Retrieving Tract Variables From Acoustics: A Comparison of Different Machine Learning Strategies

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

***Abstract*—Many different studies have claimed that articulatory information can be used to improve the performance of automatic speech recognition systems. Unfortunately, such articulatory information is not readily available in typical speaker-listener situations. Consequently, such information has to be estimated from the acoustic signal in a process which is usually termed "speech-inversion." This study aims to propose and compare various machine learning strategies for speech inversion: Trajectory mixture density networks (TMDNs), feedforward artificial neural networks (FF-ANN), support vector regression (SVR), autoregressive artificial neural network (AR-ANN), and distal supervised learning (DSL). Further, using a database generated by the Haskins Laboratories speech production model, we test the claim that information regarding constrictions produced by the distinct organs of the vocal tract (vocal tract variables) is superior to flesh-point information (articulatory pellet trajectories) for the inversion process.**

***Index Terms*—Articulatory phonology, articulatory speech recognition (ASR), artificial neural networks (ANNs), coarticulation, distal supervised learning, mixture density networks, speech inversion, task dynamic and applications model, vocal-tract variables.**

I. Introduction

PERFORMANCE of the current state-of-the-art automatic speech recognition (ASR) systems suffer in casual or spontaneous speech. This problem stems from the fact that spontaneous speech typically has an abundance of variability, a major part of which arises from contextual variation commonly known as coarticulation. Phone-based ASR systems represent speech as a sequence of non-overlapping phone units [89] and contextual variations induced by coarticulation [86] are typically encoded by unit combinations (e.g., di- or tri-phone). These di- or tri-phone [54]-based models often suffer from data sparsity. It has been observed [70], [71] that coarticulation affects the basic contrasting distinctive features between phones. Hence, an ASR system using phone-based acoustic models may be expected to perform poorly when faced with coarticulatory effects. Moreover, di- or tri-phone-based models limit the contextual influence to only the immediately close neighbors and as a result, they are limited in the degree of coarticulation that can be captured [58]. For example, in casual productions of the word "strewn," anticipatory rounding throughout the /str/ sequence can occur due to the vowel /u/. That is, coarticulatory effects can reach beyond adjacent phonemes and, hence, such effects are not covered by traditional tri-phone inventories.

Coarticulation has been described in a variety of ways including the spreading of features from one segment to another (also called assimilation). However, coarticulation can be better understood as a property that occurs from a sequence of overlapping discrete actions in the human vocal tract [38]. Articulatory phonology [5], [6], [98] treats the variability in speech (specifically coarticulation) from the speech production point of view, using speech gestures [73] as primitive speech production units. It has been shown [4]–[10] that the gesture-based speech production model effectively accounts for speech variations such as coarticulation effects by allowing gestural overlap[1] in time and gestural reduction in space.

Speech gestures are constriction actions produced by distinct organs (lips, tongue tip, tongue body, velum, and glottis) along the vocal tract [shown in Fig. 1(a)]. Speech gestures can be defined in terms of eight vocal tract constriction variables also known as tract variables (TVs), as shown in Table I. TVs describe geometric features of the shape of the vocal tract tube in terms of constriction degree and location. An active gesture is specified by activation onset and offset times and parameter values for a set of critically damped, second-order differential equations [11], shown in (1), where M, B, and K are mass, damping coefficient, and stiffness parameters of each TV (represented by z) and z_0 is the target position of the gesture:

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

Each TV involves its own set of associated articulators. Given a time varying pattern (or constellation) of gestural activity, the trajectories of the TVs are derived using the TAsk-Dynamic and

Manuscript received December 15, 2009; accepted February 18, 2010. Date of publication September 13, 2010; date of current version November 17, 2010. This work was supported by National Science Foundation (NSF) under Grants IIS-0703859, IIS-0703048, IIS-0703782, and NIH-NIDCD grant DC-02717. 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 Dr. Li Deng.

V. Mitra and C. Y. Espy-Wilson are with the Department of Electrical and Computer Enginering, Institute of 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 Laboratory, New Haven, CT 06511 USA (e-mail: nam@haskins.yale.edu).

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

L. Goldstein with the Department of Linguistics, University of Southern California, Los Angeles, CA 90089-1693 USA, and also with 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/JSTSP.2010.2076013

[1]The span of such overlap can be segmentally extensive [37], [86], [94] but may not be more than 250 ms [36]. A consonantal duration can often be less than 100 ms, which suggests that in consonantal context, coarticulatory effects can theoretically spill-over to more than a tri-phone context.

1932-4553/$26.00 © 2010 IEEE

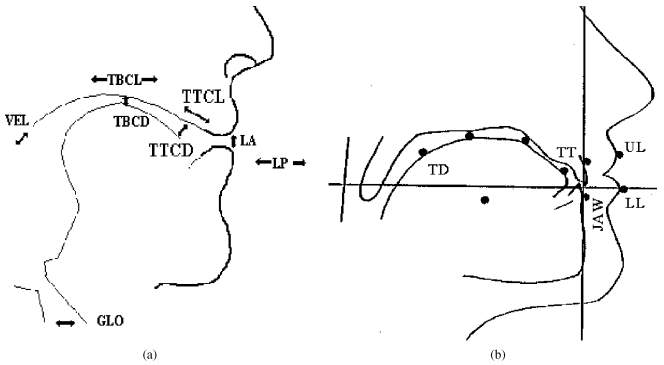


Fig. 1. (a) Tract variables (TVs) from different constriction locations. (b) Pellet placement locations according to [115].

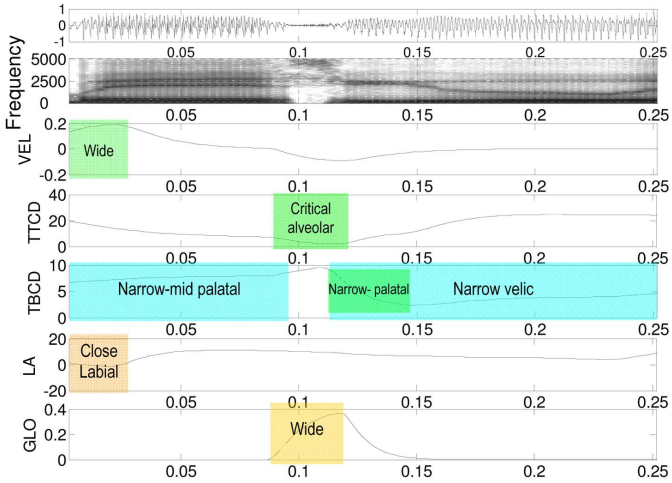


Fig. 2. Gestural activations for the utterance "miss you." Active gesture regions are marked by rectangular solid (colored) blocks. Smooth curves represent the corresponding tract variable (TV) trajectories.

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

Constriction organ	Vocal tract variables (TVs)	*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

Applications (TADA) model [84], which is a computational implementation of articulatory phonology. Fig. 2 shows the gestural pattern of the utterance "miss you," the respective gestural scores and corresponding TVs as computed by TADA.

Fig. 3 show the waveforms (or portions thereof) of two different utterances of the same word pair "perfect memory" spoken by the same person (adapted from [9]). In Fig. 3(a), the words "perfect" and "memory" are uttered with a slight

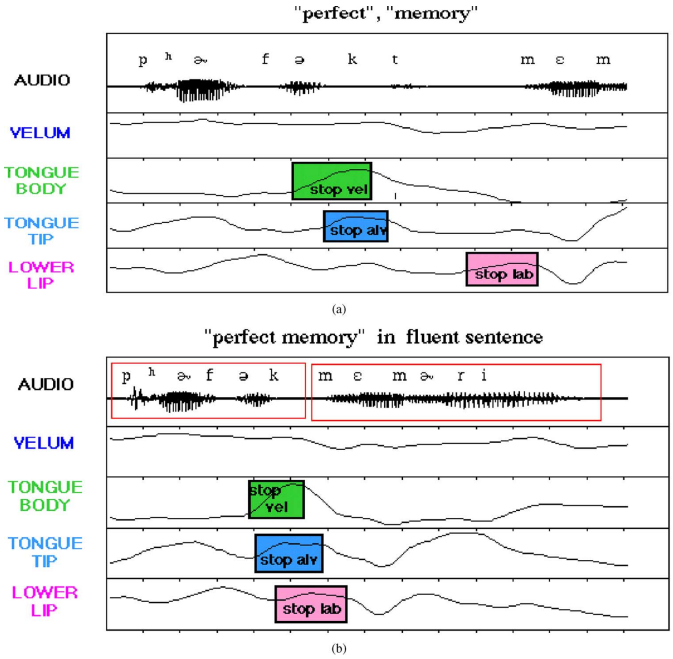


Fig. 3. Example of "perfect memory" adapted from [43], showing the acoustic signal and the recorded articulatory data. (a) Shows the case where "perfect" and "memory" are uttered as two different words (note, the /t/ burst is clearly visible. (b) Shows the utterance of "perfect memory" in a fluent sentence where the /t/ burst is reduced in the acoustic waveform.

pause between them, i.e., as isolated words. In Fig. 3(b), the words "perfect" and "memory" are uttered more fluently with no pause between the words. Comparing the waveforms at the end of the word "perfect" shows that the /t/ burst of the more carefully articulated utterance in part (a) is absent from the more casually spoken utterance in part (b). This apparent "deletion" of the phone /t/ is due to cross word-boundary coarticulation in the more casual utterance, that is, the speaker starts to articulate the /m/ in the word "memory" before the speaker has finished articulating the /t/ in the word "perfect." This coarticulation is evident from the articulatory information displayed for the tongue body, tongue tip, and lower lip. The curves show how the vertical displacement for these articulators (which can be understood as the reverse of the constriction degrees of the relevant gestures) changes as a function of time. While the vertical displacements for the different articulators (tongue body, tongue tip and lower lip) are similar for these two utterances during the /k/ and /t/ at the end of "perfect" and during the /m/ at the beginning of "memory,' the timing is substantially different. For the more fluently spoken utterance, the closure gesture for the /m/ (labeled as "stop lab") overlaps with the tongue tip constriction gesture for the /t/ (labeled as "stop alv").

However, this overlap does not occur for the utterance in part (a). What is most important to note is that, although the acoustic waveform for the more fluent utterance does not show a /t/ burst because of the overlapping gesture for the /m/, the closure gesture for the /t/ is still made by the speaker. Thus, the complex variability (and sometimes relatively discrete changes) that can occur in the acoustic signal is reduced to simple changes in relative timing at the gestural representation level. For this reason, we hypothesize that articulatory gestures will be able

to better capture and model coarticulation than phone-based sub-word units (di- or tri-phone) for ASR.

Speech variability has been an intrinsic problem with ASR systems and Stevens [103] first pointed out that such problems can be alleviated by incorporating anatomical or neuro-physiological level of speech representation which may help to closely simulate the process of human speech perception in the ASR systems. Since then many researchers have ventured to realize a speech production-based ASR architecture as presented in the following subsections.

A. Feature Based ASR Systems

Early attempts to exploit speech production knowledge in ASR systems were very limited in scope. From the late 1970s to the early 1990s most of the research [18], [44], [67], [83] was focused on trying to decipher appropriate features from the acoustic signal. Phonetic features provide descriptive information to account for phonetic differences between speech sounds [17], [66] and may be based on articulatory movements, acoustic events, or perceptual effects [16]. A given feature may be limited to a particular segment but may also be longer (suprasegmental) or shorter (subsegmental) than a segment span. Features that try to capture articulatory events are commonly known as articulatory features (AF) or articulator-bound features. The articulatory feature (AF) concept in literature parallels the "distinctive features" (DF) concept of phonological theory [15]. Although there exist some strong similarity between the AFs and DFs, there are some subtle differences as well. DFs consist of both articulator-free and articulator-bound features [106] defining phonological feature bundles that specify phonemic contrasts used in a language. On the contrary, AFs define more physiologically motivated features based on speech production; hence, they are fully articulator-bound features. One of the earliest systems trying to incorporate AFs was proposed by Schmidbauer [99], which was used to recognize German speech using 19 AFs that described the manner and place of articulation. The AF vectors were used as input to phonemic hidden Markov models (HMMs) and an improvement of 4% was observed over the baseline for a small database. The AF features were also found to be robust against speaker variability and showed less variance in the recognition accuracy of different phonemic classes as compared to the standard HMM-MFCC baseline. Deng [21] proposed an ASR system inspired by a speech-production model, in which the HMM states generated trended-sequence of observations that were piece-wise smooth and continuous. Deng *et al.* [20], [22], [23], [33] performed an exhaustive study on their AF-based system, for which they used 18 multi-valued features to describe place of articulation, vertical and horizontal tongue body movement, and voice information. In their system, the speech signal was modeled using a rule-based combination of AFs where the features at transitional regions were allowed to assume any intermediate target value between the preceding and succeeding articulatory target values. Each individual AF vector was modeled using HMM states, and the transition and emission of a single ergodic HMM was trained using all possible vectors. They reported an average improvement of 26% over the conventional phone-based HMM architecture for a speaker independent classification tasks. Phone recognition for the TIMIT dataset showed a relative improvement of about 9% over the baseline system. For speaker-independent word recognition using a medium sized corpus, they reported a relative improvement of 2.5% over a single-component Gaussian mixture phone recognizer. A phonetic-feature classification architecture was presented in [119], where 18 features were detected using a time-delay neural network. The outputs were used to obtain phoneme probabilities for ALPH English spelling database. A hybrid artificial neural network (ANN)–HMM architecture was proposed by Elenius *et al.* [31], [32] for phoneme recognition; comparing spectral representations against AF they reported an advantage of the articulatory feature-based classifier for speaker independent phoneme recognition. However, for a speaker-dependent task, they observed that the spectral representation performed better than the articulatory features. King *et al.* [60] used ANNs to recognize and generate articulatory features for the TIMIT database. They explored three different feature systems: binary features proposed by Chomsky *et al.* [15], traditional phonetic features defining manner and place categories, and features proposed by Harris [47]. The recognition rates of the three feature systems were similar. In a different study, Kirchhoff *et al.* [62], [63] used a set of heuristically defined AFs and showed that incorporating articulatory information in an ASR system helps to improve its robustness. ANNs have been extensively used in AF recognition from the speech signal. Wester *et al.* [116] and Chang *et al.* [13] proposed separate place classifiers for each manner class. Omar *et al.* [88] used a maximal mutual information approach to obtain a subset of acoustic features for the purpose of AF recognition. HMMs have also been researched widely for AF recognition. Metze *et al.* [75] proposed context-dependent HMM phone models to generate an initial AF set, which were later replaced by a set of feature detectors that used a likelihood combination at the phone or state level. They showed a word error rate (WER) reduction of 1.8% for the Broadcast news database and 1.6% for the Verbmobil task. Dynamic Bayesian Networks (DBN) have also been explored for the purpose of AF recognition. The major advantage of DBNs is their capability to model explicitly the inter-dependencies between AFs. Also, a single DBN can perform both the task of AF recognition and word recognition. One of the earlier works incorporating DBNs for the task of AF recognition was performed by Frankel *et al.* [41], who showed that modeling inter-feature dependencies improved AF recognition accuracy, raising the overall frame-wise feature classification accuracy from 80.8% to 81.5%. However, tying AF features to phone level information overlooks the temporal asynchrony between the AFs. To address this issue, an embedded training scheme was proposed by Wester *et al.* [117], which was able to learn a set of asynchronous feature changes from data. Cole *et al.* [19] showed that the model proposed in [117] provided a slight increase in accuracy for a subset of the OGI number corpus over a similar model trained on phone-derived labels. Frankel *et al.* [42] proposed a hybrid ANN/DBN architecture, where the Gaussian mixture model (GMM) observations used by the DBNs are replaced by ANN posteriors. This hybrid ANN/DBN architecture combined the discriminative training

power of ANN and the inter-feature dependency modeling capability of DBNs. The feature recognition accuracy reported in their paper for the OGI Number corpus was 87.8%. In a different study, Cetin *et al.* [12] proposed a tandem model of MLP and HMM as an ASR system. The MLPs were used for AF classification and the HMM outputs used a factored observation model. Their proposed tandem model using AFs was found to be as effective as the phone-based model. Also, the factored observation model used in their research was found to outperform the feature concatenation approach, indicating that the acoustic features and tandem features yield better results when considered independently rather than jointly. At the 2006 Johns Hopkins University Workshop, Livescu *et al.* [68] investigated the use of AFs for the observation and pronunciation models for ASR systems. They used the AF classifier outputs in two different ways: 1) as observations in a hybrid HMM/ANN model and 2) as a part of the observation in a tandem approach. In this paper, they used both audio and visual cues for speech recognition and the models were implemented as DBNs. They used Switchboard [61] and the CUAVE audio-visual digits corpus to test their approach. They observed that the best ASR performance came from the tandem approach whereas, although the hybrid models could not offer the best accuracy, they require very little training data. They predicted that the hybrid model-based approaches may hold promises for multilingual systems. Hasegawa-Johnson *et al.* [48] exploited the asynchrony between phonemes and visemes to realize a DBN-based speech recognition system. They noted that the apparent asynchrony between acoustic and visual modalities can be effectively modeled as the asynchrony between articulatory gestures corresponding to the lips, tongue and glottis/velum. Their results show that combining visual cues with acoustic information can help reduce the WER at low SNR and the WER is found to further reduce if the asynchronies amongst gestures are exploited.

B. Direct Articulatory Information Retrieval

Typically hypothesized or abstract articulatory features have been used widely in ASR research aiming to incorporate speech production models. Another distinct line of research deals with using direct articulatory (recorded or estimated) trajectories. In a typical ASR framework, the only known parameter is the acoustic speech signal and recorded articulatory data is not readily available (such data may be available for research purposes, but cannot be assumed to be available for real-world applications); hence, such information needs to be estimated from the acoustic observations. There are few ASR results in the literature using direct articulatory information owing to the difficulty in reliably predicting such articulatory dynamics from the speech signal. An alternative is to use actual articulatory recordings directly in the ASR system, but such a setup is not desirable for real-world applications. In an interesting study by Frankel *et al.* [40], a speech recognition system was developed that uses a combination of acoustic and articulatory features as input. They showed that using articulatory data from direct measurements in conjunction with MFCCs resulted in a significant improvement in performance (4% in [39] and 9% in [40]) over the baseline system. However, the phone classification accuracies from using estimated articulatory data reported in their work did not show any improvement over the baseline ASR system, which indicates that a significant amount of effort still needs to be directed toward efficiently estimating articulatory information from speech. The process of retrieving articulatory information from the speech signal is usually termed "speech-inversion." Speech inversion has been a widely researched topic in the last 35 years. One of the earliest as well as ubiquitously cited works in this area was by Atal *et al.* [1] in which information in the acoustic space was used to predict corresponding vocal tract configuration. Rahim *et al.* [92], [93] used an articulatory synthesis model to generate a database of articulatory-acoustic vector pairs and they trained multi-layered perceptrons (MLPs) to map from acoustic data to the vocal tract area functions. Shirai *et al.* [101] proposed an analysis-by-synthesis approach, which they termed as "Model Matching," where speech was analyzed to generate articulatory information and then the output was processed by a speech synthesizer such that it had minimal distance from the actual speech signal in the spectral domain. Kobayashi *et al.* [64] proposed a feed-forward MLP architecture with two hidden layers that uses the same data as used in [101] to predict the articulatory parameters and showed faster performance and better estimation accuracy. Regression techniques have been explored a number of times for speech inversion. Ladefoged *et al.* [65] used linear regression to estimate the shape of the tongue in the midsagittal plane, using the first three formant frequencies in constant vowel segments. Exploiting neural networks gained popularity after the work of Papcun *et al.* [90], in which MLPs were used to obtain articulatory motions for six English stop consonants. Richmond [95] used mixture density networks (MDNs) to obtain the articulator trajectories as conditional probability densities of the input acoustic parameters. He showed that the articulations with critical constrictions show less variability in the probability density functions than the noncritical articulatory trajectories. He also used ANNs to perform the speech inversion task and showed that the MDNs tackle the non-uniqueness of the speech inversion problem more appropriately than ANNs. Non-uniqueness is a problem in speech inversion because different vocal tract configurations can yield similar acoustic realizations. However, separate studies by Qin *et al.* [91] and Neiberg *et al.* [85] show that the majority of normal speech is produced with a unique vocal tract shape and there are only a few instances of non-uniqueness; suggesting that non-uniqueness may not be so critical an issue. One-to-many mappings (or non-uniqueness) can be of the following types: 1) a given speaker may be able to produce multiple articulatory configurations for a given phone (e.g., bunched versus retroflex for /r/ [34], [35]); 2) a given acoustic observation could potentially be generated from many different possible sets of vocal tract area functions. However, the human vocal tract is highly constrained and, as a result such type-2 non-uniqueness is well suppressed [85], [91], a result that is supported by our analysis as well. The data used in this paper may contain type-2 non-uniqueness; we do not aim to analyze type-1 non-uniqueness here.

In a different study of speech inversion, Hogden *et al.* [51] used vector quantization to build a codebook of articulatory-

acoustic parameter pairs. They built a lookup table of articulatory configurations and used the lookup table along with the codebook to estimate articulator positions given acoustic information. They reported an overall average root mean square error (RMSE) of approximately 2 mm. A similar codebook approach was pursued by Okadome *et al.* [87] who used data recorded from three Japanese male speakers which was considerably larger than the dataset used in [51]. They also augmented the codebook search process by making use of phonemic information of an utterance. The average RMSE reported by their algorithm was around 1.6 mm when they used phonemic information to perform the search process.

Efforts have also been made in implementing dynamic models for performing speech inversion. Dusan [30] used extended Kalman filtering (EKF) to perform speech inversion by imposing high-level phonological constraints on the articulatory estimation process. In his approach, the speech signal is segmented into phonological units and constructed trajectories based on the recognized phonological units; the final estimate was performed by using Kalman smoothing. Dynamic model-based approaches are typically found to work well for vowels but often fail for consonants [53].

C. Vocal Tract Resonances for ASR

Apart from articulatory variables, other sources of information such as vocal tract shapes and vocal tract resonances (VTR) can be used to capture the dynamics of natural speech. Deng *et al.* [24] and Deng [25] proposed a statistical paradigm for speech recognition where phonetic and phonological models are integrated with a stochastic model of speech incorporating the knowledge of speech production. In such an architecture, the continuous and dynamic phonetic information of speech production (in the form of vocal tract constrictions and VTRs) is interfaced with a discrete feature-based phonological process. It is claimed [25] that such integration helps to globally optimize the model parameters that accurately characterize the symbolic, dynamic, and static components in speech production and also contribute in separating out the sources of speech variability at the acoustic level. Their work shows [24] that synergizing speech production models with a probabilistic analysis-by-synthesis strategy may result in automatic speech recognition performance comparable to the human performance. Deng *et al.* [26], [69] proposed a statistical hidden dynamic model to account for phonetic reduction in conversational speech, where the model represents the partially hidden VTRs and is defined as a constrained and simplified nonlinear dynamical system. Their algorithm computes the likelihood of an observation utterance while optimizing the VTR dynamics that account for long term context-dependent or coarticulatory effects in spontaneous speech. In their work, the hidden VTR dynamics are used as an intermediate representation for performing speech recognition, where many fewer model parameters had to be estimated as compared to tri-phone-based HMM baseline recognizers. Using the Switchboard dataset, they have shown reduction [26], [69] in word error rates when compared with baseline HMM models. Togneri *et al.* [110] used the hidden-dynamic model to represent speech dynamics and explored EKF, comparing its performance with the expectation–maximization (EM) algorithm to perform joint parameter and state estimation of the model. Deng *et al.* [27] proposed an efficient VTR tracking framework using adaptive Kalman filtering, and experiments on the Switchboard corpus demonstrated that their architecture accurately tracks VTRs for natural, fluent speech. In a recent study, Deng *et al.* [28] showed that a structured hidden-trajectory speech model exploiting the dynamic structure in the VTR space can characterize the long-term contextual influence among phonetic units. The proposed hidden-trajectory model [28] showed improvement in phonetic recognition performance on the TIMIT database for the four broad phone classes (sonorants, stops, fricatives, and closures) when compared with the HMM baseline.

D. Generative Models Using Deep Architectures

The first-order Markov chain assumption and the conditional independence assumption deter the HMM-based acoustic model's capabilities to account for most of the variability seen in natural speech. To account for the limited representability of the HMM-based acoustic models, generative models [49] with deeper architectures are currently being explored. Such deeper architectures have the capability to model streams of mutually interacting knowledge sources by representing them in multiple representation layers. A recent study by Mohamed *et al.* [81] has proposed a deep belief network [50] based acoustic model that can account for variability in speech stemming from the speech production process. A deep belief network is a probabilistic generative model consisting of multiple layers of stochastic latent variables [81]. Restricted Boltzmann machines (RBMs), owing to their efficient training procedure are used as the building block for deep belief networks. These authors applied a phone recognition task to the TIMIT corpus using MFCCs with delta (velocity) and delta-delta (acceleration) as the acoustic features and reported a phone error rate of 23%, compared to 25.6% obtained from Bayesian triphone HMM model reported in [76]. They have also shown that their system offers the least phone error rate compared to some previously reported results. Another recent study by Schrauwen *et al.* [100] proposed using a temporal reservoir machines (TRMs) which is a generative model based on directed graphs of RBMs. Their model uses a recurrent ANN to perform temporal integration of the input which is then fed to an RBM at each time step. They used the TRM to perform word recognition experiments on the TI46 dataset (subset of TIDIGITS corpus) and have used the Lyon passive ear model to parameterize the speech signal into 39 frequency bands. The smallest WER reported in their paper is 7%.

E. Articulatory Gesture Motivated Features for ASR Systems

Several efforts have been made [72], [107], [108] to design a speech recognition system that exploits articulatory information (akin to articulatory gestures) based on the human speech production mechanism. In particular, Sun *et al.* [107], [108] showed improvement in ASR performance by using an overlapping feature-based phonological model defined by general articulatory dynamics. Gestural activation recovery from the acoustic signal

has been performed by [2], [57] using a temporal decomposition method, where the gestural activations represent the time interval where a gesture is active. However, since it is the values of the dynamic parameters of active gestures (such as stiffness and target), that serve to distinguish utterances in a gesture-based lexicon [7], [10], estimating only gestural activation does not provide sufficient information for lexical access. To address this, Zhuang *et al.* [122] proposed a gestural pattern vector (GPV) as a recognition unit (which is an instance of gestural activation and corresponding dynamic parameters) and a model to predict the GPVs from the TVs. In their work [123], they assumed *a priori* knowledge of the TVs; using that knowledge they correctly recognized GPVs 80% of the time and reported that the estimated GPVs yielded a word recognition rate of 85% for a dictionary of 139 words. Unlike Zhuang *et al.*'s work, we do not explicitly assume *a priori* knowledge of the TVs; hence, we have explored [77] the feasibility and accuracy of estimating TVs from the speech signal, the major part of which is reported in this paper. TVs are not only beneficial for accurately recognizing gestures but also we have shown [78] that they can help in improving noise-robustness of ASR systems.

In our study, we use TVs (constriction degree and locations at the distinct constricting organs in the vocal tract) as articulatory information (instead of pellet trajectories) to model speech dynamics. The benefits of using TVs as opposed to the x and y coordinates of transducers attached to the articulators are threefold. First, as McGowan [74] pointed out, the TVs specify the salient features of the vocal tract area functions more directly than the articulators. Second, it is constrictions in TV space that articulatory gestures directly control [84], [98], and which embody the speaker's phonological goals. There is a one-to-many relation between TV values and pellet positions (both within and across speakers), and it is the TV value that is more informative in terms of phonological category and lexical access. There may be one TV specification in terms of constriction degree and location that can have many different sets of pellet positions in terms of Cartesian coordinates that represent the same vocal tract constriction. This difference is due to the fact that the pellets are absolute measurements whereas the TVs are relative measurements. 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. Thus, TVs can be expected to bear a relation to speech acoustics that is closer to one-to-one than does the complete area function, and help to reduce the non-uniqueness of speech inversion. Finally, we have shown in a different study that incorporating TV information (estimated from the acoustic signal) improves the performance of gesture recognition [80]. Hence, better and accurate ways of TV estimation would directly aid gestural recognition performance and in turn would aid in realizing an ASR system that uses speech gestures as sub-word units. As mentioned before, we intend to use the estimated TVs to recognize speech articulatory gestures. Further, we aim to realize an ASR architecture that uses these gestures as the sub-word level lexical representation of speech. In our gesture-based ASR architecture we intend to use pseudo-TVs (that should be speaker independent but will follow articulatory dynamics closely) as hidden intermediate variables between acoustic observations and articulatory gestures, thereby providing a cross-modal bridge between the continuous acoustic regime and the discrete articulatory regime (i.e., the gestural score). The estimation of TVs presented in this paper is the initial step in determining an appropriate model for the proposed task. We have previously proposed SVR architecture [77] for TV estimation and have shown that smoothing or low-pass filtering of the estimated TVs improved the result. In another study [79], we have shown that neural networks can be efficiently used for TV estimation, and the optimality of acoustic observation contextual information plays a critical role. In this paper, we perform a more extended study, deploying several machine-learning approaches and analyzing their performance for the TV estimation task. We also compare the performance of TV estimation with that of articulatory pellet trajectory estimation and show that the former is relatively more accurate than the latter. Not only can TVs contribute to a gesture-based ASR architecture, but they should also have applications in different areas such as in assistive devices (e.g., visual speech for the hearing impaired), audio-visual speech, speech production and synthesis, language acquisition, education, etc. We have also shown [78] that estimated TVs can help to improve noise robustness of ASR architecture.

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 explores several different machine learning strategies that we used for TV estimation: support vector regression (SVR), feedforward artificial neural network (FF-ANN), and autoregressive (AR) ANN, distal supervised learning (DSL), trajectory mixture density networks (TMDNs), and Kalman smoothing. Section IV presents the experiments, results and discussions, which are in two parts: 1) comparison of estimation performance between TVs and pellet trajectories and hence evaluation of the relevance of the TVs over pellet trajectories as articulatory information for acoustic-to-articulatory mapping and 2) a detailed description of our TV estimation procedure. The conclusion is given in Section V.

II. Dataset and Signal Parameterization

We aim to model speech using overlapping articulatory gestures, where the degree and extent of overlap between the gestures are determined by coarticulatory effects. Unfortunately, the spontaneous speech databases available for ASR do not come with any gestural specification. For this reason, TADA along with HLsyn [45], [46] (a parametric quasi-articulator synthesizer developed by Sensimetrics, Inc.) is used in this research (as shown in Fig. 4) to generate a database that contains synthetic speech along with their articulatory specifications. From text input, TADA generates gestural scores (time functions of gesture activation), TV time functions and simulated pellet trajectories. The simulated pellet trajectories correspond to the flesh-point locations specified in Fig. 1(b). It also generates a set of parameters that can be used by HLsyn to create synthetic speech. The synthetic database used in this research was generated by inputting the text for the 420 unique words found in the X-ray microbeam corpus [115]. The output

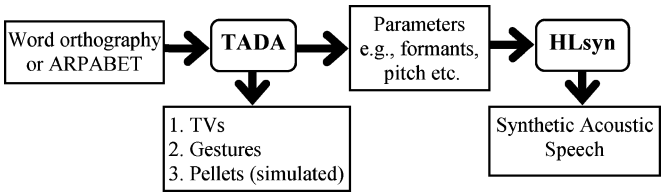


Fig. 4. Flow diagram for generating synthetic speech and the associated articulatory information using TADA and HLSyn.

synthetic speech was sampled at 10 kHz and the TV time functions and gestural scores were sampled at 200 Hz. Seventy-five percent of the data were used for training, ten percent for validation, and the rest for testing. It should be noted here that the target of the "critical" tract variable (e.g., LA for /b/) for a given phonological unit is invariant in TADA's lexicon, and therefore in the gestural score. However, the actual TV values generated are not invariant due to contextual dependency by coproduction. Hence, the TV values are not "predefined" for a given phoneme. In TADA, it is possible to implement different relative amounts of articulator contribution to TV constriction by modulating the articulator weights. For example, the same LA trajectory could be produced by different amounts of the contributions of the upper lip, lower lip, and jaw. In this study, however, only a single set of articulator weights for a given gesture was used. In our future studies, we intend to explore varying sets of articulator weights and expect that the results will show even more strongly that TVs for a given phonological unit are less variant than the pellets (flesh-point articulatory information).

The speech signal was parameterized as acoustic parameters (APs) and mel-frequency cepstral coefficients (MFCCs). APs [14], [56], [104] are knowledge-based acoustic-phonetic feature sets that provide phonetic information, such as formant values, pitch information, mean Hilbert envelope, energy onsets and offsets, and periodic and aperiodic energy in different subbands [29]. The APs were measured using a 10-ms window with a frame rate of 5 ms. For the APs, the feature dimension was much higher compared to the MFCCs; 40 different APs were selected (based upon their relevance). For the MFCCs, 13 cepstral coefficients were extracted. Each of these acoustic features was measured at a frame rate of 5 ms (time-synchronized with the TVs) with window duration of 10 ms. The acoustic features and the target articulatory information (the TVs and the simulated pellet trajectories) were z-normalized and then scaled such that their dynamic range is confined within $[-0.95, +0.95]$, except for SVR where the dynamic range is scaled between $[-1, +1]$. It has been observed [86], [95] that incorporating dynamic information helps to reduce the non-uniqueness problem for the speech inversion task; hence, the input features are contextualized in all the experiments reported in this paper. The feature contextualization is defined by the context-window parameter $\hat{C}$, where the current frame (with feature dimension d) is concatenated with $\hat{C}$ frames from before and after the current frame (with a frame shift of 2 or time shift of 10 ms), generating a concatenated feature vector of size $(2\hat{C}+1)d$. From our prior research [79], we have identified that the optimal context parameter $\hat{C}$ for the MFCCs is 8 (context duration of 170 ms) and for the APs is 9 (context duration of 190 ms) which will be used in the experiments presented in the rest of the paper.

III. Machine Learning Approaches for Speech Inversion

The process by which articulators in the human vocal tract produce the acoustic speech signal can be represented by a function f as

$$f : t \rightarrow x \tag{2}$$

where x is a vector that represents the acoustic speech signal, t is a vector representing the configuration of the articulators, and f is the function that defines the forward mapping from the articulatory domain to the acoustic domain. Thus, given a vector t_{a}, representing a specific articulatory configuration, we can obtain a specific speech output x_{a}, given f is known. In recognition tasks, the acoustic speech signal x_{a} is available to us with little or no articulatory data except what we can infer from the speech signal. If we define a function g such that

$$g : x \rightarrow t \tag{3}$$

then the articulatory configuration t_{b} can be obtained from the speech signal sample x_{b} using the function g. Thus, g is the inverse of function f and (3) represents the task of acoustic to articulatory speech inversion. Given the data-pair $[t_{\mathrm{b}}, x_{\mathrm{b}}]$, if g is estimated directly, then the resultant model is termed a direct inverse model. There are several indirect inverse model estimation approaches which do not seek to directly estimate g from the data-pair $[t_{\mathrm{b}}, x_{\mathrm{b}}]$.

Several machine learning techniques have been implemented for the task of speech inversion. Toutios *et al.* [112], [113] have used SVR to estimate electromagnetic midsagittal articulography (EMA) [97] trajectories for the MOCHA database and their results were found to be quite similar to that of the ANN-based approached proposed in [95]. ANN is widely known for its versatility in nonlinear regression problems. However, they fall short in ill-posed regression problems where the ill-posedness is due to one-to-many mapping. To address the one-to-many mapping scenarios, Jordan *et al.* [55] proposed supervised learning with distal teacher or distal supervised learning (DSL) and Bishop [3] proposed mixture density networks. While SVR and ANN-based approaches fall in the category of direct-inverse model, the DSL and the TMDN approaches can be identified as indirect inverse models. This section introduces the various machine learning techniques that we will explore in our speech inversion experiments.

A. Hierarchical Support Vector Regression

The support vector regression [102] is an adaptation of Vapnik's support vector classification algorithm [114] to the regression case. Given a set of N training vectors x_i and a target vector t such that $t_i \in \mathbb{R}$, the SVR algorithm seeks to find an optimal estimate (in terms of structural risk minimization) for the function $t = g(x)$, which has at most ε deviation from the actually obtained targets t_i for all the training data and at the

same time is as flat as possible. The ε-SVR algorithm defines that estimate as

$$g(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)\, k(x_i, x) + \beta \tag{4}$$

where $k(\,,\,)$ is the kernel used, β is the bias terms, and α_i, α_i^* are the coefficients obtained from the solution of the quadratic problem

$$\max\left[W(\alpha,\alpha^*)|0\leq\alpha,\alpha^*\leq C;\; i=1:N;\; \sum_{i=1}^{N}(\alpha_i^*-\alpha_i)=0\right] \tag{5}$$

where

$$W(\alpha,\alpha^*) = \sum_{i=1}^{N}[(\alpha_i^*-\alpha_i)t_i - \varepsilon(\alpha_i^*+\alpha_i)] - \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)k(x_i,x_j).$$

The constant C is the tradeoff between the flatness of g and the amount up to which deviations larger than ε are tolerated in the solution. $C>0$ and $\varepsilon\geq 0$ are parameters that are user-defined. C can be as high as infinity, while usual values for ε are 0.1 or 0.01. The kernel function $k(\,,\,)$ is used to transform the data into a high dimensional space to induce nonlinearity in the estimate function. SVR performs nonlinear regression by projecting the data into a high dimensional space via $k(\,,\,)$ and then performs linear regression in that space. We have used radial basis function (RBF) kernel with user-defined γ parameter

$$k(x,y) = \exp(-\gamma||x-t||^2). \tag{6}$$

B. Feedforward Artificial Neural Networks (FF-ANN)

Since Papcun *et al.* [90] used MLPs (layered ANNs using perceptron rule) to estimate articulatory trajectories for six English stop consonants, the potential of ANNs for the speech inversion task has been enthusiastically investigated. Zachs *et al.* [121] and Richmond [95] have studied the potential of ANNs for performing speech inversion. Once trained, ANNs require comparatively low computational resources compared to other methods both in terms of memory requirements and execution speed [79], [95]. ANN has the advantage that it can have M inputs and N outputs; hence, a complex mapping of M vectors into N different functions can be achieved. In such an architecture, the same hidden layers are shared by all the output TVs (shown in Fig. 5), which endows the ANNs with the implicit capability to exploit any cross-correlation that the TVs may have amongst themselves [79]. The FF-ANNs were trained with backpropagation using scaled conjugate gradient (SCG) algorithm [82].

C. Autoregressive Artificial Neural Networks (AR-ANN)

The estimated articulatory trajectories from SVR and FF-ANN-based direct inverse models were found to be corrupted by estimation noise. Human articulator movements are predominantly low pass in nature [52] and the articulatory trajectories usually have a smoother path, defined by one that does

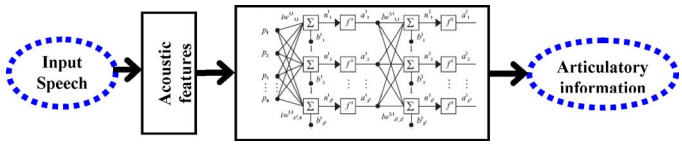


Fig. 5. Architecture of the ANN-based direct inverse model.

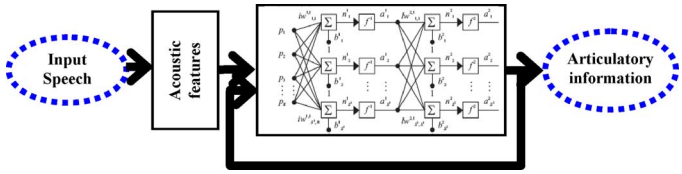


Fig. 6. Architecture of the AR-ANN-based direct inverse model.

not have any Fourier components over the cutoff frequency of 15 Hz. Nonlinear AR-ANN shown in Fig. 6, has a feedback loop connecting the output layer with the input, which helps to ensure smoother trajectories for the articulatory trajectories. The output of AR-ANN can represented as

$$\hat{y}(t) = g(\hat{y}(t-1), \hat{y}(t-2), \ldots, \hat{y}(t-d), u(t)). \tag{7}$$

The AR-ANN has its own disadvantages: 1) the architecture has to be trained with dynamic-backpropagation or backpropagation in time, which is computationally very expensive, 2) a single architecture cannot be trained easily for all the articulatory trajectories[2]; hence, a single AR-ANN has to be trained for each articulatory trajectory.

Both FF-ANN and AR-ANN are trained based on minimization of the sum-of-squares error approach. Given a set of training and target data set $[x,t]$ and a set of neurons with weights and biases defined by w and b, respectively, the sum-of-squares error is defined by

$$E_{SE}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{c}\left[g_k\left(x^i,w,b\right)-t_k^i\right]^2 \tag{8}$$

where $g_k\left(x^i,w,b\right)$ defines the network output, where the network is defined by weights w and biases b. Considering a dataset of infinite size, i.e., $N\to\infty$, (8) can be written as

$$E_{SE}(w,b) = \lim_{N\to\infty}\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{c}[g_k(x^i,w,b)-t_k^i]^2 \tag{9}$$

$$E_{SE}(w,b) = \frac{1}{2}\sum_{k=1}^{c}\int\int[g_k(x,w,b)-t_k]^2 p(t,x)dtdx. \tag{10}$$

The minimization of the error function E_{SE} with respect to $g_k(x,w,b)$ gives the following [3]:

$$\frac{\partial E_{SE}}{\partial g_k\,(x,w,b)} = 0. \tag{11}$$

Using (11) it can be shown that

$$g_k\,(x,w^*,b^*) = \mathrm{E}\,[t_k|x] \tag{12}$$

where $E[A|B]$ is the conditional expectation of A conditioned on B, w^* and b^* are the weights and biases of the network

[2]This may be because the dynamics of the different trajectories are different in nature and may not correlate so strongly with one another.

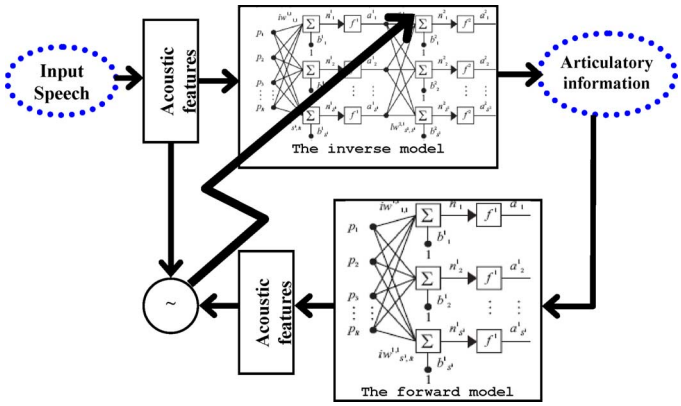


Fig. 7. Distal supervised learning approach for obtaining acoustic to TV mapping.

after training. Hence, (12) shows that networks that are optimized based on sum-of-squares approach generate average of the target data points conditioned on the input. Hence, direct inverse models obtained from supervised learning algorithms resolve one-to-M (where $M > 1$) inconsistencies by averaging [3], [55] across all the M candidates. If the set of M possible candidates form a non-convex set, then the average of the M candidates does not necessarily belong to that set; hence, the solution obtained is not necessarily the correct inverse solution.

D. Distal Supervised Learning (DSL)

To address the issues with conventional supervised learning architectures for one-to-many mapping cases, Jordan *et al.* [55], proposed supervised learning with a distal teacher or DSL. In the DSL paradigm, there are two models placed in cascade with one another: 1) the forward model (which generates acoustic features given the articulatory trajectories, hence M-to-1 mapping) and 2) inverse model (which generates the articulatory trajectories from acoustic features, hence 1-to-M mapping). Given a set of $[x_{\mathrm{b}}, y_{\mathrm{b}}]$ pairs, DSL first learns the forward model, which is unique but not necessarily perfect. DSL learns the inverse model by placing it in cascade with the forward model as shown in Fig. 7. The DSL architecture can be interpreted as an "analysis-by-synthesis" approach, where the forward model is the synthesis stage and the inverse model is the analysis stage. In the DSL approach, the inverse model is trained (its weights and biases updated) using the error that is backpropagated through the forward model whose previously learned weights and biases are kept constant.

Considering a forward mapping between an input vector x and an output vector y, using a vector of network weights and biases, w and b, the relationship can be expressed as

$$\hat{t} = g(x, w, b). \tag{13}$$

Learning the forward model is based on the following cost function [55]:

$$L = \frac{1}{2} E\left[(t-\hat{t})^T (t-\hat{t})\right] \tag{14}$$

where t is the desired target for a given input. For the inverse model, [55] defined two different approaches, a local optimization approach and an optimization along the trajectory approach. The local optimization approach necessitates using an online learning rule, whereas the optimization along trajectory requires recurrency in the network (hence, error minimization using backpropagation in time), both of which significantly increase the training time and memory requirements. In this paper, we propose a global optimization approach, which uses the tools of DSL as proposed in [55], but instead uses batch training in the feedforward network. The cost function that the DSL tries to minimize is represented as

$$J = \frac{1}{2N} \sum_{k=1}^{N} \left[(t_k^* - t_k)^T (t_k^* - t_k)\right] \tag{15}$$

where N is the total number of training samples, t_k is the target vector for the kth training sample and t_k^* is the actual target output from the network. The weight update rule is as follows:

$$w[n+1] = w[n] - \eta \nabla_w J_n \tag{16}$$

where η is the learning rate, $w[n]$ represents the weights of the network at time index n. The gradient can be obtained from (15) using the chain rule

$$\nabla_w J_n = \frac{1}{N} \sum_{k=1}^{N} \left(-\frac{\partial x_k{}^T}{\partial w} \frac{\partial t^{*T}_{k,n}}{\partial x_k} (t_k - t^*_{k,n}) \right) \tag{17}$$

where $t^*_{k,n}$ is the estimated target vector for the kth training sample at the nth time instant.

E. Trajectory Mixture Density Networks (TMDN)

Mixture density networks (MDNs) [3] combine the conventional feedforward ANNs with a mixture model. In MDN architectures the ANN maps from the input vector x to the parameters of a mixture model (shown in Fig. 8) to generate a conditional pdf of the target t, conditioned on the input x. Typically, a GMM is used in the MDN setup because of their simplicity and the fact that a GMM with appropriate parameters can approximate any density function. A Gaussian kernel is represented as

$$k_i(t|x) = \frac{1}{(2\pi)^{0.5c} \sigma_i(x)^c} \exp\left[-\frac{\|t - \mu_i(x)\|^2}{2\sigma_i(x)^2} \right] \tag{18}$$

where x and t are the input and the target vector, $\mu_i(x)$ is the center of the ith kernel, and $\sigma_i(x)$ is the spherical covariance (this assumption can be relaxed by considering either a diagonal or a full covariance) for each Gaussian kernel and c is the input dimension. In this setup, the probability density of the target data conditioned on the input using a GMM with m mixtures can be represented as

$$p(t|x) = \sum_{i=1}^{m} \alpha_i(x) k_i(t|x) \tag{19}$$

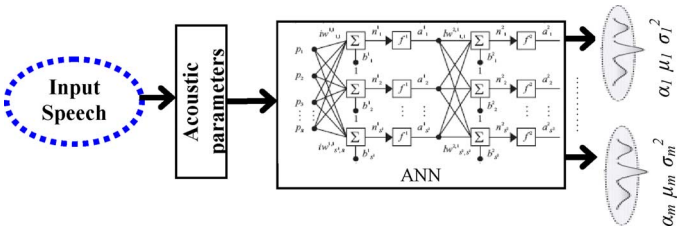


Fig. 8. MDN architecture.

where $\alpha_i(x)$ is the prior probability and $k_i(t|x)$ is the conditional probability density given the ith kernel. To satisfy the following conditions for the prior probabilities

$$\sum_{i=1}^{m} \alpha_i(x) = 1 \text{ and } 0 \leq \alpha_i(x) \leq 1. \tag{20}$$

The following "softmax" function is used to define $\alpha_i(x)$ [3]

$$\alpha_i = \frac{\exp(z_i^{\alpha})}{\sum_{l=1}^{m} \exp(z_l^{\alpha})} \tag{21}$$

where z_i^{α} is the ANN output corresponding to the prior probability for the ith mixture of the GMM component. The variances and means of the GMM model are related to the ANN outputs as follows:

$$\sigma_j = \exp(z_j^{\sigma}) \quad \text{and} \quad \mu_{jk} = z_{jk}^{\mu} \tag{22}$$

where z_i^{σ} and z_i^{μ} are the ANN outputs corresponding to the variance and the mean of the jth mixture. The MDN is trained by minimizing the following cost function:

$$E_{MDN} = -\ln\left[\sum_{i=1}^{m} \alpha_i(x^N) k_i(t^N|x^N)\right]. \tag{23}$$

As seen in Fig. 8, the ANN part of MDN generates the GMM parameters which are used to estimate the cost function E_{MDN}. The cost function E_{MDN} is minimized with respect to the ANN weights and biases.

The derivative of the cost function is evaluated separately with respect to the priors, means and variances of the mixture model that are back-propagated through the network to yield the derivative of the cost function with respect to the network weights and biases, more details available at [3]. The standard MDN architecture provides the conditional probability density of the targets conditioned on the input. To estimate the articulatory trajectories from the conditional probability densities, a maximum-likelihood parameter generation (MLPG) algorithm was proposed in [111]. The MLPG algorithm was used with MDN architecture in [96] and the resulting architecture was named as the trajectory MDN or (TMDN). In TMDN architecture, the target vector is augmented with dynamic information to yield a vector sequence O as shown as follows:

$$O = [o_1^T, o_2^T, \ldots o_n^T, \ldots, o_T^T]^T, \text{ where } o_n = [t_n^T, \Delta t_n^T, \Delta\Delta t_n^T]^T. \tag{24}$$

In our work, the dynamic target vectors are calculated as

$$\Delta t_n = \sum_{\tau=-T/2}^{T/2} w(\tau)\, t_{n+\tau} \tag{25}$$

$$\Delta\Delta t_n = \sum_{\tau=-T/2}^{T/2} w(\tau)\, \Delta t_{n+\tau} \tag{26}$$

where $(T+1)$ is the total duration of the window and the window is defined as

$$\begin{aligned} w(\tau) &= m(\tau)\omega_{ham}(\tau) \\ \text{where, } m(\tau) &= \begin{cases} -1, & \text{if } \tau < 0 \\ +1, & \text{otherwise} \end{cases} \\ \text{and, } \omega_{ham}(\tau) &= \left[0.54 - 0.46\cos\left(\frac{2\pi\tau}{\mathrm{T}}\right)\right] \end{aligned} \tag{27}$$

where $\omega_{ham}(\tau)$ is a hamming window. The vector O can be related to the target vector by the following relation, where the details about the transformation matrix W can be found from [109], [111].

$$\begin{aligned} O &= WT \\ T &= [t_1, t_2, \ldots, t_N]^T \\ W &= [w_1, w_2, \ldots, w_N]^T. \end{aligned} \tag{28}$$

In TMDN architectures the augmented feature vector O is used to train the MDN models, where O is derived from the target vector T using the transformation matrix W. The MDN in such a case gives the following conditional density $\mathrm{P}(o_n|x_n)$. For the simplest case, where the GMM in the MDN has a single mixture, the target trajectory is generated by maximizing $\mathrm{P}(O|\lambda)$ or $\mathrm{P}(WT|\lambda)$ with respect to T as shown in (29), where λ is the mixture sequence:

$$\frac{\partial \log P(WT|\lambda)}{\partial T} = 0. \tag{29}$$

A set of linear equations are generated (detailed derivation given in [111]) from (29), as

$$W^T\Sigma^{-1}WT = W^T\Sigma^{-1}M^T \tag{30}$$

where

$$\begin{aligned} \Sigma^{-1} &= diag\left[\Sigma_{\lambda_1}^{-1}, \Sigma_{\lambda_2}^{-1}, \ldots \Sigma_{\lambda_n}^{-1}, \ldots \Sigma_{\lambda_N}^{-1}\right] \\ M &= \left[\mu_{\lambda_1}^T, \mu_{\lambda_2}^T, \ldots \mu_{\lambda_n}^T, \ldots \mu_{\lambda_N}^T\right]^T \end{aligned} \tag{31}$$

$\mu_{\lambda 1}$ and $\Sigma_{\lambda 1}^{-1}$ are the 3×1 mean vector and the 3×3 diagonal covariance matrix (for a single mix GMM). Solving (30) for T gives the required maximum-likelihood trajectory. For MDNs with multiple mixtures, the approximation with suboptimal mixture sequence technique discussed in [109] is used.

F. Kalman Smoothing

The estimated articulatory trajectories were found to be corrupted with estimation noise from all except the AR-ANN

model. It was observed that smoothing the estimated articulatory trajectories improved estimation quality and the correlation and reduced the RMSE. This is a direct consequence of the observation made in [52], which claimed that articulatory motions are predominantly low pass in nature with a cutoff frequency of 15 Hz. This led us to introduce a Kalman smoother-based postprocessor in the architectures discussed above. Since articulatory trajectories are physical quantities, they can be approximately modeled as the output of a dynamic system. For the proposed architecture, we selected the following state-space representation

$$\begin{aligned} y_n &= F y_{n-1} + w_{n-1} \\ t_n &= H y_n + v_n \end{aligned} \tag{32}$$

with the following model parameters:

$$\begin{aligned} F &= \begin{bmatrix} 1 & \Gamma \\ 0 & 1 \end{bmatrix} \text{ and } H = [1 \quad 0] \\ y_0 &\sim \mathrm{N}(y_0, \hat{y}_0, \Sigma_0) \\ w_n &\sim \mathrm{N}(w_n, 0, Q) \\ v_n &\sim \mathrm{N}(v_n, 0, R) \end{aligned} \tag{33}$$

where Γ is the time difference (in milliseconds) between two consecutive measurements, $y_n = [y_n^{\mathrm{p}} y_n^{\mathrm{v}}]^T$ is the state vector and contains the position and velocity of the articulatory trajectories at time instant n.t_n is the estimated articulatory trajectory which is considered as noisy observation of the first element of the state y_n. The variables w_n and v_n are process and measurement noise, which have zero mean, known covariance Q and R, and are considered to be Gaussian. The goal is to find the smoothed estimate of the state $y_{n|N}$ given the observation sequence $T = \{t_1, t_2, t_3 \ldots, t_N\}$, that is

$$y_{n|N} = E[y_n \mid t_1, t_2 \ldots, t_N]. \tag{34}$$

Although, F and H are known parameters of the state space representation, the unknown parameter set $\Theta = \{Q, R, \hat{y}_0, \Sigma_0\}$ should be learned from the training set. After learning the unknown parameter set $\Theta = \{Q, R, \hat{y}_0, \Sigma_0\}$ the smoothed state $y_{n|N}$ is estimated by the Kalman Smoother in an optimal sense.

IV. Experiments, Results, and Discussion

In our experiments, we demonstrate that given a speech signal, tract variables can be estimated with a high accuracy. We begin our experiments by comparing the performance of TV estimation with pellet trajectory estimation, where we will show that TVs can be estimated more accurately than the pellet trajectories. Next in Section III, we perform a detailed analysis of TV estimation using the machine learning algorithms. The speech signal was parameterized as MFCCs and APs and then contextualized as discussed in Section II. The shape and dynamics of the estimated articulatory trajectories were compared with the actual ones using three quantitative measures: the root mean-squared (rms) error, mean normalized rms error [59] and the Pearson product-moment correlation (PPMC) coefficient. The RMSE gives the overall difference between the actual and the estimated articulatory trajectories, whereas the PPMC gives a measure of amplitude and dynamic similarity between them. The RMSE and the PPMC are defined as follows:

TABLE II
Optimal Number of Neurons for Each Articulatory Trajectory for 1-Mix MDN

TVs	MFCC	AP	Pellets	MFCC	AP
GLO	60	45	ULx	15	45
VEL	90	60	ULy	90	90
LA	60	45	LLx	60	90
LP	15	45	LLy	105	30
TBCL	105	30	JAWx	90	75
TBCD	45	15	JAWy	15	105
TTCL	60	60	TTx	105	15
TTCD	60	30	TTy	75	60
			TDx	30	15
			TDy	45	30

$$RMSE = \sqrt{\frac{1}{N}(e-t)^T(e-t)} \tag{35}$$

$$r_{PPMC} = \frac{N\sum_{i=1}^{N} e_i t_i - \left[\sum_{i=1}^{N} e_i\right]\left[\sum_{i=1}^{N} t_i\right]}{\sqrt{N\sum_{i=1}^{N} e_i^2 - \left(\sum_{i=1}^{N} e_i\right)^2}\sqrt{N\sum_{i=1}^{N} t_i^2 - \left(\sum_{i=1}^{N} t_i\right)^2}} \tag{36}$$

where e represents the estimated TV vector and t represents the actual TV vector having N data points. The RMSE provides a performance measure in the same units as the measured articulatory trajectories. Some of the TVs have a different measuring unit (e.g., TBCL and TTCL are measured in degrees) from the pellet trajectories (all pellet trajectories are measured in mm). Thus, to better summarize the inversion performance for all articulatory trajectories, we use the non-dimensional mean normalized RMSE, RMSE_{nrm} [59] and its average, RMSE_{nrm_avg} defined by

$$\begin{aligned} RMSE_{nrm\_i} &= \frac{RMSE_i}{\sigma_i} \\ RMSE_{nrm\_avg} &= \frac{1}{\mathrm{N}}\sum_{i=1}^{N} RMSE_{nrm\_i} \end{aligned} \tag{37}$$

where N is the number of articulatory trajectories considered (8 for TVs and 14 for pellet trajectories).

A. Comparing TV and Pellet Trajectory Estimates

TMDN has been used by Richmond [96] to estimate articulatory pellet trajectories for the multichannel articulatory MOCHA dataset [120]. Results from [96] indicate that TMDN offers much better accuracy over ANN for pellet trajectory estimation. Using a similar approach as laid out in [96], we trained individual MDN models for each articulatory trajectory, where the articulatory trajectories were augmented with static, delta, and delta-delta features as shown in (24). The MDN was built such that it generated the parameters of a GMM model with diagonal covariance matrix; yielding the parameters for a 3-D Gaussian mixture (one dimension for each feature stream of static, delta, and delta-delta features). The models were trained with one to four mixture components, but increasing

TABLE III
PERFORMANCE COMPARISON BETWEEN TV AND PELLET TRAJECTORY ESTIMATION

	TVs				Pellets trajectories			
	MFCC		*AP*		*MFCC*		*AP*	
	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$	$RMSE_{nrm\_avg}$	$PPMC_{avg}$
1-hidden ANN	*0.462*	0.881	*0.465*	0.886	*0.507*	0.838	*0.507*	0.849
TMDN	*0.443*	0.891	*0.456*	0.891	*0.493*	0.846	*0.499*	0.854
3-hidden FF-ANN	*0.313*	0.948	*0.317*	0.944	*0.410*	0.889	*0.407*	0.898

the number of mixtures did not show any appreciable improvement of the results in our case; hence, we will be presenting the results from the single mixture MDN only. The MDNs were built with a single hidden layer architecture, where the number of neurons in the hidden layer was optimized using the validation set. Table II shows the optimal number of neurons for each articulatory trajectory for each acoustic feature type. The networks were trained with the SCG algorithm using a maximum of 4000 training iterations. After the MDNs were trained, the MLPG algorithm was run ad-hoc on the resulting sequence of MDN generated pdfs for the validation set. The RMSE between the estimated and the groundtruth articulatory trajectory was used as the validation error.

The mean of the static features generated by the MDN should be equivalent to the output of a single hidden layer ANN [96] having linear activation functions, as noted from (12); these outputs are considered as single-hidden layer ANN outputs. The TMDN as well as the ANN outputs for each articulatory trajectory were processed with a Kalman smoother and the results are shown in Table III. The Kalman smoother was found to improve the PPMC on an average by 3% for both TVs and pellets.

In addition, 3-hidden layer FF-ANN architectures with tan-sigmoid activation were implemented for both the TVs and pellet trajectories. The FF-ANN architectures had as many output nodes as there are articulatory trajectories (eight trajectories for TVs and 14 trajectories for pellet data). Single 3-hidden layer FF-ANN architecture was realized for each articulatory information type (i.e., TVs and Pellet trajectories) and for each feature type (MFCC or AP). The number of neurons in each hidden layer was optimized by analyzing the RMSE from the validation set. During the optimization stage we observed that the performance of the articulatory trajectory estimation improved as the number of hidden layers was increased. It may be the case that additional hidden layers incorporated additional nonlinear activation functions into the system, which increased the potential of the architecture to cope with the high nonlinearity inherent in a speech-inversion process. However the number of hidden layers was confined to three because 1) the error surface becomes more complex (with many spurious minima) as the number of hidden layers are increased, thereby increasing the probability that the optimization process finds a local minimum and 2) increasing the number of hidden layers increases the training time as well as the network complexity. The optimal ANN architectures for the MFCCs and APs were found to be 150-100-150 and 250-300-250,[3] where the numbers represent the number of neurons in each of the three hidden layers. The 3-hidden layer FF-ANNs were trained with a target epoch of 5000 and the estimated trajectories were processed with a Kalman smoother. Post processing with Kalman smoothing decreased the RMSE on an average by 9%.

Table III shows the $RMSE_{nrm\_avg}$ and PPMC of all the TVs and Pellet trajectories from the three approaches discussed above. Note that lower RMSE and higher PPMC indicate better performance of the estimation. Table III shows that overall, the 3-hidden layer FF-ANN offered both lower RMSE and higher PPMC in both TV and pellet estimation tasks compared to the TMDN and 1-hidden layer ANN. Some of the TVs involve articulator movements that should be observed in particular pellet trajectories, whereas the others are not comparable to the pellet data at all. For example, the TV GLO represents the vibration of the vocal folds thereby distinguishing voiced regions from unvoiced ones. There is no such information present in the pellet trajectories as it is almost impossible to insert pellet transducers within the vocal chords. The TV-pellet sets that are closely related to one another are as follows:$\{\mathrm{LP} : \mathrm{ULx}, \mathrm{LLx}\}$; $\{\mathrm{LA} : \mathrm{ULy}, \mathrm{LLy}\}$, $\{\mathrm{TTCL}, \mathrm{TTCD} : \mathrm{TTx}, \mathrm{TTy}\}$, and $\{\mathrm{TBCL}, \mathrm{TBCD} : \mathrm{TDx}, \mathrm{TDy}\}$. Table IV lists the obtained PPMC for the related TV and pellet trajectory estimates from the 3-hidden layer FF-ANN when MFCCs are used as the acoustic features.

There are several important observations from Table III: 1) overall the TV estimates offered better PPMC coefficients and mean normalized rms error ($RMSE_{nrm\_avg}$) than the pellet trajectories, 2) TMDN always showed improvement over the 1-hidden layer ANN model having the same number of neurons with linear activation function, and 3) the 3-hidden layer FF-ANN with nonlinear activation showed overall the best performance.

Observations from Table III are further confirmed in Table IV, which shows that for the best performing architecture, that is the 3-hidden layer ANN, the estimated TVs overall offered higher PPMC coefficient as compared to the relevant pellet trajectory estimates. It should be pointed out here that the average PPMC for 3-hidden layer FF-ANN shown in Tables III and IV are not the same, as Table III shows the average across all the TVs/pel-

[3]The optimal number of neurons in the hidden layers was found to be very similar for TV and pellet estimation for a given acoustic feature; hence, we have used the same configuration for both the types of speech inversion task.

TABLE IV
COMPARISON OF PPMC BETWEEN RELEVANT ARTICULATORY PELLETS AND TVS FOR 3-HIDDEN LAYER ANN USING MFCC

TVs	PPMC	Pellets	PPMC
LP	0.927	LLx	0.788
		ULx	0.918
LA	0.894	LLy	0.889
		ULy	0.738
TTCL	0.951	TTy	0.945
TTCD	0.949	TTx	0.929
TBCL	0.968	TDy	0.974
TBCD	0.962	TDx	0.969
Avg	**0.942**	*Avg*	**0.894**

lets and Table IV shows the average across only the relevant set of TVs/pellets as specified above. The results are indicative of the fact that the TVs can be estimated more accurately from the speech signal than the pellet trajectories. Two reasons may explain this difference. First, according to [74], the TVs specify acoustically salient features of the vocal tract area functions more directly than the pellet information. Second, the TVs (i.e., the constriction location and degree) are intrinsically relative measures, whereas the pellet trajectories provide arbitrary flesh-point location information in the 2-D Cartesian coordinate system and are required to go through normalization [95]. Since the normalization process is sensitive to the nature of data, the relative nature of the information is not effectively captured. It should be noted, however, that such pellet-trajectory-associated problems were not overly severe in our experiment because, unlike the case of natural speech, there were no distortion in the data (as the data was synthetically generated using TADA) introduced by intra- and inter-speaker variability. Finally, note that better performance of TVs does not seem to hold for the tongue body TVs. This can be possibly attributed to the different roles played by the tongue body in speech. Tongue body TVs are controlled primarily for vowels which do not usually involve very narrow constrictions in the vocal tract (although velar consonants (e.g., /k/ and /g/) do employ it). It can thus be said that TVs are superior for representing articulations with narrow constrictions (consonants), since such constrictions will have a disproportionate influence on the acoustics [105]. For example, TB constriction for a coproduced vowel will produce little modulation of the acoustics of stop closure or fricative noise, while consonantal constriction will have a very large influence, determining if there is silence or turbulence. Also note that our main goal in retrieving articulatory information is to incorporate that information for the purpose of articulatory gesture estimation. Since articulatory gestures are action units that inherently define constriction location and degree along the vocal tract, it can be surmised that the TVs would be more appropriate intermediate entities between acoustic observations and articulatory gestures rather than flesh-point pellet trajectories. Thus, even if pellet-trajectories are recovered more accurately than the TVs (which is not found to be the case here), they could not be expected to perform as good as the TVs in the estimation of articulatory gestures.

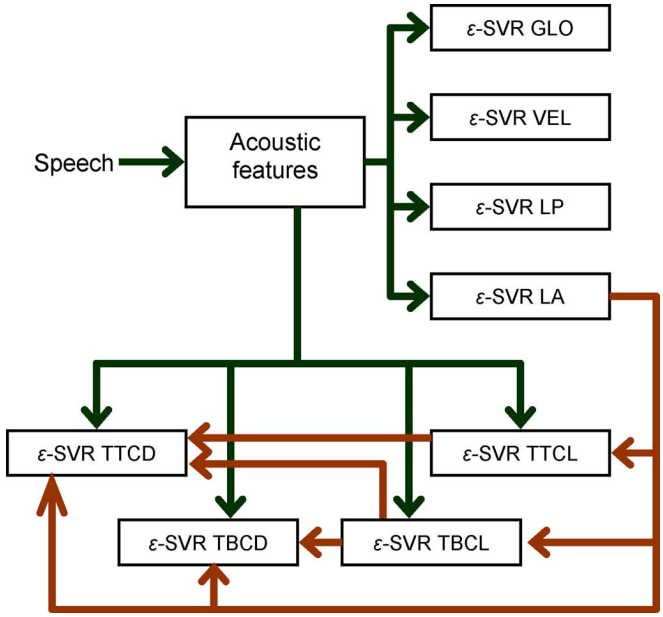


Fig. 9. Hierarchical ε-SVR architecture for generating the TVs.

B. TV Estimation: Additional Details

In this section, we will provide a more detailed analysis of the TV estimation processes. Apart from the machine learning approaches explored in the last section, we will examine SVR, AR-ANN and finally DSL for TV estimation and then compare their performance with that of the MDN and FF-ANN architectures presented in last section.

1) Hierarchical SVR: We have previously proposed [77] a nonlinear regression using a support vector regression (SVR) framework for TV estimation using APs as the acoustic feature. In the current work, we analyze the SVR performance for both MFCCs and APs and contextualize them as stated in Section II. Separate SVR models with RBF kernel were trained for each TV, where the set of APs[4] for each model was selected based upon their relevance. We observed that certain TVs (TTCL, TBCL, TTCD, and TBCD) are known to be functionally dependent upon other TVs, while the remaining TVs (GLO, VEL, LA, and LP) are relatively independent and can be obtained directly from the acoustic features. This dependency is used to create the hierarchical architecture shown in Fig. 9. From the results of the validation set the optimal value of C was found to be 1.5 and γ was set equal to $1/\mathrm{d}$ based on [112], [118], where d = dimension of the input feature set.

2) AR-ANN: The estimated TVs from TMDN, FF-ANN, and SVRs were found to be fairly noisy, which necessitated the use of Kalman smoother postprocessing. As articulatory movements are inherently low pass in nature, maintaining smoother trajectories is a desired task in speech inversion task. Using an autoregressive architecture is suitable for such an application, as the feedback loop helps to retain the smoothness of the estimated trajectories. Individual AR-ANN models were trained separately for each of the TVs. A 2-hidden layer AR-ANN model with tan-sigmoid activation, SCG training (using 5000 epochs) with dynamic backpropagation was used.

[4]The number of pertinent APs for each TV is shown in [77]

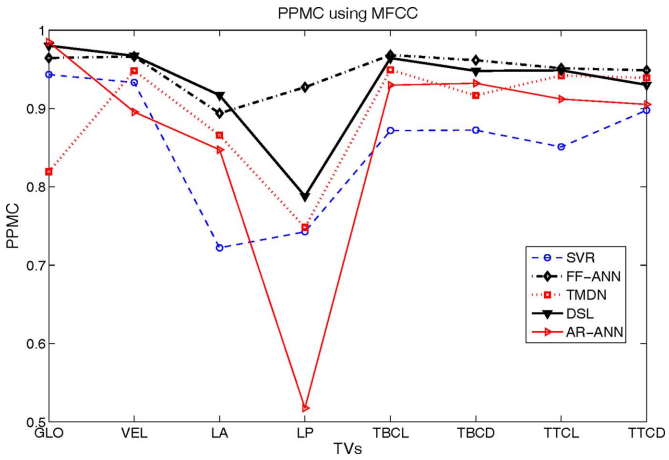


Fig. 10. PPMC for TV estimation from different architectures using MFCC.

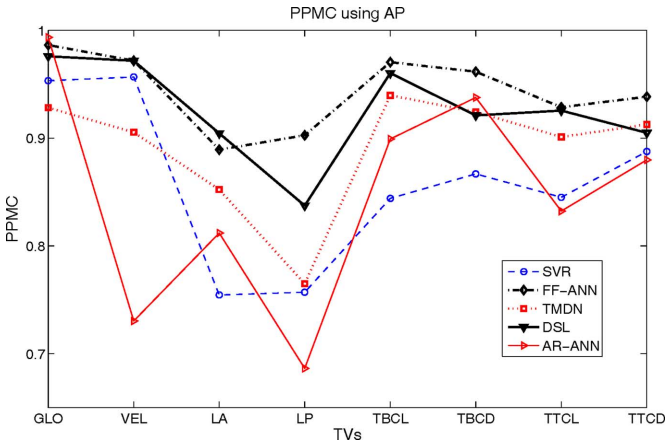


Fig. 11. PPMC for TV estimation from different architectures using AP.

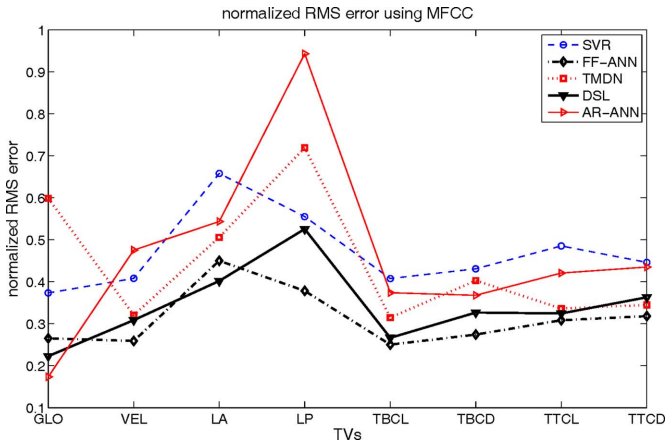


Fig. 12. Normalized RMSE for TV estimation from different architectures using MFCC.

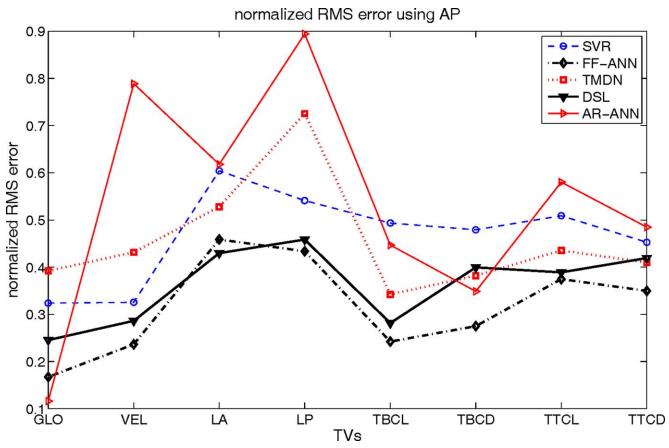


Fig. 13. Normalized RMSE for TV estimation from different architectures using AP.

The number of neurons in each hidden layer was optimized and for all the models the number of neurons within each hidden layer was confined within 25 to 200. A unit delay[5] was used in each of the AR-ANN architecture. The TV estimates from the AR-ANNs were not noisy hence were not postprocessed with the Kalman smoother.

3) DSL Architecture: A single DSL architecture was trained for all the eight TV trajectories for each acoustic feature. The forward models were created using single hidden-layer FF-ANN and trained using SCG algorithm. The number of neurons in the hidden layer was optimized using the rms error over the validation set. The inverse models were built using a 3-hidden-layer network and the number of neurons in each layer was optimized using the rms error on the validation set. The DSL models were trained using gradient descent learning algorithm (with a variable learning rate), momentum learning rule (momentum $= 0.9$) and mean squared predicted performance error [55] with regularization as the optimization criteria (regularization parameter $= 0.4$). The number of neurons in the forward model was 350 and 400 and in the inverse model were 150-100-150 and 250-300-250 for MFCC and AP, respectively.

4) Comparison of TV Estimation Architectures and Their Performance: The TV estimation results from TMDN, 3-hidden layer FF-ANN, SVR, AR-ANN, and DSL are shown in Figs. 10–13 for both APs and MFCCs. It can be observed from the plots that the 3-hidden layer FF-ANN architecture overall offered superior performance over the other approaches, closely followed by the DSL technique. For LA, DSL always performed better than the 3-hidden layer FF-ANN. The worst performance was observed from SVR and AR-ANN architectures. The feedback loop in the AR-ANN architecture helps to maintain the inherent smoothness of the articulatory trajectories but at the same time can be a source of progressive error introduction. If the AR-ANN model makes a significant error at any time instant, that error gets fed back to the system, resulting in progressive error in subsequent estimates. The TMDN results though were not as good as the 3-hidden layer FF-ANN, but were much better most of the time than the SVR and AR-ANN architectures.

Table V presents the RMSE and PPMC coefficients for all the TVs, obtained from the 3-hidden layer FF-ANN architecture for both the acoustic features. As noted from Table I, different TVs have different measuring units and dynamic ranges; hence, accordingly the RMSE needs to be interpreted. For example GLO and VEL have a very small dynamic range and hence very small

[5]Multiple delays were also tested, but were not found to yield appreciable improvement in performance.

TABLE V
RMSE AND PPMC FROM 3-HIDDEN LAYER FF-ANN

	MFCC		AP	
	RMSE	PPMC	RMSE	PPMC
GLO	0.0305	0.9645	0.0192	0.9863
VEL	0.0172	0.9663	0.0157	0.9718
LA	1.5962	0.8939	1.6266	0.8893
LP	0.3663	0.9272	0.4196	0.9026
TBCL	6.9464	0.9683	6.7244	0.9704
TBCD	1.0125	0.9617	1.0145	0.9616
TTCL	4.8963	0.9514	5.9456	0.9286
TTCD	2.3367	0.9487	2.5679	0.9384

TABLE VI
PPMC FOR FF-ANNS WITH DIFFERENT NUMBER OF HIDDEN LAYERS FOR MFCC

	GLO	VEL	LA	LP	TBCL	TBCD	TTCL	TTCD
1-hidden layer	0.942	0.951	0.872	0.928	0.956	0.946	0.929	0.928
2-hidden layer	0.960	0.961	0.885	0.925	0.967	0.960	0.940	0.939
3-hidden layer	0.965	0.966	0.894	0.927	0.968	0.962	0.951	0.949

RMSE. On the contrary, TBCL and TTCL are measured in degrees and have a larger dynamic range than the others; hence, their RMSE is in degrees and the values are larger than the others.

Table V shows that the APs overall offered better accuracy for GLO and VEL, whereas for the other TVs, the MFCCs provided better results. The APs have specific parameters for detecting voicing (e.g., periodic and aperiodic energies at different subbands) and nasalization (ratio of the energy in BW [0 to 320 Hz] and energy in BW [320 to half the sampling rate] measured in dB). Thus, GLO and VEL are better captured using the APs.

The different architectures described in this paper targeted different aspects of the speech inversion process. For example, AR-ANN targeted the inherent smoothness (low-frequency nature) of the TVs and the DSL and TMDN architecture were designed to explicitly address the non-uniqueness involved in speech inversion, whereas the 3-hidden layer FF-ANN targeted the nonlinearity of the speech inversion task. The better performance of the 3-hidden layer FF-ANN suggests that nonlinearity is the most critical aspect of TV estimation from the speech signal. The nonlinearity in the FF-ANNs is imparted by the tan-sigmoid activations used in the hidden layers. We observed that increasing the number of hidden layers in the FF-ANN architecture resulted in an increase in the PPMC and simultaneous decrease in the RMSE, as shown in Table VI, where the FF-ANN had eight output nodes (one for each TV). From Table VI it can be seen that increasing the number of hidden layers increased the PPMC consistently for all but LP.

From these observations, we reiterate Qin *et al.*'s [91] claim that non-uniqueness may not be a critical problem for speech inversion although their work was focused on pellet-trajectory-based speech inversion. McGowan [74] pointed out that the non-uniqueness problem with speech inversion is ameliorated by the use of TVs as there may be one articulatory specification

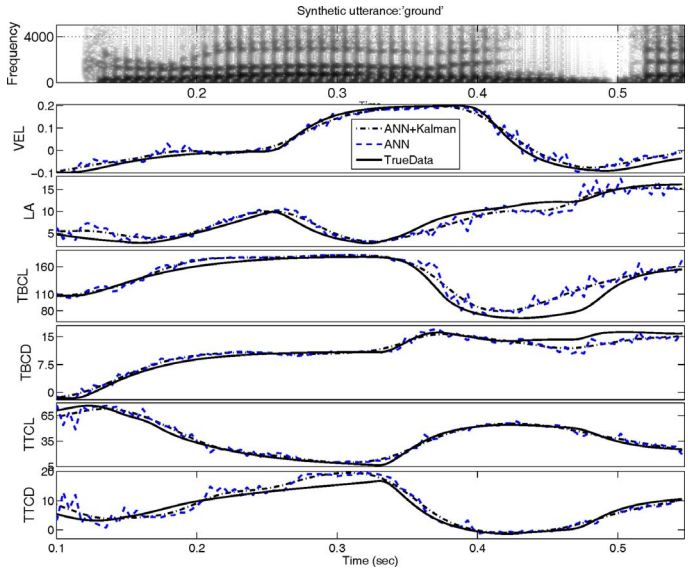


Fig. 14. Actual and estimated TVs from ANN and ANN+Kalman using MFCC as the acoustic feature.

(in terms of constriction degree and location) which can have many different sets of articulatory location (in Cartesian coordinates) that represent the same vocal tract constriction. Hence, for TVs we can expect a further (if at all any) reduction in non-uniqueness for the speech inversion task. It is well known that speech to articulatory inversion is a primarily nonlinear problem [95] and this fact could be the driving force behind the success of the 3-hidden layer FF-ANN. The DSL approach uses a similar architecture as the 3-hidden layer FF-ANN, but its inability to match the performance of the latter can be due to the inaccuracies in the forward model. As pointed out before, the DSL topology is more like an analysis-by-synthesis architecture, where the performance of synthesis part entirely depends upon the accuracy of the forward model. To ensure a highly accurate forward model, exhaustive data is typically required to ensure the forward model has examples of all possible pairs of articulatory data and acoustic observation. However, in a real-world scenario such an exhaustive data may not be always practical rendering the inaccuracy of the forward model. An example of the predicted trajectories from the 3-hidden layer FF-ANN for five different TVs (VEL, LA, TBCL, TBCD, TTCL, and TTCD) is shown in Fig. 14, for the synthetic utterance "a ground." It can be seen that the raw trajectories from the FF-ANN architecture are much noisier and the Kalman-smoothing helped to reduce that noise efficiently.

V. CONCLUSION

We have demonstrated using a TADA generated dataset that TV estimation can be done with overall better accuracy than estimation of articulatory pellet trajectories. This result suggests that TVs may be better candidates than pellet trajectories for articulatory feature-based ASR systems. Analysis of different approaches to TV estimation suggests that for the synthetic dataset we used, nonlinearity is the governing factor rather than non-uniqueness for speech inversion using TVs. We draw this conclusion since the 3-hidden layer FF-ANN architecture,

which models well the nonlinearity inherent in speech inversion, offered much better accuracy over the other competing approaches. The 3-hidden layer FF-ANN is simpler to construct and even simpler to execute when trained; hence, it would be an ideal candidate for TV estimation in a conventional ASR system or gesture-based ASR system. Currently, none of the natural speech corpora contain TV information. If and/or when such a database becomes available, similar analyses need to be performed to validate the applicability of the FF-ANN architecture for TV estimation.

ACKNOWLEDGMENT

The authors would like to sincerely thank the Associate Editor, Dr. L. Deng and the two anonymous reviewers who have helped with their valuable comments and suggestions to improve the quality and clarity of this paper.

REFERENCES

[1] 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.

[2] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Proc. ICASSP*, 1983, pp. 81–84.

[3] C. Bishop, "Mixture density networks," Neural Computing Research Group, Dept., Comput. Sci., Aston Univ., Birmingham, U.K., Tech. Rep. NCRG/4288.

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

[5] C. P. Browman and L. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140–155, 1988.

[6] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonol.*, vol. 6, pp. 201–251, 1989.

[7] C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *J. Phonetics*, vol. 18, no. 3, pp. 299–320, 1990.

[8] C. P. Browman and L. Goldstein, "Representation and reality: Physical systems and phonological structure," *J. Phonetics*, vol. 18, pp. 411–424, 1990.

[9] C. P. Browman and L. Goldstein, "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Lab. Phon. I: Between the Grammar and the Physics of Speech*, J. Kingston and M. E. Beckman, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1991, pp. 341–376.

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

[11] D. Byrd and E. Saltzman, "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," *J. Phonetics*, vol. 31, no. 2, pp. 149–180, 2003.

[12] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *Proc. ICASSP*, 2007, vol. 4, pp. 645–648.

[13] S. Chang, M. Wester, and S. Greenberg, "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," *Speech Commun.*, vol. 47, no. 3, pp. 290–311, Nov. 2005.

[14] S. Chen and A. Alwan, "Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position," in *Proc. ICSLP*, 2000, vol. 4, pp. 113–116.

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

[16] J. Clark and C. Yallop, *An introduction to Phonetics and Phonology*, 2nd ed. Oxford, U.K.: Blackwell, 1995.

[17] G. N. Clements and E. V. Hume, "The internal organization of speech sounds," in *Handbook of Phonological Theory*, J. A. Goldsmith, Ed. Cambridge, U.K.: Blackwell, 1995.

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

[19] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proc. 4th Euro. Conf. Speech Commun. Technol.*, 1995, vol. 1, pp. 821–824.

[20] L. Deng and K. Erler, "Microstructural speech units and their HMM representations for discrete utterance speech recognition," in *Proc. ICASSP*, 1991, pp. 193–196.

[21] 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.

[22] 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.

[23] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulator features for all classes of English sounds," in *Proc. ICASSP*, 1994, pp. 45–47.

[24] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Spec. Iss. Speech Prod. Modeling, Speech Commun.*, vol. 22, no. 2, pp. 93–112, 1997.

[25] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, no. 4, pp. 299–323, 1998.

[26] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," *J. Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[27] L. Deng, L. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. ICASSP*, 2004, pp. I557–I560.

[28] L. Deng, D. Yu, and A. Acero, "Structured speech modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1492–1504, Sep. 2006.

[29] O. Deshmukh, C. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: Detection of the periodicity and aperiodicity profile of speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 776–786, Sep. 2005.

[30] S. Dusan, "Statistical estimation of articulatory trajectories from the speech signal using dynamical and phonological constraints," Ph.D., Univ. of Waterloo, Dept. of Elect. Comput. Eng., Waterloo, ON, Canada, 2000.

[31] K. Elenius and G. Tacacs, "Phoneme recognition with an artificial neural network," in *Proc. Eurospeech*, 1991, pp. 121–124.

[32] K. Elenius and M. Blomberg, "Comparing phoneme and feature based speech recognition using artificial neural networks," in *Proc. ICSLP*, 1992, pp. 1279–1282.

[33] 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.

[34] C. Y. Espy-Wilson and S. E. Boyce, "The relevance of F4 in distinguishing between different articulatory configurations of American English /r/," *J. Acoust. Soc. Amer.*, vol. 105, no. 2, p. 1400, 1999.

[35] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of American English /r/," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 343–356, 2000.

[36] C. A. Fowler and E. Saltzman, "Coordination and coarticulation in speech production," *Lang. Speech*, vol. 36, pp. 171–195, 1993.

[37] C. A. Fowler and L. Brancazio, "Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation," *Lang. Speech*, vol. 43, pp. 1–42, 2000.

[38] C. A. Fowler, "Speech production and perception," in *Handbook of Psychology*, A. Healy and R. Proctor, Eds. New York: Wiley , 2003, vol. 4, Experimental Psychology, pp. 237–266.

[39] 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.

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

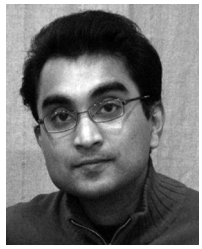
[41] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," in *Proc. Int. Conf. Spoken Lang. Process.*, Korea, 2004, pp. 1202–1205.

[42] J. Frankel and S. King, "A hybrid ANN/DBN approach to articulatory feature recognition," in *Proc. Eurospeech, Interspeech*, 2005, pp. 3045–3048.

[43] O. Fujimura, S. Kiritani, and H. Ishida, "Computer controlled radiography for observation of movements of articulatory and other human organs," *Comput. Biol. Med.*, vol. 3, pp. 371–384, 1973.

[44] O. Fujimura, "Relative invariance of articulatory movements: An iceberg model," in *Invariance & Variability of Speech Processes*, J. S. Perkell and D. Klatt, Eds. Mahwah, NJ: Lawrence Erlbaum Assoc., 1986, ch. 11, pp. 226–242.
[45] H. M. Hanson, R. S. McGowan, K. N. Stevens, and R. E. Beaudoin, "Development of rules for controlling the HLsyn speech synthesizer," in *Proc. ICASSP*, 1999, vol. 1, pp. 85–88.
[46] 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.
[47] J. Harris, *English Sound Structure*. Oxford, U.K.: Blackwell, 1994.
[48] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko, "Audiovisual speech recognition with articulator positions as hidden variables," in *Proc. ICPhS*, Saarbrucken, Germany, 2007, pp. 297–302.
[49] X. He and L. Deng, *Discriminative Learning for Speech Processing*, G. H. Juang, Ed. San Mateo, CA: Morgan & Claypool, 2008.
[50] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.
[51] J. Hogden, A. Löfqvist, 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.
[52] J. Hogden, D. Nix, and P. Valdez, "An articulatorily constrained, maximum likelihood approach to speech recognition," Los Alamos National Laboratory, Los Alamos, NM, 1998, Tech. Rep. LA-UR–96-3945.
[53] J. Hogden, P. Rubin, E. McDermott, S. Katagiri, and L. Goldstein, "Inverting mappings from smooth paths through Rn to paths through Rm. A technique applied to recovering articulation from acoustics," *Speech Commun.*, vol. 49, no. 5, pp. 361–383, 2007.
[54] F. J. Huang, E. Cosatto, and H. P. Graf, "Triphone based unit selection for concatenative visual speech synthesis," in *Proc. ICASSP*, Orlando, FL, 2002, vol. 2, pp. 2037–2040.
[55] M. I. Jordan and D. E. Rumelhart, "Forward models—Supervised learning with a distal teacher," *Cogn. Sci.*, vol. 16, pp. 307–354, 1992.
[56] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, Univ. of MD, College Park, 2004.
[57] T. P. Jung, A. K. Krishnamurthy, S. C. Ahalt, M. E. Beckman, and S. H. Lee, "Deriving gestural scores from articulator-movement records using weighted temporal decomposition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 2–18, 1996.
[58] 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.
[59] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 3, pp. 411–422, Mar. 2009.
[60] 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, 2000.
[61] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from Switchboard 1," in *Proc. Interspeech*, 2005, pp. 3385–3388.
[62] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Univ. of Bielefeld, Bielefeld, Germany, 1999.
[63] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Commun.*, vol. 37, pp. 303–319, 2002.
[64] T. Kobayashi, M. Yagyu, and K. Shirai, "Application of Neural networks to articulatory motion estimation," in *Proc. ICASSP*, 1985, pp. 1001–1104.
[65] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1027–1035, 1978.
[66] J. Laver, *Principles of Phonetics*. Oxford, U.K.: Oxford Univ. Press., 1994.
[67] 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.
[68] 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," in *Proc. ICASSP*, 2007, vol. 4, pp. 621–624.
[69] J. Ma and L. Deng, "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech," *Comput., Speech, Lang.*, vol. 14, pp. 101–104, 2000.
[70] S. Y. Manuel and R. A. Krakow, "Universal and language particular aspects of vowel-to-vowel coarticulation," *Haskins Lab. Star. Rep, Speech Res. SR-77/78*, pp. 69–78, 1984.
[71] S. Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *J. Acoust. Soc. Amer.*, vol. 88, pp. 1286–1298, 1990.
[72] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Commun.*, vol. 48, pp. 161–175, 2006.
[73] A. Martinet, "Phonetics and linguistic evolution," in *Manual of Phon.*, B. Malmberg, Ed. Amsterdam, The Netherlands: North-Holland, 1957, pp. 252–272.
[74] 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, 1994.
[75] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP*, 2002, pp. 2133–2136.
[76] J. Ming and F. J. Smith, "Improved phone recognition using Bayesian Triphone Models," in *Proc. ICASSP*, 1998, pp. 409–412.
[77] 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.
[78] 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*, U.K., 2009, pp. 2759–2762.
[79] V. Mitra, H. Nam, and C. Espy-Wilson, "A step in the realization of a speech recognition system based on gestural phonology and landmarks," in *Proc. 157th Meeting ASA*, Portland, 2009, vol. 125, J. Acoust. Soc. Amer., p. 2530.
[80] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Recovering speech gestures using a cascaded neural network," *J. Acoust. Soc. Amer.*, submitted for publication.
[81] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.
[82] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, pp. 525–533, 1993.
[83] 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. 24, no. 5, pp. 365–379, Oct. 1976.
[84] 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. 5-2, p. 2430, 2004.
[85] D. Neiberg, G. Ananthakrishnan, and O. Engwall, "The acoustic to articulation mapping: Non-linear or Non-unique?," in *Proc. Interspeech*, 2008, pp. 1485–1488.
[86] S. E. G. Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. Amer.*, vol. 39, pp. 151–168, 1966.
[87] T. Okadome, S. Suzuki, and M. Honda, "Recovery of articulatory movements from acoustics with phonemic information," in *Proc. 5th Seminar Speech Production*, Bavaria, Germany, 2000, pp. 229–232.
[88] M. K. Omar and M. Hasegawa-Johnson, "Maximum mutual information based acoustic features representation of phonological features for speech recognition," in *Proc. ICASSP*, 2002, vol. 1, pp. 81–84.
[89] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE Auto. Speech Recog. Understanding Workshop*, 1999, vol. 1, pp. 79–83.
[90] 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.
[91] 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.
[92] M. G. Rahim, W. B. Kleijn, J. Schroeter, and C. C. Goodyear, "Acoustic-to-articulatory parameter mapping using an assembly of neural networks," in *Proc. ICASSP*, 1991, pp. 485–488.
[93] 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.

[94] D. Recasens, "Timing constraints and coarticulation: Alveolo-palatals and sequences of alveolar + [j] in Catalan," *Phonetica*, vol. 41, pp. 125–139, 1984.
[95] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D., Univ. of Edinburgh, Edinburgh, U.K., 2001.
[96] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," *Lecture Notes in Comput. Sci.*, vol. 4885/2007, pp. 263–272, 2007.
[97] J. Ryalls and S. J. Behrens, *Introduction to Speech Science: From Basic Theories to Clinical Applications*. Boston, MA: Allyn & Bacon, 2000.
[98] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 332–382, 1989.
[99] O. Schmidbauer, "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations," in *Proc. ICASSP*, 1989, pp. 616–619.
[100] B. Schrauwen and L. Buesing, "A hierarchy of recurrent networks for speech recognition," in *Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.
[101] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Commun.*, vol. 5, pp. 159–170, 1986.
[102] A. Smola and B. Scholkhopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.
[103] K. N. Stevens, "Toward a model for speech recognition," *J. Acoust. Soc. Amer.*, vol. 32, pp. 47–55, 1960.
[104] K. N. Stevens, S. Manuel, and M. Matthies, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production," in *Proc. Int. Cong. Phon. Sci.*, 1999, vol. 2, pp. 1117–1120.
[105] K. N. Stevens, *Acoustic Phonetics (Current Studies in Linguistics)*. Cambridge, MA: MIT Press, 2000.
[106] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1872–1891, 2002.
[107] J. Sun and L. Deng, "Annotation and use of speech production corpus for building language-universal speech recognizers," in *Proc. 2nd Int. Symp. Chinese Spoken Lang. Processi. ISCSLP*, Beijing, China, Oct. 2000, vol. 3, pp. 31–34.
[108] 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, 2002.
[109] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of speech parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[110] R. Togneri and L. Deng, "Joint state and parameter estimation for a target-directed nonlinear dynamic system model," *IEEE Trans. Signal Process.*, vol. 51, no. 12, pp. 3061–3070, Dec. 2003.
[111] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Jun. 2000, vol. 3, pp. 1315–1318.
[112] A. Toutios and K. Margaritis, "A support vector approach to the acoustic-to-articulatory mapping," in *Proc. Interspeech*, 2005, pp. 3221–3224.
[113] A. Toutios and K. Margaritis, "Learning articulation from cepstral coefficients," in *Proc. SPECOM*, 2005.
[114] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[115] J. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook*. Madison: Univ. of Wisconsin, 1994.
[116] M. Wester, S. Greenberg, and S. Chang, "A dutch treatment of an elitist approach to articulatory-acoustic feature classification," in *Proc. Eurospeech*, 2001, pp. 1729–1732.
[117] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *Proc. Inst. Electronics, Info, Commun. Engi. Beyond HMM Workshop*, 2004, vol. 104, pp. 37–42, SP2004-81-95.
[118] J. Weston, A. Gretton, and A. Elisseeff, *SVM Practical Session – How to Get Good Results Without Cheating*. Tuebingen, Germany: Machine Learning Summer School, 2003.
[119] C. Windheuser, F. Bimbot, and P. Haffner, "A probabilistic framework for word recognition using phonetic features," in *Proc. ICSLP*, 1994, pp. 287–290.
[120] A. Wrench, The MOCHA-TIMIT Articulatory Database 1999 [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html
[121] J. Zachs and T. R. Thomas, "A new neural network for articulatory speech recognition and its application to vowel identification," *Comput., Speech, Lang.*, vol. 8, pp. 189–209, 1994.
[122] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "The entropy of articulatory phonological code: Recognizing gestures from tract variables," in *Proc. Interspeech*, 2008, pp. 1489–1492.
[123] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "Articulatory phonological code for word classification," in *Proc. Interspeech*, 2009, pp. 2763–2766.



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

He is currently a Research Assistant for the Speech Communication Laboratory, Institute of Systems Research (ISR), University of Maryland. His research interest is 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 at 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, 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, in 1981, 1984, and 1987, respectively.

She is a Professor in the Electrical and Computer Engineering Department 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 serving as 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 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 Ph.D. degree from the Institute of Child Development, University of Minnesota, Minneapolis, 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 taught at Yale University in linguistics from 1980 to 2007. Since 2007, he has been a Professor of linguistics at the University of Southern California, Los Angeles. 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.

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