-warping procedure (more details in [24]). Note, that the gestural annotation was performed only for the training sets of Aurora-2 and XRMB databases.

III. The Gesture based DBN Architecture

The G-DBN architecture models the gestural activations (i.e. whether a gesture at a given constriction site is active or not) as discrete binary random variables that were observed during training and hidden during testing. The initial version of the G-DBN was presented in [20]. Compared to HMMs, DBNs have the flexibility to realize multiple hidden variables at a given time, which enables the DBN to model articulatory gestures as individual state variables, one for each articulatory gesture. DBN can also explicitly model the interdependencies amongst the gestures and can simultaneously perform gesture recognition and word recognition, eliminating the necessity to perform gesture recognition as a prior separate step before word recognition. For our DBN implementation we used the Graphical Models Tool-Kit (GMTK) [25], where conditional probability tables (CPT) are used to describe the probability distributions of the discrete random variables (RVs) given their parents, and Gaussian mixture models (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

$$w = \arg\max_i P(w_i \mid o) = \arg\max_{w_i} \frac{P(w_i)P(o \mid w_i)}{P(o)} = \arg\max_{w_i} P(w_i)P(o \mid w_i) \quad (1)$$

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 the words w are equally probable. Hence we are left with $P(o|w_i)$ which can be further simplified as

$$P(o \mid w) = \sum_s P(s, o \mid w) = \sum_s P(s \mid w)P(o \mid s, w) \approx \sum_s P(s_1 \mid w)P(o_1 \mid s_1, w)\prod_{i=2}^{n} P(s_i \mid s_{i-1}, w)P(o_i \mid s_i, w) \quad (2)$$

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 and then (2) can be reformulated as

$$P(o \mid w) \approx \sum_s P(s_1 \mid w)P(o_1 \mid s_1, a_1, w) \times \prod_{i=2}^{n} P(s_i \mid s_{i-1}, w)P(a_i \mid a_{i-1}, s_i)P(o_i \mid s_i, a_i, w) \quad (3)$$

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 or phone 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 issue we slightly modified equation (3) as follows

$$P(o \mid w) \approx \left[\sum_s P(s_1 \mid w)P(a_1 \mid w)P(o_{1,1} \mid s_1, w)P(o_{2,1} \mid a_1, w)\right] \times \prod_{i=2}^{n}\left[P(s_i \mid s_{i-1}, w)P(a_i \mid a_{i-1}, w)P(o_{1,i} \mid s_i, w)P(o_{2,i} \mid a_i, w)\right] \quad (4)$$

Equation (4) is based on the assumption that the gestural states and the word states (or phone states) are individual entities tied directly to the word (or phone) RVs. This is represented by the graphical model shown in Fig 3. We name the DBN architecture obtained from equation (4) as the G-DBN II architecture.

The G-DBN II 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 RVs (A_1 to A_N). The prologue and the epilogue in Fig. 3 denote the initial and the final frame(s) and the center represents the intermediate frames, which are unrolled in time to match the duration of a given utterance. For word (or phone) model, S represents the word (or phone) state, W represents the word (or phone) RV, P represents the word (or phone) position RV and T represents the word (or phone) transition RV. Note that in G-DBN II the word (or phone) state (S) is a functional parent of the observation O_1, which in turn is functionally linked to the word (or phone) RV (W) through the word (or phone) position RV (P). The gestural states (RVs A_1 to A_N) are now functional parents of observation O_2 and they in turn are functionally tied directly with the word (or phone) RV (W) instead of the word (or phone) states (S), which is the main difference between G-DBN in [20] and G-DBN II presented here. Another important modification in the G-DBN II architecture is the temporal contextualization of the observation set O_2, which is a consequence of our observation from [18], where we saw that temporal contextualization of the acoustic observations helped to improve the gesture recognition performance.

In the G-DBN II shown in Fig. 3, square/circular nodes represent discrete/continuous RVs, and shaded/unshaded nodes represent observed/hidden RVs. The continuous observed RV O_1 is the 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. The other continuous observed RV O_2 consists of only the cepstral coefficients of O_1 (ignoring the Δs and Δ^2s) with time contextualization. The 13D cepstral coefficients were mean subtracted and variance normalized and concatenated with 8D estimated TVs (obtained from a neural network based TV estimator, discussed in section IV), then

contextualized by stacking cepstral coefficients from nine frames (selecting every 4th frame) where the 5th frame is centered at the current time instant. The resulting contextualized acoustic feature vector had a dimensionality of 189 (=9x21) which constitutes the second observation set O_2.

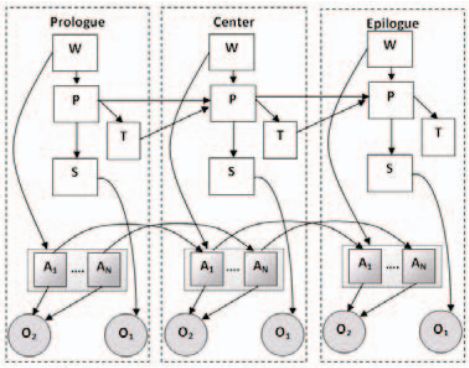

Fig. 3 G-DBN II graphical model for a word

We modelled 6 articulatory gestures as hidden RVs in the G-DBN II architecture, so *N* (the subscript of *A*) in Fig. 3 is 6. 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 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. Since the TVs were used as a set of observation and the TVs by themselves contain coarse target specific information about the gestures, it can be expected that the system has gestural target information to some extent. Both G-DBN and G-DBN II use 6 gesture RVs: GLO, VEL, LA, LP, TT and TB. Note that the gestural activations for TTCL and TTCD are identical, hence were replaced by a single RV TT (tongue tip) and the same was true for TBCL and TBCD, which were replaced by TB (tongue body).

IV. Experiments and Results

A. The TV Estimator

Our prior study [18] has shown that articulatory gestures can be recognized with a higher accuracy if the knowledge of TVs is used in addition to the acoustic parameters (MFCCs) as opposed to using the acoustic parameters alone. However in a typical ASR setup, the only available observable is the acoustic signal, which is parameterized as acoustic features. This indicates the necessity to estimate the TVs from the acoustic features. We have performed an exhaustive study using synthetic speech in [15] and observed that a feed-forward ANN can be used to estimate the TV trajectories from acoustic features.

The TV estimator was trained and optimized using the acoustic signals and TVs from speaker 12 of XRMB database. The groundtruth TVs were obtained from the annotated gestural scores, using the Haskins Laboratories' TADA [25] synthesizer. 76.8% of the data were used to train the TV estimator, 10.7% was used as a validation set and the rest was used for testing. The TV estimator takes in speech waveform sampled at 8KHz. The acoustic signal is parameterized as 13D MFCCs. The acoustic features and the TV trajectories were z-normalized (with std. dev. = 0.25). The resulting acoustic coefficients were scaled such that their dynamic range was confined within [-0.95, +0.95]. The acoustic features were created by stacking the scaled and normalized acoustic coefficients from nine frames (selecting every other frame), where the 5th frame is centered at the current time instant. The resulting contextualized acoustic feature vector had a dimensionality of 117 (= 9×13). Table III presents the correlation scores (Pearson's product-moment coefficient: PPMC) obtained between the estimated and the groundtruth TVs for the test set.

The TV estimator is a 4-hidden layer feed-forward ANN using tanh-sigmoid activation function with 117 input nodes and 8 outputs nodes (corresponding to 8 TVs shown in Table I). The number of hidden layers was restricted to 4 in order to reduce the training time as well as the network complexity. Indeed, no appreciable improvement of performance was observed with further addition of hidden layers. The ANN was trained for up to 4000 iterations and the number of neurons in each hidden layer of the ANN was optimized using the validation set, where the optimal number of neurons in each of the four hidden layer was found to be 225, 150, 225 and 25. Note that we have also explored other cepstral features, such as RASTA-PLP, PLPCC, acoustic phonetic features etc [26, 27] in addition to MFCCs and none of them strongly outperformed the MFCCs.

TABLE III
PPMC for the estimated TVs

GLO	VEL	LA	LP	TBCL	TBCD	TTCL	TTCD
0.853	0.854	0.801	0.834	0.860	0.851	0.807	0.801

B. G-DBN and G-DBN II

Both the G-DBN and G-DBN II architectures were tested on Aurora-2 database, using MFCC features and cepstral features from the ETSI-advanced [21] frontend. With G-DBN we trained and tested only 8 states/word models because of the complexities mentioned earlier. We trained a total of 11 whole word models (zero to nine and oh) and 2 models for 'sil' and 'sp'; where 'sil' had 3 states and 'sp' had one state. The maximum number of mixtures allowed per state was four with vanishing of mixture-coefficients allowed for weak mixtures.

The ETSI-advanced frontend has been proposed for the Distributed Speech Recognition (DSR) setup. ETSI-advanced frontend uses MFCCs, where the speech signal is sampled at 8 kHz, analyzed in blocks of 200 samples with an overlap of 60%, pre-emphasized with a factor of 0.9, and Hamming windowed for FFT-computation. It uses a power spectrum estimate before performing the filterbank integration and has a built-in noise reduction module. The advanced frontend uses the cepstral coefficients C_1 to C_{12} and the weighted combination of the frame-wise log-energy measure and the C_0

coefficient, resulting in a 13D feature vector. The final ETSI advanced feature consists of this 13D feature vector along with their Δ and the Δ^2 coefficients, yielding a 39D feature vector. Table IV presents the results obtained from the G-DBN and HMM backend using MFCC and ETSI-advanced as frontend on Aurora-2. Note that with G-DBN we could train 8 states/word model whereas the MFCC-HMM model had 16 states/word.

With G-DBN II we were able to train and test 16 state/word models. The maximum number of mixtures/state used was four with vanishing of mixture-coefficients allowed for weak mixtures. The results from the G-DBN II are shown in Table IV, compared to some state-of-the art results on Aurora-2 reported in the literature.

The MFCC+TV+Gesture-posterior system in Table IV is from our prior gesture-based HMM system [18], which uses artificial neural network to estimate TV-trajectories and gesture posteriors and feed the resultant along with the MFCCs to a left-to-right HMM word recognizer as input. Note that the TV estimator and gesture recognizers used in [18] were trained using a synthetic speech database; hence the models were not as accurate. The SME and SME+MVN results are borrowed from [28]. The MVA frontend processing was performed using the approach laid out in [29] (with ARMA order of 3) and the ETSI-advanced frontend was obtained from the ETSI portal [30]. For both of them, the word recognition experiments were carried out in house. The results from maximum likelihood linear regression (MLLR1 and MLLR2) and feature compensation (FC) are borrowed from [31]. Note that [31] does not report -5dB SNR results; hence that column is kept empty. Table IV shows that for both noisy cases, the ETSI-advanced provided the best word recognition accuracy when used with the G-DBN II frontend.

TABLE IV
AURORA-2 WORD RECOGNITION ACCURACIES FROM THE PROPOSED APPROACH V/S SOME STATE-OF-THE-ART RESULTS

		Clean	0-20dB	-5dB
HMM	MFCC	99.12	58.02	6.99
	MFCC+TV+Gesture posteriors [18]	98.56	73.49	16.36
	Soft Margin Estimation (SME) [28]	99.64	67.44	11.70
	SME + Mean and Variance Normalization (MVN) [28]	**99.68**	86.01	24.9
	Mean, Variance Normalization and ARMA filtering (MVA) [29]	99.18	83.75	24.72
	MLLR1 [31]	97.35	77.95	-
	MLLR2 [31]	98.95	76.76	-
	Feature Compensation (FC) [31]	99.00	83.50	-
	ETSI-advanced [21]	99.09	86.13	27.68
G-DBN	MFCC	98.52	78.77	17.42
	ETSI-advanced [21]	98.51	80.92	21.40
G-DBN II	MFCC	99.27	84.00	26.56
	ETSI-advanced [21]	99.19	**86.41**	**29.64**

At clean, however, SME+MVN showed the best result. Both MFCCs and ETSI-advanced frontend showed a relative performance improvement of 37.76% and 0.63% respectively over the HMM backend when they are used with G-DBN II backend.

To analyse how the proposed G-DBN II architecture performs for a phonetically more diverse dataset than Aurora-2, we performed phone recognition experiments on the XRMB dataset. The forced alignment data for XRMB from the Penn Phonetics Lab Forced Aligner contained around 64 distinct phones. Even though the forced aligner uses CMU pronunciation dictionary with vowel lexical stress, the stress information was ignored in our experiments, which resulted in 39 distinct phonemes and a symbol 'sp' for speech pause. The training and testing set consists of 5.12 million frames (≈ 7.17 hours) and 1.16 million frames (≈ 1.62 hours). The phone models were trained with clean data and evaluated with clean testing data and their noisy counterparts.

The G-DBN II model was modified for this experiment to have 3 state/phone (with two dummy states) context free monophone models. Each state was allowed to have up to 4 mixtures per state, with vanishing of weak mixtures allowed. Along with the G-DBN II based monophone model, two more DBN based context free monophone models were trained which had the identical setup as the G-DBN II model but only lacked the gesture RVs. The first model used only the 39D MFCCs as the acoustic observation (DBN-MFCC monophone model). The second model used the 39D MFCCs along with the 8D TV trajectories as input (DBN-MFCC-TV monophone model). The performance of the models was obtained by comparing the phone state sequences inferred by Viterbi decoding to the alignments from the Penn forced aligner. The phone error rates (PER) were obtained using the HTK toolkit and the results are shown in Table V.

Note that the DBN-MFCC and DBN-MFCC-TV monophone models are theoretically similar to HMM monophone models due to the lack of any gestural RV with temporal dependency. The G-DBN II model demonstrated a 13.9% and 2.5% absolute reduction in PER at clean and noisy conditions. These reductions demonstrate that articulatory gestures not only contain sufficient details that help to distinguish one phonetic category from another but also such representations are robust to noise. Also note that the gestural RVs in the G-DBN II architecture takes in contextualized MFCCs as input which may also be a significant contributor to the robustness of the monophone models. Table V shows that simply appending the TVs to the MFCCs helped to bring down the error rate of the monophone DBN models by 2% for noisy conditions. This result is consistent to our previous observations [33], in which the TVs have always demonstrated noise robustness in word recognition tasks.

TABLE V
XRMB PHONE ERROR RATES

		PER (%)	
		Clean	0-20dB
DBN	MFCC	51.33	78.87
	MFCC+TV	51.45	76.16
G-DBN II	MFCC	**37.57**	**73.63**

V. Conclusion

We presented an articulatory gesture based DBN architecture that models the gestures as discrete hidden random variables. Acoustic observations in the form of MFCCs, estimated TVs, and their contextualized counterpart were used as input. Our results show that the proposed architecture significantly improves the recognition performance over traditional MFCC-HMM system both in word recognition and phone recognition experiments. For word recognition experiments, the ETSI-advanced frontend showed the best recognition accuracy with our G-DBN II word recognizer and its performance was found to be the best amongst some of the state-of-the-art noise-robust techniques shown in Table IV.

Gestures by definition are multi-channel discrete features and the DBN architecture is ideal for modeling articulatory gestures. DBNs by virtue of their generic modeling capability can simultaneously model gestures, their temporal dependency and the acoustic model. Unfortunately the challenge in using a DBN is its degree of freedom, which may make a model overly complex and tediously slow if not impossible to decode. Hence, using simple graphical structures and efficient triangulations are crucial to designing and testing a DBN.

The phone recognition results validate that the proposed system can be generalized to a large vocabulary database. Unfortunately, the XRMB and Aurora-2 are the only databases that we have been able to gesturally annotate so far [24]. An important future direction would be to annotate large vocabulary speech databases such as the Wall Street Journal or Switchboard, so that the proposed technique could be explored on them. Also note that typically a gestural span is much larger than phone duration; hence phones may not be the best sub-word units for a gesture based speech recognizer. Syllable or demi-syllable models will be explored in future studies.

References

[1] K. Kirchhoff, G. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition", Speech Comm., vol.37, pp. 303-319, 2000.

[2] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information", PhD Thesis, University of Bielefeld, 1999.

[3] M. Richardson, J. Bilmes and C. Diorio, "Hidden-articulator Markov models for speech recognition", Speech Comm., 41(2-3), pp. 511-529, 2003.

[4] O. Schmidbauer, "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations", Proc. of ICASSP, pp.616-619, 1989.

[5] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal", Sig. Proc., 27(1), pp.65-78, 1992.

[6] L. Deng and D. Sun, "A statistical approach to ASR using atomic units constructed from overlapping articulatory features", J. of Acoust. Soc. Am., 95, pp.2702–2719, 1994.

[7] K. Erler and L. Deng, "Hidden Markov model representation of quantized articulatory features for speech recognition", Comp., Speech & Lang., Vol.7, pp.265–282, 1993.

[8] J. Frankel and S. King, "ASR - Articulatory Speech Recognition", Proc. of Eurospeech, pp.599-602, Denmark, 2001.

[9] J. Frankel, K. Richmond, S. King and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces", Proc. of ICSLP, Vol.4, pp.254-257, 2000.

[10] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop", Proc. of ICASSP, Vol. 4, pp. 621-624, 2007.

[11] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein and E. Saltzman, "Articulatory Phonological Code for Word Classification", Proc. of Interspeech, pp.2763-2766, UK, 2009.

[12] R. Daniloff and R. Hammarberg, "On defining coarticulation", J. of Phonetics, Vol.1, pp. 239-248, 1973.

[13] C. Browman and L. Goldstein, "Articulatory Gestures as Phonological Units", Phonology, 6: 201-251, 1989.

[14] C. Browman and L. Goldstein, "Articulatory Phonology: An Overview", Phonetica, 49: 155-180, 1992.

[15] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, "Retrieving Tract Variables from Acoustics: a comparison of different Machine Learning strategies", IEEE J. of Selected Topics on Sig. Proc., Vol. 4(6), pp. 1027-1045, 2010.

[16] P. Mermelstein, "Articulatory model for the study of speech production", J. Acoust. Soc. of Am., 53(4), pp.1070–1082, 1973.

[17] C. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures", J. of Phonetics, Vol. 18, pp. 299-320, 1990.

[18] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, "Robust word recognition using articulatory trajectories and Gestures", Proc. of Interspeech, pp. 2038-2041, Japan, 2010.

[19] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In Proc. ISCA ITRW ASR2000, pp. 181-188, Paris, France, 2000.

[20] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, "Gesture-based Dynamic Bayesian Network for Noise robust Speech Recognition", Proc. of ICASSP, pp. 5172-5175, 2011.

[21] ETSI ES 202 050 Ver. 1.1.5, (2007). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithms.

[22] Westbury "X-ray microbeam speech production database user's handbook", Univ. of Wisconsin, 1994.

[23] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus", J. Acoust. Soc. of Am., 123(5), pp. 3878, 2008.

[24] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Espy-Wilson and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from natural speech", Proc. of Interspeech, pp. 30-33, Japan, 2010.

[25] H. Nam, L. Goldstein, E. Saltzman and D. Byrd, "Tada: An enhanced, portable task dynamics model in matlab", J. Acoust. Soc. of Am., 115(5), 2, pp. 2430, 2004.

[25] J. Bilmes, "GMTK: The Graphical Models Toolkit", SSLI Laboratory, Univ. of Washington, October 2002.

[26] V. Mitra, I. Özbek, H. Nam, X. Zhou and C. Espy-Wilson, "From Acoustics to Vocal Tract Time Functions", Proc. of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4497-4500, 2009.

[27] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, "Speech Inversion: Benefits of Tract Variables over Pellet Trajectories", Proc. of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5188-5191, Prague, Czeck Rep., 2011.

[28] X. Xiao, J. Li, E.S. Chng, H. Li and C. Lee, "A Study on the Generalization Capability of Acoustic Models for Robust Speech Recognition" , IEEE Trans. Audio, Speech & Lang. Process, 18(6), pp. 1158-1169, 2010.

[29] C. Chen and J. Bilmes, "MVA Processing of Speech Features", IEEE Trans. on Audio, Speech and Lang. Processing, 15(1), pp.257-270, 2007.

[30] http://portal.etsi.org/stq/kta/DSR/dsr.asp

[31] X. Cui and A. Alwan, "Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR", IEEE Transs. on Speech and Audio Processing, Vol. 13(6), pp. 1161-1172, 2005.

[32] M. Cooke, J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition", Journal of Acoustic Society of America, Vol. 120, pp 2421-2424, 2006.

[33] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Tract variables for noise robust speech recognition", IEEE Trans. on Audio, Speech and Language Processing, pp. 1913-1924, 2011