Articulatory Information for Noise Robust Speech Recognition

Vikramjit Mitra, *Student Member, IEEE*, Hosung Nam, *Member, IEEE*, Carol Y. Espy-Wilson, *Senior Member, IEEE*, Elliot Saltzman, and Louis Goldstein

***Abstract*—Prior research has shown that articulatory information, if extracted properly from the speech signal, can improve the performance of automatic speech recognition systems. However, such information is not readily available in the signal. The challenge posed by the estimation of articulatory information from speech acoustics has led to a new line of research known as "acoustic-to-articulatory inversion" or "speech-inversion." While most of the research in this area has focused on estimating articulatory information more accurately, few have explored ways to apply this information in speech recognition tasks. In this paper, we first estimated articulatory information in the form of vocal tract constriction variables (abbreviated as TVs) from the Aurora-2 speech corpus using a neural network based speech-inversion model. Word recognition tasks were then performed for both noisy and clean speech using articulatory information in conjunction with traditional acoustic features. Our results indicate that incorporating TVs can significantly improve word recognition rates when used in conjunction with traditional acoustic features.**

***Index Terms*—Articulatory phonology, articulatory speech recognition, artificial neural networks (ANNs), noise-robust speech recognition, speech inversion, task dynamic model, vocal-tract variables.**

I. Introduction

SPONTANEOUS speech typically has an abundance of variability, which poses a serious challenge to current state-of-the-art automatic speech recognition systems (ASR). Such variability has three major sources: 1) the environment, introducing different background noises and distortions, 2) the speaker, introducing speaker-specific variations such as dialectical–accentual–idiosyncratic contextual variation, and 3) the recording device, which introduces channel variations and other signal distortions.

Manuscript received March 15, 2010; revised July 06, 2010 and October 12, 2010; accepted December 08, 2010. Date of publication December 30, 2010; date of current version July 15, 2011. This work was supported in part by the National Science Foundation under Grants IIS0703859, IIS-0703048, and IIS0703782. V. Mitra and H. Nam contributed equally to this work. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

V. Mitra and C. Y. Espy-Wilson are with Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: vmitra@glue.umd.edu; espy@glue.umd.edu).

H. Nam is with the Haskins Laboratories, New Haven, CT 06511 USA (e-mail: nam@haskins.yale.edu).

E. Saltzman is in joint appointment with the Department of Physical Therapy and Athletic Training, Boston University, MA 02215 USA and also with the Haskins Laboratories, New Haven, CT 06511 USA (e-mail: esaltz@bu.edu).

L. Goldstein is with the Department of Linguistics, University of Southern California, Los Angeles, CA 90089 USA and also with the Haskins Laboratories, New Haven, CT 06511 USA (e-mail: louisgol@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2103058

A. Noise Robustness in ASR

Several approaches have been proposed to incorporate noise robustness into ASR systems, which can be broadly grouped into three categories: 1) the front-end based approach; 2) the back-end based approach; and 3) the missing feature theory.

The front-end based approaches usually aim to generate relatively contamination-free information for the back-end classifier or model. Such approaches can be grouped into two sub-categories. First, the noisy speech signal is enhanced by reducing the noise contamination (e.g., spectral subtraction [1], computational auditory scene analysis [2], modified phase opponency [3], speech enhancement and auditory modeling using the ETSI system [4], etc.). Second, features effective for noise robustness are employed in ASR systems (e.g., RASTAPLP [5], Mean subtraction, Variance normalization and ARMA filtering (MVA) post-processing of cepstral features [6], cross-correlation features [7], variable frame rate analysis [8], peak isolation [9], and more recently the ETSI basic [10] and advanced [11] front-ends, etc.).

The back-end based approach incorporates noise robustness into the back-end statistical model of the ASR system [usually a hidden Markov model (HMM)] for different speech segments. The goal of the back-end based systems is to reduce the mismatch between the training and the testing data. One such approach is to train the back-end models using data that contain different types of noise at different levels [12]. However, a shortfall to such a system is the necessity of knowledge of all possible noise type at all possible contamination levels, which renders the training data immensely huge if not unrealizable. An alternative is to adapt the back-end to the background noise. For instance, parallel model combination (PMC) [13] uses the noise characteristic and the relation between the clean and noisy speech signals to adapt the Gaussian mixture means and covariances of clean acoustic HMMs toward the true distributions of the noisy speech features. Usually such a transformation is fairly accurate but computationally expensive because the model parameters need to be updated constantly for non-stationary noise. Maximum-likelihood linear regression (MLLR) [14] performs model adaptation by rotating and shifting the Gaussian mixture means of clean HMMs using linear regression without using any prior knowledge of the background noise. Piecewise-linear transformation (PLT) was proposed [15] for a modified version of MLLR where different noise types are clustered based on their spectral characteristics and separate acoustic models are trained for each cluster at different signal-to-noise ratios (SNRs). During recognition, the best matched HMM is selected and adapted by MLLR.

The third approach is the missing feature theory [16], [17], which assumes that in noisy speech some spectro-temporal regions are so noisy that they can be treated as missing or unreliable. The missing feature approach computes a time–frequency reliability mask to differentiate reliable regions from the unreliable ones where the mask can be binary [16] or real valued [17]. Once the mask is computed, the unreliable components are dealt with by two different approaches: 2) data imputation [16] where the unreliable components are re-estimated based on the reliable components and 2) marginalization [16] where only the reliable components are used by the back-end for recognition. Bounded marginalization (BM) was proposed in [18] which generally outperform the "plain" marginalization. BM uses the knowledge that the unreliable data is bounded and the knowledge of such bounds is used to constrain the upper and lower bounds of the integral used for obtaining the likelihood of the incomplete data vector.

B. Articulatory Information for Contextual Variation

In the absence of noise, the major sources of variability in speech are speaker differences and contextual variation (commonly known as coarticulation). Typically, speaker differences are addressed by adapting the acoustic model to a particular speaker. Contextual variation is accounted for by using tri-phone or quin-phone based ASR systems that represent speech as a sequence of non-overlapping phone units [19]. However, such tri- or quin-phone models often suffer from data sparsity and capture contextual influence only from the immediate neighboring phones [20]. Indeed, coarticulation can have contextual influence beyond the immediate neighbors, and hence such models may fail to adequately account for coarticulatory effects [21].

It has been suggested [22] that the variations that occur in speech can be accounted for by incorporating speech production knowledge, which in turn may improve the performance of ASR systems. In a typical ASR application, the only known observable is the speech signal and speech production knowledge (typically articulatory dynamics) is unknown (such data may be available for research purposes, but cannot be assumed to be available for real-world applications). Hence, speech production related information needs to be estimated from the speech signal. Deciphering articulatory information from the speech signal and exploiting it in ASR has been widely researched and some of the prominent approaches are presented here.

Feature-Based Systems: Most of the initial research [23]–[25] in trying to incorporate speech production knowledge into ASR systems focused on deciphering appropriate features to capture articulatory dynamics and events, commonly known as articulatory features (AFs). One of the earliest AF-based ASR systems was proposed by Schmidbauer [26], who used 19 AFs (describing the manner and place of articulation) to perform HMM-based phone recognition of German speech and reported an improvement of 4% over the Mel-frequency cepstral coefficient (MFCC)–HMM baseline. These features showed less variance in recognizing different phonemic classes and were more robust against speaker differences as compared to the standard MFCC–HMM baseline. Deng [27] proposed an ASR system where the HMM states generated a trended-sequence of observations that were piece-wise smooth and continuous. Deng and his colleagues used 18 multi-valued AFs [28], [29] describing the place of articulation, horizontal and vertical tongue body movement, and voicing information. They reported an average classification improvement of 26% over the conventional phone-based HMM architecture for a speaker-independent task. Phone recognition on the TIMIT dataset showed a relative improvement of about 9% over the MFCC-HMM baseline.

King *et al.* [30] used artificial neural networks (ANNs) to recognize and generate AFs for the TIMIT database. They explored three different feature systems: binary features proposed by Chomsky *et al.* [31], traditional phonetic features defining manner and place categories, and features proposed in [32] and reported almost similar recognition rates for all of them. A comprehensive literature survey on the use of AFs and speech production model motivated ASR architectures is presented in [33].

Articulatory Trajectories and Their Use in ASR: Using articulatory trajectory information is more challenging than AFs in the sense that it involves retrieving articulatory dynamics from the speech signal, which is called "speech-inversion." This inverse problem is traditionally known to be ill-posed [34] as it is not only nonlinear but also non-unique. Nonlinearity arises due to the quantal-nature [35] of speech and non-uniqueness happens because different vocal tract configurations can yield similar acoustic realizations.

One of the earliest works on speech-inversion was by Atal *et al.* [36] who used temporal decomposition to predict the corresponding vocal tract configuration from acoustic signal. Multi-layered perceptrons (MLPs) also have been used by many studies [34], [37], [38] to obtain articulatory information from the speech signal. Ladefoged *et al.* [39] used linear regression to estimate the shape of the tongue in the midsagittal plane, using the first three formant frequencies in constant-vowel segments. Codebook-based approaches have also been proposed [40], [41] for speech inversion. Richmond [34] proposed mixture density networks (MDNs) to obtain flesh-point trajectories (also known as pellet trajectories) as conditional probability densities of the input acoustic parameters. He compared his results with that from ANNs and showed that MDN can directly address non-uniqueness in speech inversion. Recently, studies by Qin *et al.* [42] and Neiberg *et al.* [43] stated that "non-uniqueness" may not be so critical an issue but nonlinearity is more critical for speech inversion.

Inversion studies involving articulatory trajectories have been mostly confined to predicting such dynamics efficiently and accurately, and understanding their functional relationship with the acoustics. Due to the difficulty in estimating them, only a few ASR results [44], [45] have been known to use such articulatory dynamics. An alternative is to use actual articulatory recordings directly into the ASR system, but such a setup is not desirable for real-world applications. Frankel *et al.* [44] developed a speech recognition system that uses a combination of acoustic and articulatory features as input, where the articulatory trajectories are modeled using phone-specific linear dynamic models (LDMs). They showed that using articulatory data from direct measurements in conjunction with MFCCs resulted in a performance improvement by 9% [45] over the system using MFCCs only. Such an improvement did not hold when the articulatory data was estimated from the acoustic signal [45].

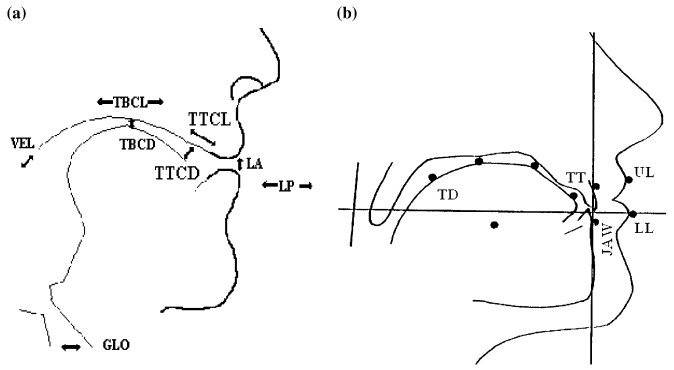


Fig. 1. (a) Eight tract variables from five distinct constriction locations. (b) Pellet placement locations according to [51].

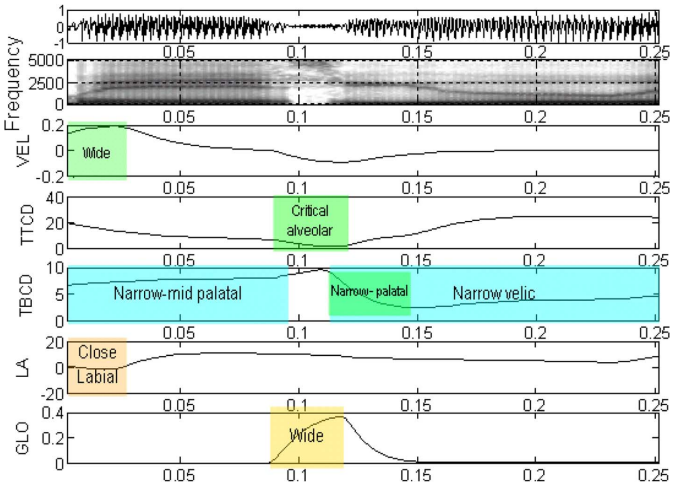


Fig. 2. Gestural score for the utterance "miss you." Active gesture regions are marked by rectangular solid (colored) blocks. Smooth curves represent the corresponding tract variable trajectories (TVs).

Articulatory Gestures and Tract Variables: Speech variations such as coarticulation have been described in several ways, including the spreading of features from one segment to another [46], influence on one phone by its neighbor and so on. Articulatory phonology [47], [48] defines coarticulation as a phenomenon that results from overlapping vocal tract constrictions or *gestures*. An utterance is represented by a constellation of gestures known as the gestural score as shown in Fig. 2. Note unlike AFs, gestural offsets are not aligned with acoustic landmarks. Gestures [49] are constricting actions produced by five distinct organs/constrictors (lips, tongue tip, tongue body, velum, and glottis) as in Fig. 1(a), and defined as eight vocal tract constriction variables (henceforth, tract variables) as in Table I. The tract variables describe geometric states of the shape of the vocal tract tube in terms of constriction degree and location of the constrictors. Each gesture is represented as a critically damped, second-order differential equation [49], shown in (1), where M, B, and K are mass, damping coefficient and stiffness parameters of each tract variable (represented by z) and z_0 is the target position of the gesture. An active gesture is specified by its activation onset and offset times and parameter values:

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

Given a gestural score, the tract variable trajectories (henceforth, TVs[1]) are derived using the TAsk-Dynamic and Applications (TADA) model [49], [50], which is a computational implementation of articulatory phonology. Fig. 2 shows the utterance "miss you," and its corresponding gestural scores and TVs as computed by TADA. Note TVs are continuous time functions whose dynamics are determined by the corresponding gestural specifications (see Fig. 2) whereas AFs are typically discrete units whose boundaries are determined from acoustic landmarks.

Articulatory pellet trajectories, which have overwhelmingly been used in the literature for speech inversion [34], are flesh-point information (e.g., electromagnetic midsagittal articulographic or EMA) representing positional (x-y coordinate) information of transducers [or pellets, Fig. 1(b)] placed on the different articulators [51]. Unlike pellet trajectories, which are absolute measures in Cartesian coordinates, TVs are relative measures and suffer less from non-uniqueness [52]. For example, TV description of a tongue tip stop will always exhibit a value of zero for TTCD (distance of tongue tip from palate), even though the pellet positions will differ depending on the location of pellets on an individual's vocal tract, the vowel context, etc. We have previously shown [53] that TVs can be estimated more accurately than pellet trajectories from speech signal.

C. Articulatory Information for Noise Robust ASR

Incorporating speech production knowledge into ASR systems was primarily motivated to account for coarticulatory variation. Kirchhoff was the first to show [54] that such information can help to improve noise-robustness of ASR systems as well. She and her colleagues [54], [55] used a set of heuristically defined AFs, which they identified as pseudo-articulatory features. Their AFs represent speech signal in terms of abstract articulatory classes such as: voiced/unvoiced, place and manner of articulation, lip-rounding, etc. However, their AFs do not provide detailed numerical description of articulatory movements within the vocal tract during speech production. They showed that their AFs in combination with MFCCs provided increased recognition robustness against the background noise, where they used pink noise at four different SNRs. They concluded that the AFs and MFCCs may be yielding partially complementary information since neither alone provided better recognition accuracy than when both used together. In a different study, Richardson *et al.* [56] proposed the hidden articulatory Markov model (HAMM) that models the characteristics and constraints analogous to the human articulatory system. The HAMM is essentially an HMM where each state represents an articulatory configuration for each di-phone context, allowing asynchrony among the articulatory features. They reported that their articulatory ASR system demonstrated robustness to noise and stated that the articulatory information may have assisted the ASR system to be more attuned to speech-like information.

In this paper, we demonstrate that articulatory information in the form of TVs estimated from the speech signal can improve the noise robustness of a word recognizer using natural speech

[1]Note we use "TVs" to refer to tract variable trajectories, i.e., the time functions of the tract variables, which should be distinguished from the tract variables themselves.

TABLE I
CONSTRICTION ORGAN, VOCAL TRACT VARIABLES, THEIR UNIT OF MEASUREMENT, AND DYNAMIC RANGE

Constriction organ	Vocal tract variables	*Unit*	*Dynamic range*	
			Max	*Min*
Lip	Lip Aperture (LA)	mm	27.00	-4.00
	Lip Protrusion (LP)	mm	12.00	8.08
Tongue Tip	Tongue tip constriction degree (TTCD)	mm	31.07	-4.00
	Tongue tip constriction location (TTCL)	degree	80.00	0.00
Tongue Body	Tongue body constriction degree (TBCD)	mm	12.50	-2.00
	Tongue body constriction location (TBCL)	degree	180.00	87.00
Velum	Velum (VEL)	-	0.20	-0.20
Glottis	Glottis (GLO)	-	0.74	0.00

when used in conjunction with the baseline acoustic features. Previously, we have shown [53] that the TVs can be estimated more accurately compared to pellet trajectories and we demonstrated that estimation of the TVs from speech is essentially a nonlinear process, where the estimation performance improves as the nonlinearity in the inversion process increases. We observed [53] that a 3-hidden layer feed-forward (FF) ANN offers reasonably accurate TV estimates compared to other machine-learning approaches (support vector regression, trajectory mixture density networks, distal supervised learning, etc.). In this paper we use the 3-hidden layer FF-ANN to estimate TVs from speech signal, and a Kalman smoother postprocessor to retain their characteristic smoothness. Our work is unique in the following ways.

1) Unlike the results reported by Frankel *et al.* [44], [45], we do not use flesh-point measurements (pellet trajectories) of the different articulators. Instead, we are using the vocal tract constriction trajectories or TVs, which are less varying than the pellet trajectories [52], [53]. None of the work available in literature evaluated the articulatory information (in the form of TVs) estimated from the speech signal under noisy conditions. In the present study, we show not only that TVs can be estimated more robustly from noise-corrupted speech compared to pellet trajectories, but also that the estimated TVs do a better job than pellet trajectories when applied to word recognition tasks under noisy conditions.
2) The work presented by Frankel *et al.* [44], [45] used LDM at different phone contexts to model the articulatory dynamics for clean speech, whereas we are using the TV estimates (without any phone context) directly into an HMM-based word recognizer for the recognition task.
3) Kirchhoff *et al.*'s work [54], [55] though uses articulatory information for noise robust speech recognition; their AFs do not capture the dynamic information about articulation but describe only the critical aspects of articulation. They are mostly hypothesized or abstract discrete features derived from acoustic landmarks or events and are not directly obtained from actual articulatory events. On the contrary, TVs provide actual articulatory dynamics in the form of location and degree of vocal tract constrictions in the production system.
4) Kirchhoff *et al.*'s work dealt with only pink noise at four SNR levels (30, 20, 10, and 0 dB), whereas we report our results on eight different real-world noise types (subway, car, babble, exhibition, train-station, street, airport, and restaurant) at six different SNRs (20, 15, 10, 5, 0, and -5 dB). Richardson *et al.* [56] used hypothetical AFs obtained at diphone context. Their noise robustness experiment was very limited in scope, and used stationary white Gaussian noise at 15-dB SNR only.
5) Finally, we present a study in which articulatory information is used across different acoustic feature sets and front-end processing methods to verify whether the benefits observed in using such articulatory information are specific to particular features or are consistent across features.

Earlier in [57], we proposed that TVs can potentially improve the noise-robustness of MFCC-based ASR systems, using only two noise types: car and subway noise. In this paper we extend that study by performing ASR experiments using six more noise types and show that the noise-robust nature of the TVs hold for other acoustic features (e.g., RASTAPLP) as well. In addition, we present ASR results from using only TVs as input and show that they outperform MFCCs at very low SNRs. The TV estimation models in this paper are more robust and well trained compared to those used in our earlier work [57]. Finally, to justify the selection of TVs, we performed ASR experiments (both in noisy and clean conditions) and evaluated the noise-robustness of the TVs compared to that obtained by using conventional pellet trajectories.

The organization of the paper is as follows. Section II provides a brief introduction to the dataset used in our experiments and their parameterization. Section III describes the 3-hidden layer FF-ANN architecture for TV estimation. Section IV presents the experiments, results and discussions followed by the conclusion in Section V.

II. DATASET AND SIGNAL PARAMETERIZATION

This study aims to obtain a proof-of-concept that estimated TVs can help to improve the noise robustness of ASR systems. To train a model for estimating TVs from speech, we require a speech database containing groundtruth TVs. Unfortunately, no such database is available at present. For this reason, TADA along with HLsyn [58] (a parametric quasi-articulator synthesizer developed by Sensimetrics Inc.) was used in our work (as shown in Fig. 3) to generate a database that contains synthetic speech along with their articulatory specifications. From text input, TADA generates TVs, simulated pellet trajectories and

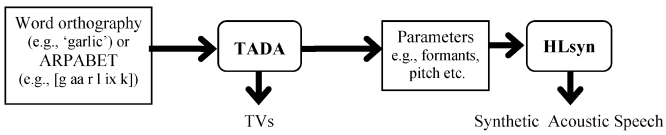


Fig. 3. Flow diagram for generating synthetic speech and the associated articulatory information using TADA and HLsyn.

other parameters, some of which are used by HLsyn to create the corresponding synthetic speech.

To create the dataset for training the TV-estimator, we selected 960 utterances from the training set of Aurora-2 [59], where each utterance contains a sequence of digits. Aurora-2 was created from the TIdigits database consisting of connected digits spoken by American English speakers, sampled at 8 kHz. The Arpabet for each digit sequence was input to TADA, which generated the corresponding TVs (refer to Table I), vocal tract area function, formant information, etc. The pitch, gender and formant information was input to HLsyn for generating the corresponding synthetic speech. The sampling rate of the synthetic speech and TVs are 8 kHz and 200 Hz, respectively. We named this dataset as AUR-SYN, where 70% of the files were randomly selected as the training-set and the rest as the testing-set. The testing files were further corrupted with subway and car noise at six different SNR levels similar to the Aurora-2 corpus. The noisy test section was created solely for testing the TV-estimator's performance under noise contamination. Executing TADA and HLsyn is expensive computationally. For example, to generate synthetic speech and its associated articulatory information for a mono-syllabic word such as 'one', TADA+HLsyn requires 85 seconds of CPU time in an AMD Athlon 64 dual-core 2.20-GHz processor with 3.5 GB of RAM. Hence, generating 8000 such mono-syllabic words would require almost 8 days of CPU processing time. Note that the training set of Aurora-2 consists of 8440 utterances, where each utterance can have more than one digit, indicating that the generation of synthetic speech and its associated articulatory information for the whole training set of Aurora-2 would result in CPU processing time of much more than 8 days. Because of these facts we had to limit the number of utterances to ensure a reasonable data generation time. Currently, we are annotating TVs and gestures [60] for the X-ray microbeam database [51] and the clean training set of Aurora-2 database, which will help us to build natural speech trained TV-estimators in future.

For TV estimation, speech signal was parameterized as MFCCs, where 13 cepstral coefficients were extracted at the rate of 200 Hz with an analysis window of 10 ms. The MFCCs and TVs were z-normalized and scaled to fit their dynamic ranges into $[-0.95, +0.95]$. It has been stated [34] and we have also observed [53], [61] that incorporating dynamic information helps to improve the speech-inversion performance, for which the input features were contextualized before being fed to the TV-estimator. The contextualized features for a given frame at t ms, were selected from a context-window of duration d ms ($d \geq 10$ ms), where feature vectors are evaluated between $[t - d/2, t)$ ms and $(t, t + d/2]$ ms and concatenated with the feature vector at t ms, where the features vectors are selected at 10-ms interval. Previously [53], we observed that the optimal TV estimation context window for the MFCCs is 170 ms and its dimension after contextualization is 221.

To perform our noise robustness experiments, we used test set A and B from Aurora-2, which contain eight different noise types at seven different SNR levels. Training in clean condition and testing in noisy scenario is used in all the experiments reported here.

III. TV Estimator

ANNs have been used by many [34], [38] for speech inversion. Compared to other architectures, ANNs have lower computational cost both in terms of memory and execution speed [34]. Further, ANNs can perform a complex nonlinear mapping of M input vectors (for our case, acoustic features, i.e., contextualized MFCCs) into N output vectors (for our case, TVs). In such architecture, the same hidden layers are shared across all the output TVs, which allows the ANN to capture any cross-correlation that TVs may intrinsically have amongst themselves [53]. The 3-hidden layer FF-ANN specification used in this paper is based on our prior analysis [53], where the number of neurons for the three layers is selected to be 150, 100, and 150, respectively. The FF-ANN is trained with back-propagation using scaled conjugate gradient as the optimization rule. A tan-sigmoid activation function is used as excitation for all of the layers.

The estimated TVs were found to exhibit substantial estimation error. We thus applied a Kalman smoother to the TV estimates, which improved the accuracy by ensuring the inherent smoothness of TVs [53], [61]. This is a direct consequence of the observation made in [62], which claimed that articulatory motions are predominantly low pass in nature with a cutoff frequency of 15 Hz.

IV. Experiments and Results

We aim to test the possibility of using the estimated TVs as input to the word recognition task on Aurora-2 and examine whether they can improve the recognition accuracies in noise. The details of the experiments are described in the following subsections. In Section IV-A, we first present the TV estimation results for the synthetic speech data for clean and noisy conditions. In Section IV-B, we then apply the synthetic-speech-trained TV-estimator on the natural utterances of Aurora-2 to estimate their corresponding TVs. In Section IV-C, we perform word recognition experiments using the estimated TV inputs, and further compared their performances when combined with acoustic features (MFCCs and RASTAPLP) and various front-end processing methods (such as MVA post-processing of acoustic features [6] and ETSI basic [10] and advanced [11] front-ends).

A. TV Estimation in Clean and Noisy Condition for AUR-SYN (Synthetic Speech)

The performance of the FF-ANN based TV-estimator is evaluated using two quantitative measures: root mean-squared error (RMSE) and Pearson product-moment correlation (PPMC) coefficient. RMSE gives the overall difference between the original and the estimated articulatory trajectories, whereas PPMC

TABLE II
RMSE AND PPMC FOR THE CLEAN SPEECH FROM AUR-SYN

	No-smoothing		Kalman smoothed	
	RMSE	PPMC	RMSE	PPMC
GLO	0.0196	0.9873	0.0191	0.9880
VEL	0.0112	0.9874	0.0101	0.9900
LA	1.0199	0.9654	0.9054	0.9734
LP	0.2257	0.9795	0.1986	0.9841
TBCL	2.2488	0.9966	2.0097	0.9973
TBCD	0.4283	0.9882	0.3841	0.9907
TTCL	2.9758	0.9806	2.8108	0.9830
TTCD	1.2362	0.9893	1.1722	0.9905

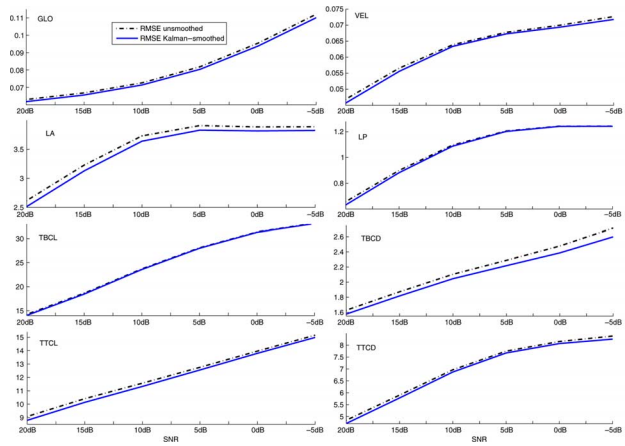


Fig. 4. RMSE of estimated TVs for AUR-SYN (synthetic speech) at different SNRs for subway noise.

indicates the strength of a linear relationship between two variables:

$$RMSE=\sqrt{\frac{1}{N}(e-t)^T(e-t)} \tag{2}$$

$$r_{\text{PPMC}}=\frac{N\sum_{i=1}^{N}e_it_i-\left[\sum_{i=1}^{N}e_i\right]\left[\sum_{i=1}^{N}t_i\right]}{\sqrt{N\sum_{i=1}^{N}e_i^2-\left(\sum_{i=1}^{N}e_i\right)^2}\sqrt{N\sum_{i=1}^{N}t_i^2-\left(\sum_{i=1}^{N}t_i\right)^2}} \tag{3}$$

where $\hat{\tau}$ and τ represent the estimated and the groundtruth TV vector having N data points. RMSE provides a performance measure in the same units as the measured articulatory trajectories. The FF-ANN TV-estimator was trained with the training set of AUR-SYN and the results are obtained using the test-set. Table II presents RMSE and PPMC of the estimated TVs for the clean set of AUR-SYN with and without using the Kalman smoothing. Table II shows that using the Kalman smoother helped to reduce RMSE and increase PPMC for the clean test set. As evident from Table I, some TVs have different measuring units (e.g., TBCL and TTCL are measured in degrees) from others (e.g., LA, LP, TBCD, and TTCD are all measured in mm), which should be considered while interpreting the RMSE values.

Figs. 4 and 5 show RMSE and PPMC plots, respectively, of the estimated TVs at different SNRs from the test set of AUR-SYN corrupted with subway noise. As SNR decreases,

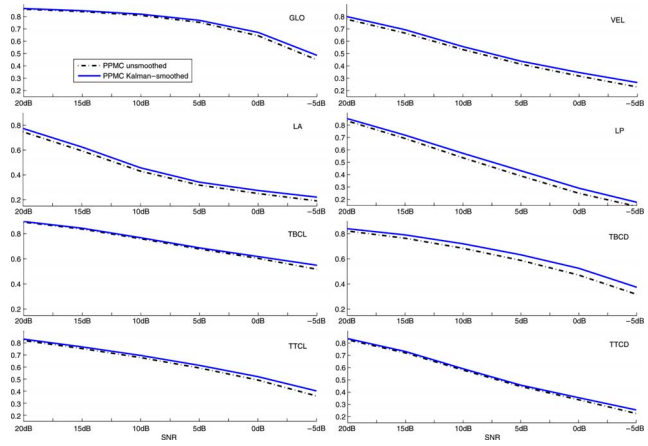


Fig. 5. PPMC of estimated TVs for AUR-SYN (synthetic speech) at different SNRs for subway noise.

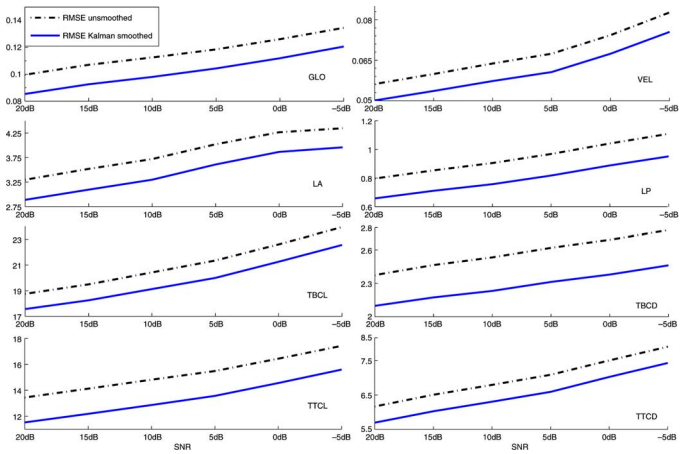


Fig. 6. RMSE (relative to clean condition) of estimated TVs for Auora-2 (natural speech) at different SNRs for subway noise.

the RMSE of the estimated TVs increases and their PPMC decreases, this indicates that the estimation deteriorates with decrease in SNR. Using Kalman smoothing results in lower RMSE and higher PPMC at a given SNR. The car noise part of the AUR-SYN test-set shows a similar pattern.

B. TV Estimation in Clean and Noisy Condition for Aurora-2 (Natural Speech)

The FF-ANN TV-estimator presented in the last section (which was trained with the clean synthetic speech from AUR-SYN) was used to estimate TVs for the natural speech of the Auroa-2 database. The estimated TVs were then Kalman-smoothed. Since there is no known groundtruth TVs in Aurora-2, RMSE and PPMC cannot be computed directly. We instead compared the unsmoothed or Kalman-smoothed estimated TVs from different noise types and levels to the corresponding unsmoothed or Kalman-smoothed estimated TVs from clean utterances, to obtain the relative RMSE and PPMC measures. Figs. 6 and 7 show that the relative RMSE increases and the PPMC decreases as SNR decreases for the subway noise section of Aurora-2, and Kalman smoothing helps to improve the relative RMSE and the PMMC. Note that the TV estimates for the natural utterances showed a relatively lower PPMC compared to those of the synthetic utterance (see Figs. 4 and 5). This may be due to the mismatch between the

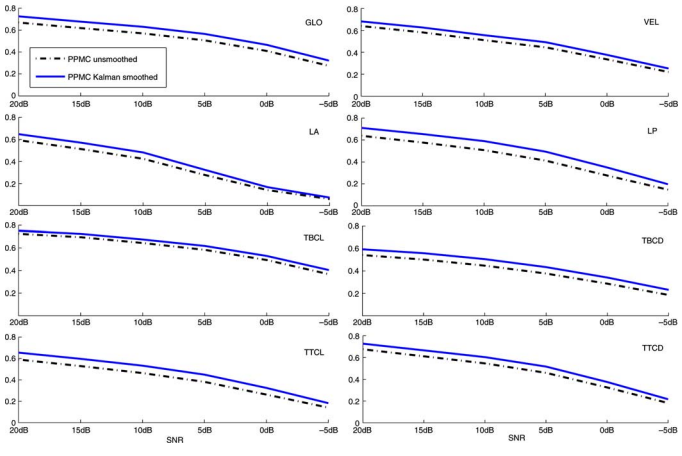


Fig. 7. PPMC (relative to clean condition) of estimated TVs for Auora-2 (natural speech) at different SNRs for subway noise.

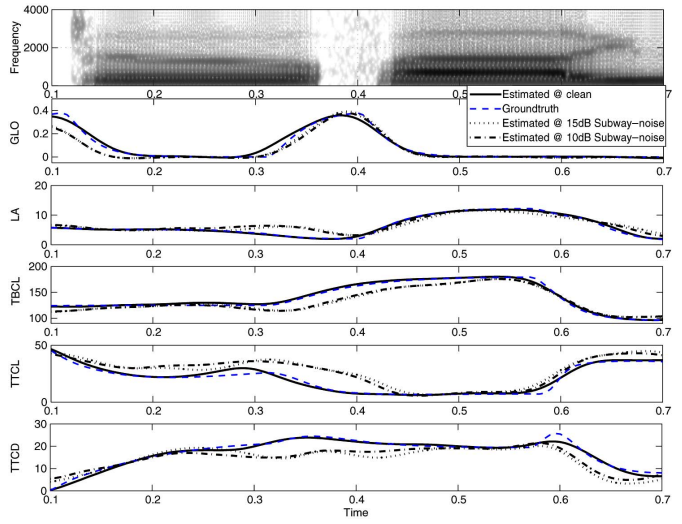


Fig. 8. Spectrogram of synthetic utterance "two five," along with the ground truth and estimated (at clean condition, 15 dB and 10 dB subway noise) TVs for GLO, LA, TBCL, TTCL, and TTCD.

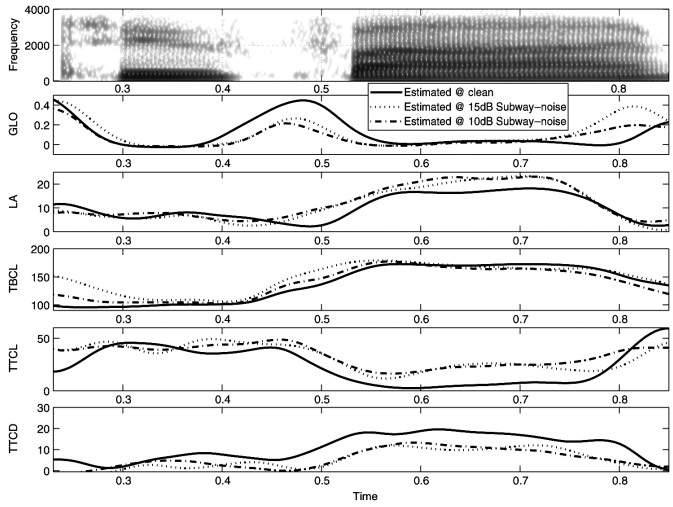


Fig. 9. Spectrogram of natural utterance "two five," along with the estimated (at clean condition, 15 dB and 10 dB subway noise) TVs for GLO, LA, TBCL, TTCL. and TTCD.

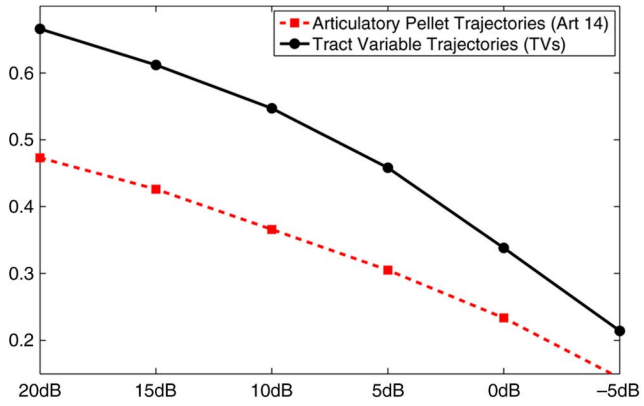


Fig. 10. Average PPMC (relative to clean condition) of the estimated TVs and pellet trajectories (after Kalman smoothing) for Auora-2 (natural speech) at different SNRs for subway noise.

training data (synthetic data of AUR-SYN) and testing data (natural utterances of Aurora-2).

Figs. 8 and 9 show how the estimated TVs from natural speech look compared to those for the synthetic speech. Fig. 8 shows the groundtruth TVs (GLO, LA, TBCL, TTCL, and TTCD) and the corresponding estimated TVs for the synthetic utterance "two five" from AUR-SYN for clean condition, 15 dB and 10 dB SNR subway noise contaminated speech. Fig. 9 shows the same set of TVs estimated from the natural utterance "two five" from Aurora-2 for clean condition, 15 dB and 10 dB SNR. Note that, since we do not know the groundtruth TVs for this natural utterance, it cannot be shown in the plot. Comparing Figs. 8 and 9 we observe that the estimated TVs for both the natural and synthetic speech show much similarity in their dynamics at clean condition, with noise addition the dynamic characteristics of the trajectories starts to deviate away from that at clean condition.

In earlier work [53], we showed that TVs can be estimated relatively more accurately than flesh-point pellet trajectories for clean synthetic speech. To further validate the TV's relative estimation superiority over pellet trajectories for noisy speech, we trained a 3-hidden layer FF-ANN pellet-estimation model using TADA-simulated pellet trajectories for the AUR-SYN data. Seven pellet positions were considered: Upper Lip, Lower Lip, Jaw, and four locations on the Tongue; since each position was defined by its x- and y-coordinates, this gave rise to a 14-dimensional data trajectory which we named Art-14. The pellet trajectory estimation model was deployed on the test set of the Aurora-2 data and the estimated pellet trajectories were smoothed using a Kalman filter. Fig. 10 shows the average relative PPMC across all the components of the Kalman-smoothed TV and pellet trajectory estimates for the subway noise section of Aurora-2.

It can be observed from Fig. 10 that the TV estimates offer a higher average relative PPMC at all noise levels compared to the pellet-trajectory estimates, indicating the relative noise-robustness of the TVs.

C. Noise Robustness in Word Recognition Using Estimated TVs

In this section, we performed ASR experiments using the estimated TVs inputs to examine if they help to improve the ASR noise-robustness. We employed the HTK-based speech recognizer distributed with the Aurora-2 [59], which uses eleven whole word HMMs with three mixture components per state and

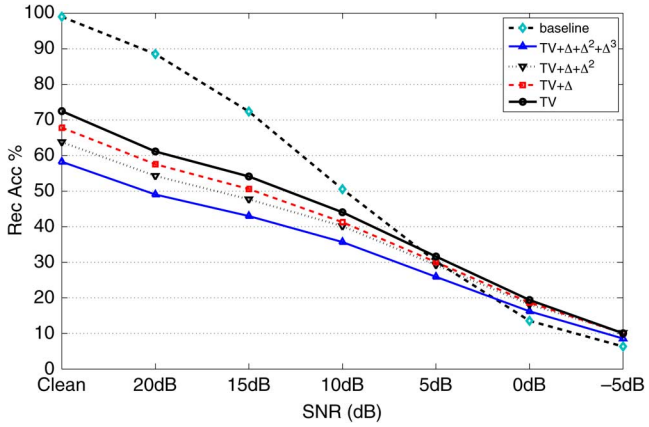


Fig. 11. Average word recognition accuracy (averaged across all the noise types) for the baseline and TVs with different Δs.

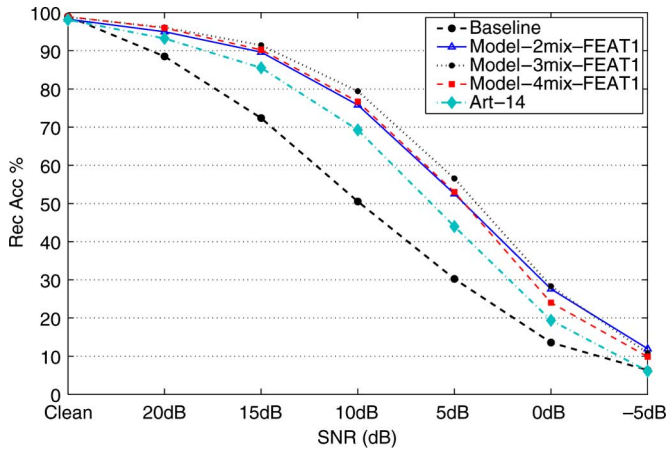


Fig. 12. Average word recognition accuracy (averaged across all the noise types) for the baseline, MFCC+TV using the three different number of Gaussian mixture components per state, and MFCC+Art14 using a 3 Gaussian mixture component per state model.

two pause models for "sil" and "sp" with six mixture components per state. The ASR experiment was based on training on clean condition and testing on multi-SNR noisy data. The following subsections report ASR results obtained from using the estimated TVs in different input conditions.

1) Use of TVs and Their Contextual Information in ASR: We first needed to examine if variants of TVs, or their Δs[4] can help for better ASR performance, and tested four different feature vectors[2] as ASR inputs: 1) TVs; 2) TVs and their velocity coefficients $(\text{TV} + \Delta)$;[3]; 3) TVs and their velocity and acceleration coefficients $(\text{TV} + \Delta + \Delta^2)$; and 4) TVs and their velocity, acceleration, and jerk coefficients $(\text{TV} + \Delta + \Delta^2 + \Delta^3)$. Fig. 11 shows their word recognition accuracies along with a baseline defined by using the MFCC feature vector.[4] The recognition accuracy from using TVs and/or their Δs in the clean condition is much below the baseline recognition rate, which indicates that TVs and their Δs by themselves may not be sufficient for word recognition. However, at 0 and -5 dB, TVs and their Δs offered better accuracy over MFCCs (significance was confirmed at the 1% level, using the significance-testing procedure described in [63]). Our observation for the clean condition is consistent with Frankel *et al.*'s observation [44], [45] that using estimated articulatory information by itself resulted in much lower recognition accuracy as compared to acoustic features.

We also observed that TVs' contextual information (their Δs) in conjunction with TVs did not show better accuracies than TVs alone (at the 5% significance level). This may be because the TV-estimator already uses a large contextualized (context window of 170 ms) acoustic observation (as specified in Section II) as the input; hence, the estimated TVs by themselves should contain sufficient contextual information and further contextualization may be redundant.

2) TVs in Conjunction With the MFCCs: Frankel *et al.* [44], [45] noticed a significant improvement in recognition accuracy when the estimated articulatory data was used in conjunction with the cepstral features, which we also observed in our prior work [57]. We used the MFCCs along with the estimated TVs for the ASR experiments. Here we considered three different models by varying the number of word (digit) mixture components per state from 2 to 4, identified as "Model-2mix," "Model-3mix," and "Model-4mix," where "Model-3mix" is the baseline model distributed with Aurora-2. Fig. 12 compares the recognition accuracy[5] of MFCC+TV from the different word models to the baseline accuracy using MFCC only. Adding TVs to MFCCs resulted in significant improvement in the word recognition accuracy compared to the baseline system using MFCCs only. The improvement is observed at all noise levels for all noise types. Note the baseline here is the result from the Model-3mix,[6] which showed the best performance among the models using MFCC+TV as shown in Fig. 12. Also in Fig. 12 we show the performance of the 14 flesh-point pellet trajectories (Art-14) when used in addition to the MFCCs, where the back-end uses 3-mixture components per state. Fig. 12 clearly shows the superiority of TVs over Art-14 for improving the noise-robustness of a word-recognizer. Although Art-14 is found to improve the noise robustness over the MFCC baseline, it fails to perform as well as the TVs.

3) Speech Enhancement: This section examines how speech enhancement will interact with the use of TV estimates and MFCCs. We used the preprocessor based MPO-APP[7] speech-enhancement architecture described in [64] to enhance the noisy speech signal from Aurora-2. Four different combinations of MFCC and TV estimates were obtained depending upon whether or not their input speech was enhanced.[8] Fig. 13 presents the average word recognition accuracies obtained from these four different feature sets. Similar to the results in the last section, articulatory information (in the form of TVs) can increase the noise robustness of a word recognition system when used with the baseline-MFCC features.

Indeed, TV estimates from enhanced speech exhibited poorer performance than TVs from noisy speech. This can be due to

[2]The dimension of TV and each of its Δs is 8.

[3]Δ, Δ^2, and Δ^3 represent the first, second, and third derivatives, respectively.

[4]The dimension of MFCC feature vector is 39: 12 MFCC + energy, 13 Δ and 13 Δ^2.

[5]The recognition accuracy here is averaged across all the noise types.

[6]We used this model for the rest of this paper.

[7]MPO: modified phase opponency and APP: aperiodic-periodic and pitch detector. The MPO-APP [3] speech enhancement architecture was motivated by perceptual experiments.

[8]The $\text{MFCC}_{\text{MPO}-\text{APP}}$ and the $\text{TV}_{\text{MPO}-\text{APP}}$ are the MFCCs and TVs that were obtained after performing MPO-APP enhancement of the speech signal.

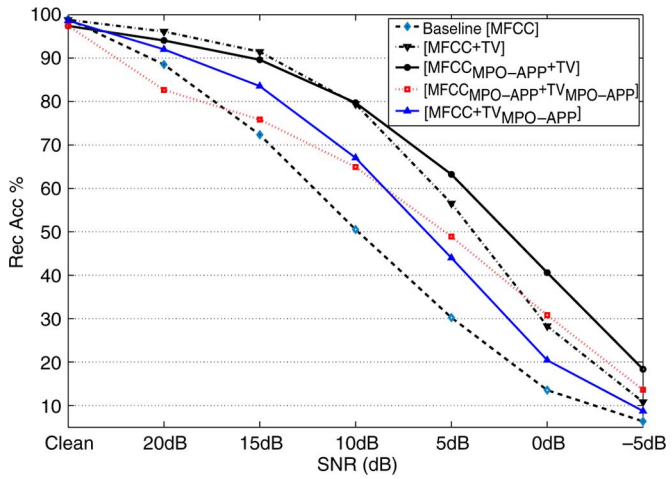


Fig. 13. Average word recognition accuracy (averaged across all the noise types) for the four different combinations of MFCCs and TVs.

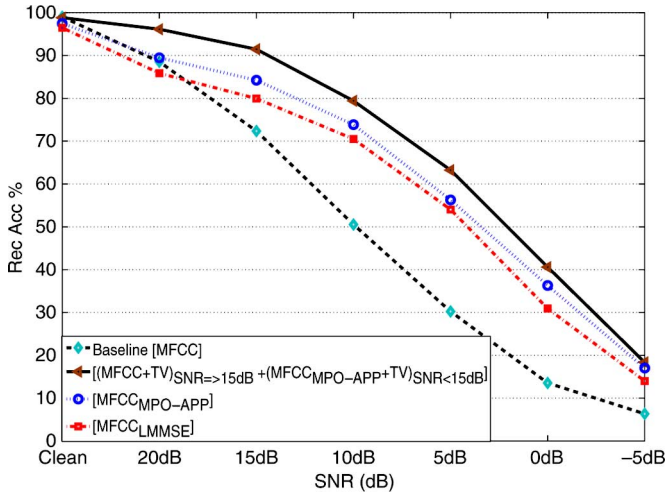


Fig. 14. Average word recognition accuracy (averaged across all the noise types) for the (a) baseline (MFCC), (b) system using $\{[\text{MFCC} + \text{TV}]_{\text{SNR}\geq 15\text{ dB}} + [\text{MFCC}_{\text{MPO}-\text{APP}} + \text{TV}]_{\text{SNR}<15\text{ dB}}\}$, system using the (c) preprocessor-based MPO-APP, and (d) LMMSE-based speech enhancement prior to computing the MFCC features (MFCC).

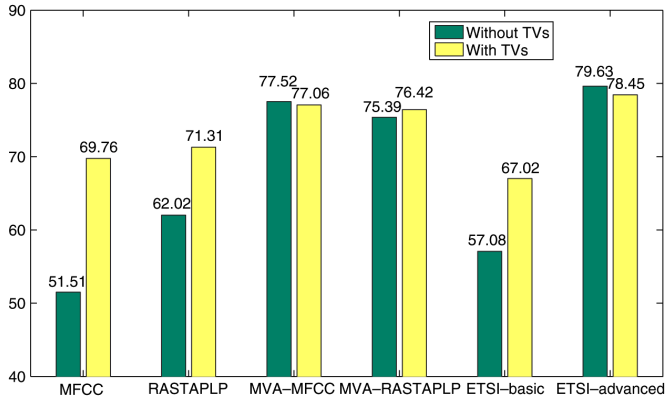


Fig. 15. Overall word recognition accuracy (averaged across all noise types and levels) for the different feature sets and frontends with and without TVs.

the fact that the MPO-APP-based speech enhancer [3] models speech as a constellation of narrow-band regions, retaining only the harmonic regions while attenuating the rest. The voiceless consonants (which are typically wideband regions) are most likely to be attenuated as a result of MPO-APP enhancement of speech. Given the attenuation of unvoiced regions in the enhanced speech, the TV-estimator may have difficulty in detecting the TVs properly at unvoiced consonant regions.

In Fig. 13, the best accuracy is found in MFCC+TV from clean condition to 15 dB, and $\text{MFCC}_{\text{MPO}-\text{APP}} + \text{TV}$ from 10 to -5 dB. Such a system can be realized by using the preprocessor-based MPO-APP architecture prior to generating the baseline MFCC features only for SNRs lower than 15 dB, which is named as $[(\text{MFCC} + \text{TV})_{\text{SNR}\geq 15\text{ dB}} + (\text{MFCC}_{\text{MPO}-\text{APP}} + \text{TV})_{\text{SNR}<15\text{ dB}}]$ feature set. Note the preprocessor-based MPO-APP [64] has an inbuilt SNR-estimator in its preprocessing module which has been used to perform speech enhancement only if the detected SNR is $<$15 dB. Fig. 14 compares $[(\text{MFCC} + \text{TV})_{\text{SNR}\geq 15\text{ dB}} + (\text{MFCC}_{\text{MPO}-\text{APP}} + \text{TV})_{\text{SNR}<15\text{ dB}}]$ with recognition rates from other referential methods that does not use TVs: $\text{MFCC}_{\text{MPO}-\text{APP}}$ (MFCCs after MPO-APP enhancement of speech [64]) and $\text{MFCC}_{\text{LMMSE}}$ (MFCCs after the log-spectral amplitude minimum mean square estimator (LMMSE)-based speech enhancer [65]). The use of articulatory information (in the form of the eight TVs) in addition to MFCCs resulted in superior performance as compared to using speech enhancement alone ($\text{MFCC}_{\text{MPO}-\text{APP}}$ and $\text{MFCC}_{\text{LMMSE}}$). This shows the strong potential of the articulatory features for improving ASR noise robustness.

4) Different Frontend Processing and Feature Sets: In Section IV-C3, we observed that TVs in word recognition task help to increase the accuracy when they are used in conjunction with the MFCCs. Word recognition accuracies were further improved at low SNRs when MPO-APP speech enhancement is performed before obtaining the MFCCs. This section examines whether the advantage of using TVs holds for other feature sets (RASTAPLP) and front-end processing (MVA and ESTI).

RelAtive SpecTrA (RASTA) [5] is a technique that performs low-pass filtering in the log-spectral domain to remove the slowly varying environmental variations and fast varying artifacts. We employed RASTAPLP as acoustic feature set instead of MFCC for the Aurora-2 word recognition task. Similar to Section IV-C2, we observed that use of TVs in addition to RASTAPLP exhibited a better accuracy than either TVs or RASTAPLP alone.

Mean subtraction, Variance normalization and ARMA filtering (MVA) postprocessing has been proposed by Chen *et al.* [6], which have shown significant error rate reduction for the Aurora-2 noisy word recognition task, when directly applied in the feature domain. We applied MVA to both MFCC and RASTAPLP and used them along with TVs as inputs for the word recognition task.

The ETSI front-ends have been proposed for the distributed speech recognition (DSR). We have considered two versions of the ETSI front-end, the ETSI basic (ETSI ES 201 108 Ver. 1.1.3, 2003) [10] and the ETSI advanced (ETSI ES 202 050 Ver. 1.1.5, 2007) [11]. Both the basic and the advanced front-ends use MFCCs, where the speech is sampled at 8 kHz, analyzed in blocks of 200 samples with an overlap of 60% and uses a Hamming window for computing the fast Fourier transform (FFT).

Fig. 15 compares the overall recognition accuracies from six different front-ends: 1) MFCC; 2) RASTAPLP; 3) MFCC

TABLE III
AVERAGED RECOGNITION ACCURACIES (0 TO 20 dB) OBTAINED FROM USING TVS AND SOME STATE-OF-THE-ART WORD RECOGNITION SYSTEMS THAT HAS BEEN REPORTED SO FAR

	Rec. Acc (%)
MFCC	51.05
MFCC+TV	74.23
RASTAPLP+TV	75.95
MVA-RASTAPLP+TV	82.05
Soft Margin Estimation (SME) [66]	67.44
SME + Mean and Variance Normalization [66]	86.01
ETSI-advanced [11]	86.13
Feature Compensation (FC) [67]	83.50
MLLR [67]	76.76

through MVA (MVA-MFCC); 4) RASTAPLP through MVA (MVA-RASTAPLP); 5) ETSI-basic; and 6) ETSI-advanced. All these conditions are further separated into cases with and without TVs. The positive effect of using TVs was consistently observed in most of the noisy scenarios of MFCC, RASTAPLP, MVA-RASTAPLP, and ETSI-basic but not in MVA-MFCC and ESTI-advanced. Note, that the TV-estimator being trained with synthetic speech does not generate highly accurate TV estimates when deployed on natural speech. The ETSI-advanced and the MVA-MFCC front-ends show substantial noise robustness by themselves; hence, the inaccuracy in the TV estimates factors in more and hence fails to show any further improvement in their performance.

Table III compares the recognition accuracies from our experiments to some of the state-of-the-art results that have been reported on Aurora-2. The entries in bold are the accuracies obtained from using the estimated TVs. The accuracy from using MVA-RASTAPLP+TV is close to those of the state-of-the art recognition.

V. DISCUSSION AND CONCLUSION

This study aimed to investigate the possibility of using TVs as noise robust ASR input. First, we evaluated how accurately articulatory information (in the form of TVs) can be estimated from noisy speech at different SNRs using a feedforward neural network. The groundtruth TVs at present are only available for synthetic dataset (we are currently working to generate TVs for the X-ray microbeam data [60]); hence, the TV-estimator was trained with the synthetic data only. Using that network we also evaluated the feasibility to estimate TVs for a natural speech dataset (Aurora-2), consisting of digits. We showed that the TV-estimator can perform reasonably well for natural speech. Second, we showed that estimated TVs for natural speech when used in conjunction with the baseline MFCC and RASTAPLP features can improve word recognition rates appreciably for noisy cases. We also observed that a speech enhancement algorithm when used prior to generating the MFCCs or RASTAPLP at low SNRs ($<$15 dB) can help to improve the noise robustness even further. These results suggest that TVs, if estimated properly, can contribute in improving the noise robustness of ASR systems. The improvement in the recognition accuracies by incorporating articulatory information in the form of TVs may indicate that the acoustic features (MFCC and RASTAPLP) and TVs are providing partially complementary information about speech; hence, neither of them alone provided better accuracy than when both used together. This observation is in line with that made by Kirchhoff [54], [55]. Note that the recognition accuracy improvement obtained as a result of using TVs in addition to the acoustic features (MFCC and RASTAPLP) is confirmed at a significance level of 0.01%.

When TVs are used in conjunction with noise robust front-end processing such as MVA-RASTAPLP or the ETSI-basic front-end, improvement in word recognition accuracy has been observed for noisy cases. No improvement is however witnessed for MVA-MFCC and ETSI-advanced, which may be due to the inaccuracy of the TV-estimator. In this paper, we have used an FF-ANN-based inverse model to estimate TVs from the speech signal, where the model was trained with a significantly small number of data (960 utterances) than that available in Aurora-2 training database (8440 training utterances). Also there exists a strong acoustic mismatch between the training (clean synthetic speech data) and testing (clean and noisy natural speech data) utterances for the TV-estimator. Despite these differences, we were able to observe improvement in word recognition accuracies in the noisy cases of the Aurora-2 dataset for acoustic features: MFCC, RASTAPLP, and noise-robust front-ends: MVA-RASTAPLP and ETSI-basic. We are currently in the process of generating a natural speech dataset that will contain TV groundtruth information which, when realized, would help to construct a more accurate and robust TV-estimator for natural speech. Doing so may, in turn, help in obtaining results comparable to some of the state-of-the-art results reported in Table III. Future research aims to realize such a TV-estimator that will be trained not only with natural speech but also with a much larger training corpus. Finally, the TV-estimator uses contextualized MFCCs as inputs and throughout the course of this paper we have noticed that the raw MFCCs showed least noise robustness. Hence, future research should also address using noise robust features like RASTAPLPs or MVA-RASTAPLPs or MVA-MFCCs that may ensure better estimation of TVs even for the noisy speech samples resulting in improvement of the recognition accuracies.

ACKNOWLEDGMENT

The authors would like to thank Dr. K. Livescu, Dr. M. Hasegawa-Johnson, and Dr. T. Pruthi for their helpful tips and suggestions, Dr. C.-P. Chen and Dr. Jeff Bilmes for their help with the MVA algorithm, the Associate Editor Dr. Nestor Becerra Yoma, and the anonymous Reviewers of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING for their valuable comments and suggestions.

REFERENCES

[1] P. Lockwood and J. Boudy, "Experiments with a non linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," in *Proc. Eurospeech*, 1991, pp. 79–82.
[2] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2130–2140, Sep. 2007.
[3] O. Deshmukh, C. Espy-Wilson, and L. H. Carney, "Speech enhancement using the modified phase opponency model," *J. Acoust. Soc. Amer.*, vol. 121, no. 6, pp. 3886–3898, 2007.
[4] R. Flynn and E. Jones, "Combined speech enhancement and auditory modelling for robust distributed speech recognition," *Speech Commun.*, vol. 50, pp. 797–809, 2008.
[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[6] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[7] T. M. Sullivan, "Multi-Microphone correlation-based processing for robust automatic speech recognition," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1996.

[8] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Proc. ICASSP*, 2004, pp. 549–552.

[9] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 451–464, Sep. 1997.

[10] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 201 108 Ver. 1.1.3, 2003.

[11] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Frontend Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 202 050 Ver. 1.1.5, 2007.

[12] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM spin evaluation system," in *Proc. ICASSP*, 2002, vol. 1, pp. I-53–I-56.

[13] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[14] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput., Speech Lang.*, vol. 9, pp. 171–185, 1995.

[15] Z. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," *Speech Commun.*, vol. 42, no. 1, pp. 43–58, Jan. 2004.

[16] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[17] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, pp. 373–376.

[18] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. Eurospeech*, 1999, vol. 6, pp. 2833–2836.

[19] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE Autom. Speech Recognition Understanding Workshop*, CO, 1999, vol. 1, pp. 79–83.

[20] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 1086–1101, Feb. 2002.

[21] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. ICASSP*, 2001, vol. 1, pp. 577–580.

[22] K. N. Stevens, "Toward a model for speech recognition," *J. Acoust. Soc. Amer.*, vol. 32, pp. 47–55, 1960.

[23] R. Cole, R. M. Stern, and M. J. Lasry, , J. S. Perkell and D. Klatt, Eds., "Performing fine phonetic distinctions: Templates versus features," in *Invariance and Variability of Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1986, ch. 15, pp. 325–345.

[24] B. Lochschmidt, "Acoustic-phonetic analysis based on an articulatory model," in *Automatic Speech Analysis and Recognition*, J. P. Hayton, Ed. Dordrecht, The Netherlands: D. Reidel, 1982, pp. 139–152.

[25] R. D. Mori, P. Laface, and E. Piccolo, "Automatic detection and description of syllabic features in continuous speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 365–379, Oct. 1976.

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

[27] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Process.*, vol. 27, no. 1, pp. 65–78, 1992.

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

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

[30] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 333–353, Oct. 2000.

[31] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper & Row, 1968.

[32] J. Harris, *English Sound Structure*. Oxford, U.K.: Wiley-Blackwell, 1994.

[33] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 723–742, 2007.

[34] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, Univ. of Edinburgh, Edinburgh, U.K., 2001.

[35] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 2000.

[36] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1535–1555, 1978.

[37] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. Sondhi, "On the use of neural networks in articulatory speech synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1109–1121, 1993.

[38] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Amer.*, vol. 92, no. 2, pp. 688–700, 1992.

[39] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Amer.*, vol. 64, no. 4, pp. 1027–1035, 1978.

[40] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Amer.*, vol. 100, no. 3, pp. 1819–1834, 1996.

[41] T. Okadome, S. Suzuki, and M. Honda, "Recovery of articulatory movements from acoustics with phonemic information," in *Proc. 5th Seminar Speech Prod.*, Bavaria, Germany, 2000, pp. 229–232.

[42] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proc. Interspeech*, 2007, pp. 74–77.

[43] D. Neiberg, G. Ananthakrishnan, and O. Engwall, "The acoustic to articulation mapping: Non-Linear or non-unique?," in *Proc. Interspeech*, 2008, pp. 1485–1488.

[44] J. Frankel and S. King, "ASR—Articulatory speech recognition," in *Proc. Eurospeech*, 2001, pp. 599–602.

[45] 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," in *Proc. ICSLP*, 2000, vol. 4, pp. 254–257.

[46] R. Daniloff and R. Hammarberg, "On defining coarticulation," *J. Phon.*, vol. 1, pp. 239–248, 1973.

[47] C. P. Browman and L. Goldstein, "Towards an articulatory phonology," *Phonol. Yearbook*, vol. 85, pp. 219–252, 1986.

[48] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Gynecol. Obstet. Invest.*, vol. 49, pp. 155–180, 1992.

[49] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 332–382, 1989.

[50] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in Matlab," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, p. 2430, 2004.

[51] J. Westbury, *X-ray Microbeam Speech Production Database User's Handbook*. Madison, WI: Univ. of Wisconsin, 1994.

[52] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Commun.*, vol. 14, no. 1, pp. 19–48, Feb. 1994.

[53] 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. Sel. Topics Signal Process.*, vol. 4, no. 6, pp. 1027–1045, Dec. 2010.

[54] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Univ. of Bielefeld, Bielefeld, Germany, 1999.

[55] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Commun.*, vol. 37, no. 3-4, pp. 303–319, Jul. 2002.

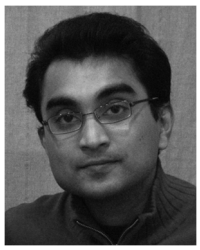
[56] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Speech Commun.*, vol. 41, no. 2-3, pp. 511–529, Oct. 2003.

[57] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Noise robustness of tract variables and their application to speech recognition," in *Proc. Interspeech*, 2009, pp. 2759–2762.

[58] H. M. Hanson and K. N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," *J. Acoust. Soc. Amer.*, vol. 112, no. 3, pp. 1158–1182, 2002.

[59] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Autom. Speech Recognition: Challenges For New Millenium, ASR-2000*, Paris, France, 2000, pp. 181–188.

[60] 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," in *Proc. Interspeech*, 2010, pp. 30–33.
[61] V. Mitra, I. Özbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustics to vocal tract time functions," in *Proc. ICASSP*, 2009, pp. 4497–4500.
[62] J. Hogden, D. Nix, and P. Valdez, An articulatorily constrained, maximum likelihood approach to speech recognition Los Alamos National Laboratory, Los Alamos, NM, Tech. Rep., LA-UR-96-3945, 1998.
[63] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, vol. 1, pp. 532–535.
[64] V. Mitra, B. J. Borgstrom, C. Espy-Wilson, and A. Alwan, "A noise-type and level-dependent MPO-based speech enhancement architecture with variable frame analysis for noise-robust speech recognition," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 2751–2754.
[65] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[66] 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.*, vol. 18, no. 6, pp. 1158–1169, Aug. 2009.
[67] X. Cui and Y. Gong, "A study of variable-parameter gaussian mixture hidden Markov modeling for noisy speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1366–1376, May 2007.



Vikramjit Mitra (S'05) received the B.E. degree from Jadavpur University, West Bengal, India, in 2000, the M.S. degree in electrical engineering with specialization in signal processing and communication from University of Denver, Denver, CO, in 2004, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park.

He is currently working as a Researcher for the Speech Communication Laboratory, Institute of Systems Research (ISR), University of Maryland. His research interests are in robust speech recognition, estimation of articulatory information from speech, language recognition, information retrieval, and machine learning.

Hosung Nam (M'09) received the M.S. and Ph.D. degrees from the Department of Linguistics, Yale University, New Haven, CT, in 2007.

He is a Linguist who is an expert in the field of articulatory phonology, a sub-discipline of linguistics that integrates the abstract symbolic aspects of phonology with its phonetic implementation in the dynamics of speech motor control. His research emphasis is on the link between speech perception and production, speech error, automatic speech recognition, sign language, phonological development, and their computational modeling. He has been a Research Scientist at Haskins Laboratories, New Haven, CT, since 2007.

Carol Y. Espy-Wilson (SM'08) received the B.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1979 and the M.S., E.E., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 1981, 1984, and 1987, respectively.

She is a Professor in the Department of Electrical and Computer Engineering and the Institute for Systems Research at the University of Maryland, College Park. She is also affiliated with the Center for Comparative and Evolutionary Biology of Hearing. Her research focuses on understanding the relationship between acoustics and articulation and it involves modeling speech production, studying speech perception, developing signal processing techniques that capture relevant information in speech, and using the knowledge gained to develop speech technologies. Current projects include single-channel speech enhancement and speaker separation, speech recognition, speaker recognition, language identification, and forensics.

Prof. Espy-Wilson is currently past Chair of the Speech Technical Committee of the Acoustical Society of America (ASA) and she is a member of the IEEE Speech and Language Technical Committee. She is an Associate Editor of the *Journal of the Acoustical Society of America* and she has served as a member of the Language and Communication study section of the National Institutes of Health and as an Associate Editor of ASA's magazine, *Acoustics Today*. She is currently serving on the National Advisory Board on Medical Rehabilitation Research which advises the National Center for Medical Rehabilitation Research, part of the Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health. She has received several awards including a Clare Boothe Luce Professorship, an NIH Independent Scientist Award, a Honda Initiation Award, and a Harvard University Radcliffe Fellowship. She is a Fellow of the Acoustical Society of America.

Elliot Saltzman received the doctorate from the Institute of Child Development, University of Minnesota, and his original training was in developmental psychology.

He is an Associate Professor in the Department of Physical Therapy and Athletic Training at Boston University's Sargent College, Boston, MA, and a Senior Research Scientist at Haskins Laboratories, New Haven, CT. During graduate studies, his focus was on the developmental relationship between early sensorimotor behaviors and emergent cognitive and linguistic skills. While at the University of Minnesota, he realized that in order to rigorously address the sensorimotor origins of cognition and language, one must first understand the nature of sensorimotor coordination and control. This realization led him to become immersed in an interdisciplinary research program that included human motor physiology, human perception and performance, robotics, engineering, and computer science. Out of this work came a deep appreciation for the multi-leveled nature of skilled behavior and the manner in which the dynamics of tasks could be used to illuminate the flexibility and adaptability that are the hallmarks of even the simplest motor skills. In particular, he developed a task dynamic model of the coordinative structures underlying skilled actions in which movement patterns are generated with reference to dynamical systems specified in abstract, task-specific coordinate spaces. As a model of speech production, task-dynamics provides a conceptual and computational rapprochement between the abstract, symbolic nature of linguistic units and their concrete implementation as sensorimotor units of articulatory control and coordination. Currently, he is extending the task-dynamic model to encompass the temporal patterning of action units in both speech and manual behaviors, and how these patterns are modulated when the units participate in higher order sequential and/or hierarchical patterns.

Louis Goldstein received the Ph.D. degree in linguistics from the University of California, Los Angeles (UCLA).

He has been a Senior Scientist at Haskins Laboratories, New Haven, CT, since 1980, and lectured in linguistics at Yale University, New Haven, from 1980 to 2007. Since 2007, he has been Professor of linguistics at the University of Southern California. His main work has been the development of articulatory phonology, a framework for modeling the phonetic and phonological structure of language which he undertook in collaboration with C. Browman. In this approach, the primitive, combinatorial units of phonology are gestures, constriction actions of the vocal tract articulators. Utterances are modeled as ensembles of these gestures, organized in time according to principles of inter-gestural coordination. His current work focuses on the dynamics of planning the relative timing of gestures in these ensembles and the relation of that dynamics to the syllable-structure organization in phonology. Specific research projects include dynamical modeling of speech variability and error, use of prosody and gestural phonology in speech recognition, cross-linguistic differences in the coordination of speech gestures, and the gestural analysis of signs in American Sign Language.

Dr. Goldstein is a Fellow of the Acoustical Society of America.