A Novel Neural-Based Pronunciation Modeling Method for Robust Speech Recognition

Guangpu Huang [1], Meng Joo Er [2]

Computer Vision Lab, Nanyang Technological University
50 Nanyang Avenue, Republic of Singapore
[1] hu0002pu@e.ntu.edu.sg
[2] emjer@ntu.edu.sg

***Abstract*—This paper describes a recurrent neural network (RNN) based articulatory-phonetic inversion (API) model for improved speech recognition. And a specialized optimization algorithm is introduced to enable human-like heuristic learning in an efficient data-driven manner to capture the dynamic nature of English speech pronunciations. The API model demonstrates superior pronunciation modeling ability and robustness against noise contaminations in large-vocabulary speech recognition experiments. Using a simple rescoring formula, it improves the hidden Markov model (HMM) baseline speech recognizer with consistent error rates reduction of 5.30% and 10.14% for phoneme recognition tasks on clean and noisy speech respectively on the selected TIMIT datasets. And an error rate reduction of 3.35% is obtained for the SCRIBE-TIMIT word recognition tasks. The proposed system qualifies as a competitive candidate for profound pronunciation modeling with intrinsic salient features such as generality and portability.**

I. Introduction

Conventional automatic speech recognition (ASR) systems treat the speech sound as a concatenation of acoustic observations during probabilistic modeling, e.g., the hidden Markov models (HMMs). This "beads-on-the-string" approach proves to be extremely problematic when dealing with the dynamic nature of speech sounds in practical applications largely due to its lack of intelligibility at phone-level and lack of robustness in noisy conditions [1].

In recent years, the articulatory and the auditory based feature representations have gained particularly attention to solve the problem in various applications with moderate success, e.g., the acoustic landmark detection systems in [2]. There is a rising body of ASR research which attempts to utilize the knowledge of human speech production and perception so as to better explain the dynamic nature of human speech [3]. This prospect is also motivated by the neurological studies with the discovery of "mirror" neurons that are potentially involved in both the speech production and perception process [4]. In fact, artificial neural networks (ANNs) based have already been applied to utilize the articulatory with the acoustic knowledge sources both in speech recognition as well as in speech synthesis [5], [6].

However, when using fixed-structure ANNs to model the nonlinearities of the correlated feature spaces: the articulatory gestures and the acoustic cues, there are two major difficulties. On the one hand, there exists multiple articulatory configurations that could produce the same acoustic output, i.e., the "many to one" problem. On the other hand, human perception of speech is "categorical", whereas acoustically variant sound units are perceived with binary-like "yes or no" decisions [7].

In this work, a modest modeling method, namely the neural network based articulatory-phonetic inversion (API) model, is proposed to utilize the multi-dimensional feature representations for improved ASR performance. Different from the pseudo-articulatory features which are simply transcribed by broad phonological classes, e.g., manner of articulation (MOA) and place of articulation (POA), as used in [5], a more reliable API procedure is proposed using recurrent neural networks (RNNs). The neural networks are excellent universal approximators with simple topological structure, and fast learning algorithms can be easily implemented because of the locally tuned neurons, which have assured their suitability for pattern classification problems [8]. In addition, the proposed API model is supplied with a specialized adaptation algorithm which embeds a set of heuristic learning rules to address the many-to-one problems of articulation and to approximate the categorical nature of audition.

What signifies the proposed method is that the two aspects of human speech: articulation and audition, are exploited collectively to form a *unified explanation* for the intrinsic and extrinsic variations of speech pronunciations. And the knowledge sources in the multi-dimensional spaces are utilized in an extended system which is ready for use in practical ASR applications.

The rest of the paper is arranged as follows. Section II explains the multi-dimensional pronunciation modeling method in detail: the theoretical basis in Section II-A, the structure of the API model in Section II-B, and the heuristic learning algorithm in Section II-C. Speech recognition experiments are presented in Section III, and results are summarized and discussion in Section IV. The paper concludes in Section V.

II. Multi-Dimensional Pronunciation Modeling

A. Theoretical Basis

In general, human speech production involves four stages in a top-down fashion: semantic conceptualization, phonological

encoding, phonetic encoding, and articulation [4]. Of particular interest is the articulatory-acoustic interface in the final three stages, where the aerodynamic and myoelastic energy in the articulatory space are transformed into acoustic sound waves. On the other hand, human speech perception involves four stages in a bottom-up fashion: peripheral perception, cochlear filtering, transduction, and semantic abstraction [9]. Here the acoustic-auditory interface in the first three stages witnesses the transformation of acoustic energies into electrical potentials across the inner hair cell (IHC) membranes.

Thus the speech sound can be treated as "the information carrier" in certain transmission channel, e.g., the air, with articulatory gestures and auditory impressions at its two terminals. The semantic aspects is omitted since the primary goal of ASR is to act as a speech-to-text "court recorder" [1].

As Chomsky has noted, the normal use of language is in many aspects a " creative activity", which is unique to human and distinguishes itself from animal communication. There are two ends on Chomsky's notion of language usage in the context of speech recognition: the "many-to-one" issue in articulation and the categorical nature of perception. In the literature, the "many-to-one" mapping between articulatory movements and acoustic cues can be approximated using various target approximation schemes which calculate "the minimum amount of articulatory movements" required for audible sound change [10]. And the categorical perception of phonemes can be approximated through boundary conditions using decision-tree based classifiers or through various Gaussian-mixture models (GMMs), which aim for "the maximum amount of cross-class phone discriminations" [11], [12].

It follows that the pronunciation variations of human speech project non-linearly into the different feature spaces. Thus we aim to implement a mapping and decoding method to recover the underlying articulatory-phonetic features (APFs) to improve feature representations of speech signal.

B. RNN-Based Articulatory-Phonetic Inversion

A schematic view of the neural network based API model is demonstrated in Fig. 1. The API model aims to infer the articulatory motor commands, e.g., muscular activities, from the input acoustic parameters, e.g., discrete formant measures or mel-frequency cepstral coefficients (MFCCs). In the proposed system, RNNs with time-delay designed using the NICO Toolkit [13] are employed to model the relation between the acoustic and the articulatory information. Each network consists of three layers: input, hidden and output layer, and the networks are fully connected. A temporal context window of ten frames was used to balance the trade-off between the number of parameters and classification accuracy as suggested in [14]. The activation function of the output layer is the softmax function:

$$f(x_i) = \frac{exp(x_i)}{\sum_{k=1}^{K} exp(x_k)}, \tag{1}$$

where K is the number of units in the output layer.

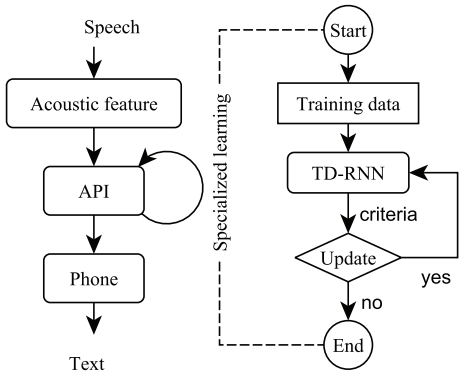


Fig. 1. Overview of the RNN based API model with the specialized learning algorithm.

The off-line inversion process is divided into two computational blocks. Firstly the input acoustic feature vectors are mapped into articulatory configurations, i.e., articulatory inversion. Secondly the articulatory patterns are mapped into phonemes, i.e., phonetic inversion. In particular, the connectionist model is globally tuned using two parental datasets before being used for on-line speech recognition. The first set of training data for articulatory inversion is supplied by a bio-mechanical articulatory synthesizer. Though there are various collections of direct physiological measurements available, e.g., the electromagnetic articulograph (EMA), and laryngograph i.e., electroglottograph (EGG) [6], they are usually too limited for ASR studies due to the difficulty of data collection. Moreover, these corpora are generally "target-specific", i.e., the speakers are well-learned, and the produced sound patterns are subjectively designed, which render them too sparse for our purposes.

Therefore, we exploited the existing sources of vastly available speech recordings and speech synthesizers. The bio-mechanical speech synthesizer used here consists of 28 muscular structures from 12 major groups with reference to [15], [16]. The tube walls of each articulator are described by width Δx, length Δy, and depth Δz. These muscular groups assimilate the human anatomy both in physiological properties and in functionality of speech production [15].

For our experiments, 10 critical muscles are closely monitored for APF retrieval. The 30 channel articulatory gestures consist of the $(\Delta x, \Delta y, \Delta z)$ coordinate of the 8 muscles: two intrinsic tongue muscles which change the shape of the tongue body, i.e., verticalis (VS), and transversus (TS), three extrinsic tongue muscles which change the position the tongue body, i.e., genioglossus (GG), hyoglossus (HG), and styloglossus (SG), three facial muscles, i.e., masseter (MS) which raises or lowers the jaw, risorius (RO) and orbicularisoris (OO) the combination of which constrict, round, or spread the lips, the levatorpalatini (LP) which controls the velopharyngeal port, or velum, the the lungs (LG) which control the pulmonic

mechanism, and the interarytenoid (IA) which controls the glottalic mechanism. The maximal and minimal moving ranges of the muscle width are normalized to be in the $[-1, +1]$ interval. A value outside this range indicates extra compressive or de-compressive myoelastic tension of the tissues but does not introduce further vocal tract deformation. For the highly elastic tongue muscles, the boundary values would result in constriction at certain parts of the vocal tract walls, i.e., POA [14]. While the LP only assumes two values for English phones, 0 for nasals and 1 for the rest.

And $500,000$ configurations of phone-level articulatory parameters are randomly selected from the synthesized 100 TIMIT SX sentences. One obvious advantage of this set of data is that articulatory-acoustic pairings are more objectively distributed due to the explicit control of the vocal apparatus movements compared to human speakers.

The second set of training data for phonetic inversion is the same 100 phonetically rich TIMIT SX sentences repeated by 10 human speakers, 5 male and 5 female, with the "Received Pronunciation" (RP), extracted from the SCRIBE corpus [17]. There are two reasons that the SCRIBE corpus is chosen. Firstly, the selected corpus gives detailed annotation by trained linguists using about 290 detailed phonetic labels, where TIMIT is annotated with 45 generalized IPA phonetic labels of English phonemes. The latter have also been widely used as the basic units of conventional HMM based acoustic models. While they may be sufficient for experienced human listeners, they are not optimal for pronunciation modeling in ASR. Secondly, SCRIBE serves the purpose of balancing between the perceptual invariance by multiple human listeners and the articulatory effort by human speakers. These details are not available in the TIMIT phone-set, and the representations are more compact and ready for use than the physiological data such as EMA measures. So it stands a better chance of knowledge preservation and accurate phonetic representations than the others.

C. Specialized API learning algorithm

To enhance the inversion processing during off-line training, a specialized API learning algorithm is implemented. The algorithm is analogous to the human experiences of speech acquisition through the language teachers, i.e., the bio-mechanical synthesizer is taught by the 20 RP speakers. In the forward learning mode, the algorithm iteratively updates the network weights to minimize the Root Mean Square (RMS) error at the output layer. The process is monitored by two heuristic measures summarized as follows.

Rule 1: Listener oriented optimization of acoustic qualities of the synthesized speech using the relative entropy, $H(\mu||\sigma)$. Since not all of the synthesized acoustic sound result in audible or meaningful outputs, a feature evaluation criterion is applied to refine the raw space based on the relative entropy measure, $H(\mu||\sigma)$, defined as:

$$H(\mu||\sigma) = H_i - H^i_{ref}, \tag{2}$$

where

$$H_i = \sum_{x}^{no.offrames} \mu_i(x) log \frac{\mu_i(x)}{\sigma_i(x)}, \tag{3}$$

in which $\mu_i(x)$ & $\sigma_i(x)$ are the centre values and the standard deviation of acoustic feature vectors for all the frames of the i_{th} phone in the reduced target regions, and H^i_{ref} is the reference value computed from the training dataset.

Rule 2: Speaker oriented minimization of articulatory cost, C_{RAT}, during regional target approximation.

In the synthesis model, the equilibrium positions and the "regional articulatory target" (RAT) are defined to smooth the dynamic articulatory transactions using the target approximation modeling methods, which have previously been used to improve the performance of speech recognizers and synthesizers in [6], [12]. The production cost in the API model is calculated as the articulatory effort required or energy consumed to move from the current place to the nearest point in the target region, or to return to the equilibrium position.

$$C_{RAT} = \frac{\partial W}{\partial E} \quad = \frac{2a}{b} \times \frac{\partial (exp(b(I_1 - 3)) - p(I_3 - 1)^2)}{\partial (FF^T - I)}, \tag{4}$$

where W is the stored strain energy described by the seven components of F, E is the Lagrangian strain tensor, I_1 and I_3 are the first and the third invariant of the deformation tensor E: $I_1 = 3 + 2Tr(E), I_3 = det(2E + I)$, $a, b\&p$ are tuning variables, and the internal force F in the muscle is defined as:

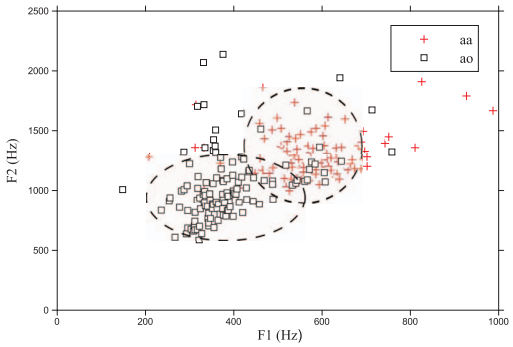
$$F = [M]\frac{\partial^2 \vec{V}}{\partial t^2} + [C]\frac{\partial \vec{V}}{\partial t} + [K]\vec{V}, \tag{5}$$

where [M] is the mass matrix, [C] the damping matrix, [K] the elasticity matrix, and $\vec{V} = [x(t), y(t), z(t)]$ is the displacement vector. The minimization of this criterion keeps the variation of the muscular activities as low as possible near the target region, and it has a tendency of bringing the articulators back to equilibrium position, which is in accordance with human speech production [18].

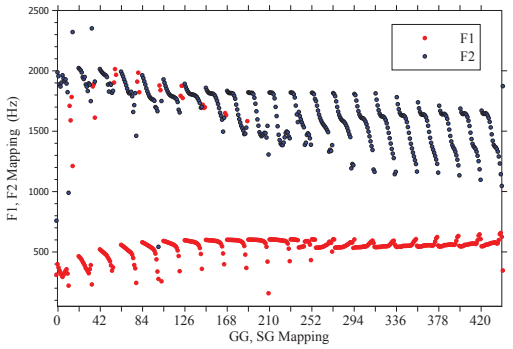
In this manner, the two criterion iteratively spans or shrinks the articulatory target, resulting in a distributed RAT for each phone in the articulatory space. The process is demonstrated in Fig. 2.

In the acoustic space, simple spectral measures, formants: F1 and F2, are plotted for two monophones, /aa/ and /ao/, 100 utterances each by 5 speakers from the TIMIT database, in Fig. 2(a). The entropy measure, H, for the two feature vectors are $[F1(622.2, 168.9), F2(1297.2, 220.7)]$ for /aa/, and $[F1(562.9, 51.4), F2(1054.4, 299.2)]$ for /ao/, all units in Herz (Hz). For continuous speech, /aa/ is often in a reduced form with a relatively large F2 variance that causes misclassification errors as shown by the overlapped concentration ellipses in Fig. 2(a). In the first articulatory inversion step, two APFs, GG and SG, which are the principle controlling parameters for tongue body forwarding and back-raising respectively, are illustrated in Fig. 2(b). For clarity, 441 pairs of (GG, SG) configurations are inverted from the bio-mechanical articulatory

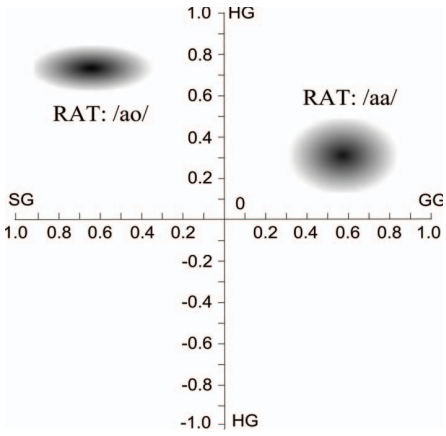
configurations using a singular width change $\Delta x = 0.1$ from -1 to 1. In the second phonetic inversion step, the final RAT at the RNNs output layer is computed as a regional distribution, where the center-point is desirable to reach but not necessarily compulsory to tolerate pronunciation variations within phone classes, as shown in Fig. 2(c) . Using the APF representation, /aa/ and /ao/ show clear distinction in the articulatory space. The center-point of the RAT, (GG_μ, HG_μ, SG_μ), is located at the central-front position for /aa/:(0.55, 0.30, 0.0), and at a back position for /ao/: (0.0, 0.71, 0.60), shown as the darkened gradient in Fig. 2(c).



(a) Acoustic formant map



(b) Articulatory inversion



(c) Phonetic inversion

Fig. 2. Illustration of the articulatory-phonetic inversion process using formant measures for two vowels: /aa/ and /ao/, uttered by 20 speakers from the TIMIT database.

III. Design of Speech Recognition Experiments

A. Acoustic Features

The baseline acoustic feature stream is provided by the standard ETSI Distributed Speech Recognition MFCCs, which consist of 13 cepstral coefficients including the log energy $C0$ [19]. MFCCs are calculated by taking the discrete cosine transform (DCT) of the log powers at each of the mel frequency bands. In this study, the 13 static cepstral coefficients ($C0 - C12$) were augmented with the dynamic delta and acceleration coefficients which are calculated using two frames of past context and two frames of future context in the HTK module, resulting in a 39-dimensional feature vector.

B. Recognizer Back End and Recognition Tasks

To test our hypothesis that the RNN based API module has indeed "learned" the articulatory-phonetic mappings and is capable of dealing with the intrinsic and extrinsic pronunciation variations of speech, the API model is firstly used as a stand-alone phoneme recognizer. Then it is integrated with the conventional HMM baseline through phone rescoring for both phoneme and word recognitions. For our experiments, the weighted product rule based on the generalized confidence score adapted from [20] is used:

$$p(o_t|A) = C \prod_{i=1}^{N} p(o_t|A^i)^{\gamma^i}, \tag{6}$$

where A is the resulting parameters in the combined system, C is a normalization constant, A^i is the set of acoustic parameters for the i^th system, $p(o_t|A^i)$ is the likelihoods computed by the i^{th} classifier, and γ^i is the interpolation weight of the i^{th} knowledge source. This multiplication task becomes a simple summation in the logarithmic domain. In the rescoring module, the posterior phoneme log-likelihood are computed at each state using the weighted scheme in (6). For the experiments here, C is set to 1, and the sum of the interpolation weights, γ^i, is forced to be 1 such as that used in [5], and [14].

The HMM recognizers are based on continuous density left-to-right HMMs with multiple Gaussian mixture models per state implemented using Cambridge's HTK tools [21]. The HMM phone recognizer uses context independent (CI) monophone models, while the word recognizer uses tied-state tri-phone acoustic models and a tri-gram language model. TIMIT dataset originally has 61 phonetic annotations, which are mapped to 45 phones based on recommendations from previous research. Details of this mapping can be found in [5] and [1] The recognizers have 5 states per phone model and 15 diagonal covariance Gaussians per state. The parameters of the state Gaussian mixture components were estimated by maximum likelihood estimation (MLE). For both recognizers, a 3-state silence model and a 1-state short pause model which shares the middle state of the silence model are used. These settings were used for easy side-by-side comparison.

Two sets of natural speech data are selected for the recognition experiments. The first set is the identical 100 TIMIT sentences recorded by the 10 RP speakers of SCRIBE, which

provides a well-matched training and testing condition, denoted as SCRIBE-TIMIT. The second set is the commonly used TIMIT database. The available 2342 sentences (SA, SI and SX) produced by 497 speakers (155 female/342 male) from six dialect regions (DR1 to DR6) are used for recognition experiments to test the accuracy of API on unfamiliar conditions. For ASR experiments, 80% of the dataset is used for training, and the other 20% for testing. The recognition performances in noisy conditions are also reported.

IV. Results and Discussion

A. Effect of the Interpolation Weight

Since the interpolation weight γ^i determines the contribution of individual feature streams, we have, therefore, evaluated a range of weights for the rescoring scheme on the well-matched training-testing set, i.e., the 100 SCRIBE-TIMIT sentences. Fig. 3 plots the phone recognition results, i.e., the phone error rate (PER), as a function of γ^1 assigned to the API baseline, for clean speech and for speech contaminated by white Gaussian noise (WGN) at $15dB$ signal to noise ratio (SNR). So during the rescoring process, the system is purely APF at $\gamma = 1$, and purely HMM (with MFCCs front end) at $\gamma = 0$. For clean testing, the best performance is obtained with a weight of 0.3, but for noisy testing, $SNR = 10dB$, the best result lies between 0.4 and 0.5. We, therefore, assigned an interpolation weight of 0.35 to the APF baseline for all other ASR experiments, which is slightly higher than the 0.2 suggested in [14].

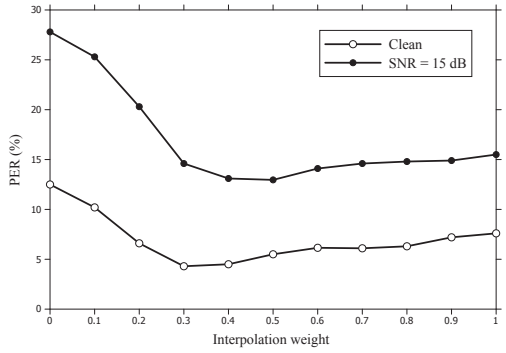


Fig. 3. Phone error rate of the combined MFCC/APF recognizer as a function of the interpolation weight assigned to the APF baseline.

B. Recognition Accuracy and Robustness

The results for the phone recognition performance as a function of SNR on the TIMIT database of the HMM baseline, the RNN based API module, and the combined system are shown in Fig. 4. Phone recognition accuracy are measured against SNR from 0 to 25 dB (clean speech) using added WGN.

Both the API module and the rescored module outperformed the HMM baseline in noisy and clean conditions. The API module gave 3.38% PER improvement on clean speech and an average 8.55% on noisy speech over plain MFCC. The

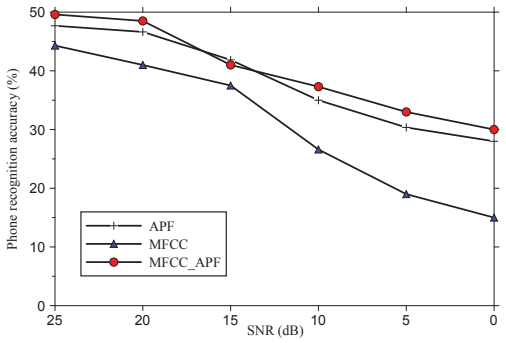


Fig. 4. Phone recognition accuracy in noise using MFCC baseline only, APF baseline only, and the combined system (MFCC_APF).

largest improvement occurred with the rescored system, which gave 5.30% and an average 10.14% relative PER improvement for clean and noisy speech respectively. It appears that not only does the API module lower the probability of wrong phone sequences, it also supplements the HMM baseline with additional sources of information, which is evidenced by the salience of the rescored system, i.e., robustness against noise contamination. It thus confirms our initial hypothesis that the phonetic information are better preserved in the API module. In other words, the phone states are better represented using the multi-dimensional features in the acoustic models. Thus we further used the hybrid system to carry out word recognition tasks to investigate the benefits of the API method. However, large vocabulary word recognition is generally more dependent on linguistic structures of the input speech, i.e., the language model. Thus the HMM context dependent (CD) tri-phone model and a simple bi-gram language model are applied for the word recognition on the smaller SCRIBE-TIMIT dataset. The API system is the same as used for phone recognition tasks. Table I summarizes the word error rate (WER) of the MLE-based HMM baseline and the rescored system. The largest WER reduction is obtained by the rescored module over the plain MFCC baseline 6.33%, which is observed for the CI uni-gram shown in the first two rows in the left column in Table I. And the smallest improvement is observed in the well constrained CD bi-gram HMM recognizer, 0.7% by the knowledge module. Average WER reduction using the API model is 3.35%.

TABLE I
WER on the SCRIBE-TIMIT word recognition task using the MFCC baseline and the MFCC baseline rescored with APF (MFCC_APF).

	CI monophone (%)	CD tri-phone (%)
MFCC (uni-gram)	22.35	14.54
MFCC_APF (uni-gram)	16.02	10.36
MFCC (bi-gram)	17.79	9.75
MFCC_APF (bi-gram)	15.62	9.05

The results on PER and WER generally agree with the findings reported in previous studies. In [5] and [14], a set

of phoneme classifiers were proposed to extract articulatory features in hybrid HMM/ANN recognition systems, where PER reduction of 5.92% on the TIMIT SA (dialectic) sentences, and WER reduction of 3.1% on the German Verbmobil corpus were achieved for clean speech. Similarly Kirchhoff also obtained an average 5.4% PER reduction for noisy testings on continuous digits recognition. In the literature, word recognition accuracy as high as 73.6% has been reported on large vocabulary conversational speech [5]. Our system is more competitive in terms of error rate reduction. The result itself is insignificant in terms of short-term ASR advances, nonetheless it clearly implies that the presence of articulatory knowledge in the front end feature space actually opens up new grounds for future deployment. One of the drawbacks is the reliability of the synthesizer, where the intrinsic physiological properties of the speaker are possibly over-generalized, e.g., the elasticity matrix [K] for the tongue muscles usually requires more sophisticated modeling methods as suggested in [6] and [18]. This will be taken into consideration in our future studies.

V. Conclusion

In this paper, a novel RNN based system is presented to retrieve and utilize articulatory-phonetic information from multiple knowledge sources for improved speech recognition. In particular, articulatory gestures are inferred using a explicitly controllable bio-mechanical speech synthesizer and a set of TIMIT sentences with "received pronunciation" annotated by professional linguists to deal with the incompetence of English pronunciation modeling in conventional ASR systems. The RNN based API model is monitored by a specialized learning algorithm which assimilates the human experiences of speech acquisition as a way to address the "many to one" problem in speech production and to exploit the benefits of categorical speech perception. It is clear from the phone recognition results that the API model are much more competent in modeling highly variant phonetic features than the widely used HMM based recognizers with the MFCC front end. Furthermore, the API system is globally tuned with the same set of speech data in an off-line learning mode, and then used for all online testings. Thus portability and computational efficiency are another two salient properties of the proposed framework. These novel qualities are further evidenced by the fact that the proposed model obtains consistent recognition improvements for the word recognition tasks when compared to the HMM recognizer with tri-phone and bi-gram language models. Yet it is perceivable that the extracted articulatory features might introduce redundancies when combined with acoustic features. This can be dealt with by applying dimension reduction procedures such as principle component analysis (PCA) or discrete cosine transform (DCT) on the input features to produce more compact articulatory-phonetic representations. These aspects are included in our ongoing research.

References

[1] J. H. Martin and D. Jurasfsky, *Speech and language processing*. Prentice Hall, 2008.

[2] K. Stevens, "Toward a model of lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, vol. 111(4), pp. 1872 – 1891, 2002.

[3] B. Kroger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793 – 809, 2009.

[4] W. J. Levelt, "Models of word production," *Trends in cognitive sciences*, vol. 3, no. 6, pp. 223 – 232, 1999.

[5] S. Siniscalchi and C. H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, no. 11, pp. 1139 – 1153, 2009.

[6] P. Birkholz, B. Kroger, and C. Neuschaefer Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2011.

[7] R. Mottonen and K. E. Watkins, "Motor representations of articulators contribute to categorical perception of speech sounds," *J. Neurosci.*, vol. 29, pp. 9819 – 9825, 2009.

[8] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679 – 691, may. 2005.

[9] W. Jeon and B.-H. Juang, "Speech analysis in a model of the central auditory system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1802 – 1817, aug. 2007.

[10] A. Kielar, L. Milman, B. Bonakdarpour, and C. Thompson, "Neural correlates of covert and overt production of tense and agreement morphology: Evidence from fmri," *Journal of Neurolinguistics*, vol. 24, no. 2, pp. 183 – 201, 2011.

[11] A. Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 833 – 841, nov 2001.

[12] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723 – 742, 2007.

[13] N. Strom, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," *The Free Speech Journal*, vol. 5, 1997.

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

[15] P. Boersma, "Functional phonology: Formalizing the interactions between articulatory and perceptual drives," Ph.D. dissertation, University of Amsterdam, 1998.

[16] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, pp. 1070 – 1082., 1973.

[17] J. Hieronymus, M. Alexander, C. Bennett, I. Cohen, D. Davies, D. Dalby, J. Laver, W. Barry, A. Fourcin, and J. Wells, "Proposed speech segmentation criteria for the scribe project," SCRIBE Project Report, Tech. Rep., 1990.

[18] P. Perrier and D. J. Ostry, "The equilibrium point hypothesis and its application to speech motor control," *Journal of Speech and Hearing Research*, vol. 39, pp. 365 – 378, 1996.

[19] M. Slaney, *Auditory toolbox*, 1998.

[20] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech recognition and utterance verification based on a generalized confidence score," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 821 – 832, nov 2001.

[21] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.