Noise Robustness of Tract Variables and their Application to Speech Recognition

Vikramjit Mitra\textsuperscript{1}, Hosung Nam\textsuperscript{2}, Carol Espy-Wilson\textsuperscript{1}, Elliot Saltzman\textsuperscript{3}, Louis Goldstein\textsuperscript{4}

\textsuperscript{1}Department of Electrical and Computer Engineering, University of Maryland, USA
\textsuperscript{2}Haskins Laboratories, New Haven, USA
\textsuperscript{3}Department of Physical Therapy and Athletic Training, Boston University, USA
\textsuperscript{4}Department of Linguistics, University of Southern California, USA

vmitra@umd.edu, nam@haskins.yale.edu, espy@umd.edu, esaltz@bu.edu, louisgol@usc.edu

Abstract

This paper analyzes the noise robustness of vocal tract constriction variable estimation and investigates their role for noise robust speech recognition. We implemented a simple direct inverse model using a feed-forward artificial neural network to estimate vocal tract variables (TVs) from the speech signal. Initially, we trained the model on clean synthetic speech and then test the noise robustness of the model on noise-corrupted speech. The training corpus was obtained from the Task Dynamics Application model (TADA [1]), which generated the synthetic speech as well as their corresponding TVs. Eight different vocal tract constriction variables consisting of five constriction degree variables (lip aperture [LA], tongue body [TBCD], tongue tip [TTCD], velum [VEL], and glottis [GLO]); three constriction location variables (lip protrusion [LP], tongue tip [TTCL], tongue body [TBCL]) were considered in this study. We also explored using a modified phase opponency (MPO) [2] speech enhancement technique as the preprocessor for TV estimation to observe its effect upon noise robustness. Kalman smoothing was applied to the estimated TVs to reduce the estimation noise. Finally, the TV estimation module was tested using a naturally-produced speech that is contaminated with noise at different signal-to-noise ratios. The estimated TVs from the natural speech corpus are then used in conjunction with the baseline features to perform automatic speech recognition (ASR) experiments. Results show an average 22% and 21% improvement, relative to the baseline, on ASR performance using the Aurora-2 dataset with car and subway noise, respectively. The TVs in these experiments are estimated from the MPO-enhanced speech.

Index Terms: Gestural phonology, Neural Networks, noise robust speech recognition, speech inversion, Kalman smoothing.

1. Introduction

Speech inversion or acoustic-to-articulatory inversion of speech has been a widely researched topic in the last 35 years. Various factors have stimulated research in this area, the most prominent being the failure of current state-of-the-art phone-based automatic speech recognition (ASR) systems during spontaneous or casual speech. There are several strong arguments for considering articulatory information in ASR systems. First, it may help to model coarticulation and reduction in a more systematic way. Second according to Kirchhoff et. al. [3], articulatory information is more robust to speaker variation and signal distortion. Finally, in [3, 4], it was demonstrated that articulatory features can significantly improve the performance of an ASR system in noisy environments. They also demonstrated that this effectiveness increases with a decrease in the signal-to-noise ratio (SNR).

A major motivation for the current study is our belief that an overlapping gesture-based architecture inspired by articulatory phonology can overcome many of the intrinsic limitations of phone-based units in articulating co-articulation. In articulatory phonology, the articulatory constriction gesture is an invariant action unit that can be decomposed into a constellation of articulatory gestures, and hence the ASR performance.

This study does not exclusively deal with the non-uniqueness of speech inversion. It has been observed [11, 12] that a static solution for speech inversion suffers largely from the non-uniqueness issue. It has been proposed that incorporating dynamic information about the acoustic data may help to disambiguate points of instantaneous one-to-

\textsuperscript{1} The first two authors have contributed equally to this paper and are in alphabetic order.
many mappings although this process would increase the difficulty of the non-linear mapping problem [11, 12]. In our approach, we use temporal context information to reduce the severity of non-uniqueness. We have implemented a simple 3-hidden layer feedforward artificial neural network (ANN) that estimates the TVs given acoustic features with contextual information as input. The configuration of the network as well as the contextual information of the acoustic feature is based upon our previous experimental results [13]. The network is trained with clean synthetic data obtained from TADA and it is evaluated against noise-corrupted (car and subway noise) versions of the synthetic data set at various SNRs. In a parallel study, the noise-corrupted speech was enhanced by an MPO-APP based speech enhancement algorithm [14] and then the enhanced speech was used to estimate the TVs. The network is then used to generate the TVs both with and without initial speech enhancement for the car and subway noise sections of the Aurora-2 [15] dataset and the estimated TVs were used in conjunction with the baseline features to observe if the use of TVs helps in improving the accuracy of the ASR system in noisy cases.

2. The Data

To develop the module for estimating the TVs, we randomly selected 960 utterances from the Aurora-2 training data. The digit sequence, mean pitch and gender information from each utterance were input to TADA. TADA then generated the TVs (see Table 1), vocal tract area function and formant information for each utterance. Finally, the pitch and formant information was input to HLSyn which generated the synthetic waveforms. Out of the 960 files, 690 of them were used for training the ANN model and the remaining were used for testing. The testing files were further corrupted with subway and car noise at the same SNR levels as used in Aurora-2.

Table 1. Constriction organ, vocal tract variables & involved model articulators

<table>
<thead>
<tr>
<th>Constriction organ</th>
<th>Tract Variables (TVs)</th>
<th>Articulators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lip</td>
<td>Lip Aperture (LA)</td>
<td>Upper/lower lip, jaw</td>
</tr>
<tr>
<td></td>
<td>Lip Protrusion (LP)</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip</td>
<td>Tongue tip constriction degree (TTCD)</td>
<td>Tongue body, tip, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue tip constriction location (TTCL)</td>
<td></td>
</tr>
<tr>
<td>Tongue Body</td>
<td>Tongue body constriction degree (TBCD)</td>
<td>Tongue body, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue body constriction location (TBCL)</td>
<td></td>
</tr>
<tr>
<td>Velum</td>
<td>Velum (VEL)</td>
<td>Velum</td>
</tr>
<tr>
<td>Glottis</td>
<td>Glottis (GLO)</td>
<td>Glottis</td>
</tr>
</tbody>
</table>

In the second part of this study, the TV estimation module developed from synthetic data was used to generate the TVs for all of the naturally-produced clean utterances in the Aurora-2 training data. In addition, TVs were estimated for the test sets that were corrupted with car and subway noise.

3. The TV estimator and its noise robustness

Several researchers [16, 17, 18] have studied the potential of ANNs for speech inversion. Compared to other architectures, ANNs have a lower computational cost both in terms of memory and execution speed [18]. In ANN-based direct inverse models, instantaneous non-uniqueness has been dealt with by using a suitable contextual window to exploit dynamic information in the input space. In this research, we have used 13 Mel-frequency Cepstral Coefficients (MFCC) as the acoustic feature, with an analysis window of 10ms and a frame advance of 5ms. The MFCCs were mean subtracted and normalized by five times the standard deviation and were scaled to be in the range of ~0.95 to +0.95. Based on our observation in [13], we selected a contextual window where 8 frames are selected before and after the current frame with a frame shift of 2 (time shift of 10 msec) between the frames giving rise to a vector of size (2*(8+1)) * 13 = 221. The TVs were similarly normalized and scaled. The feed-forward network specification is also based on [13], where the number of processing elements (i.e., the nodes) for the three layers was selected to be 150-100-150. The ANN was trained with a back-propagation algorithm with scaled conjugate gradient as the optimization rule and was trained for 5000 epochs. A tan-sigmoid function was used as the excitation for all of the layers.

The estimation of the TVs from the synthetic data with noise was done in four different ways: (1) the noisy files were used as-is to estimate the TVs (“No Process TV”), (2) the noisy files were used as-is to estimate the TVs and the estimated TVs were smoothed after estimation by a Kalman smoother (“No Process + Kalman TV”), (3) the noisy files were first enhanced and then TV estimation was performed, (“MPOAPP TV”), and (4) the noisy files were enhanced as in (3), and the estimated TVs were processed with Kalman smoothing (“MPOAPP + Kalman TV”). Figure 1 shows the plot of the root mean square error (RMSE) and Figure 2 shows the correlation of the estimated tract variable (TBCD) with respect to the ground truth decreases. However at a given SNR, the correlation is higher (RMSE is lower) for the TV estimated after MPOAPP enhancement than that without any enhancement. Thus, enhancement improves TV estimation. Also, the Kalman smoothing is found to help in the process of the estimation as compared to not using it at all. Figure 3 shows the estimated TVs for the synthetic utterance ‘two five’, using no processing with Kalman smoothing (clean

![Figure 1. RMSE of estimated TBCD TV at different SNRs for subway noise.](image-url)
condition) and MPOAPP enhancement with Kalman smoothing (15dB subway noise).

Figure 2. Correlation of estimated TBCD TV at different SNRs for subway noise.

Figure 3. The spectrogram of synthetic utterance ‘two five’, along with the ground truth and estimated (at clean and 15dB subway noise) TVs for GLO, TBCL, TTCL and TTCD.

The TV estimation module, which was trained with clean synthetic data, was then used for natural speech. Since there is no known groundtruth for the TVs of this database, the RMSE as well as the correlation cannot be computed directly. However, we can compare the estimated TVs from different noise levels to those in clean and generate the relative RMSE and correlation. Figure 4 shows that the relative RMSE (normalized) increases and the correlation decreases as the noise level gets higher.

Figure 4. Average Correlation and normalized RMSE of the estimated “No Process TVs” for car noise at different SNRs (the RMSE and correlation information are measured with respect to the estimated “No Process TVs” at clean condition).

The TV estimation module, which was trained with clean synthetic data, was then used for natural speech. Since there is no known groundtruth for the TVs of this database, the RMSE as well as the correlation cannot be computed directly. However, we can compare the estimated TVs from different noise levels to those in clean and generate the relative RMSE and correlation. Figure 4 shows that the relative RMSE (normalized) increases and the correlation decreases as the noise level gets higher.

Figure 5 shows the estimated TVs for GLO, TBCL, TTCL and TTCD for the utterance ‘two-five’ from the Aurora-2 database. The TV estimation is performed after MPOAPP enhancement followed by Kalman smoothing.

4. ASR experiments

The estimated TVs were concatenated with the 39 dimensional MFCCs for performing the word recognition task of the Aurora-2 database. The subway and the car noise subset of test set A of the Aurora-2 were used. The TVs were estimated for the training set as well; and the ASR experiment was based on training on clean and testing on multi-SNR noisy data. The recognizer trained with the Aurora-2 data-set was used for the ASR experiment. The backend uses eleven whole word HMMs, each with 16 states and each state having three Gaussian mixtures. Two pause models for “‘sil” and “‘sp”, were used, where the “‘sil” model has three states and each state has six mixtures. The “‘sp” model has a single state. The training and testing scripts provided with the Aurora-2 database were used with minor modifications accommodating for the testing only on car and subway noise as well as the change in the feature dimension because of augmentation with the TVs. Figures 6 and 7 shows the word recognition accuracy for the car and subway noise using TV estimation with and without MPOAPP enhancement, but both with Kalman smoothing post-processing. From the plots it is evident that the TVs helped to improve the recognition rates over the baseline. The improvement is found to be even better for the case where the TVs were estimated after MPOAPP enhancement of the noisy data.

Figure 5. The spectrogram of natural utterance ‘two five’ from Aurora-2, along with the reconstructed TVs (at clean and 15dB car noise) for GLO, TBCL, TTCL and TTCD.
Two main tasks were performed in this research. The first task involved evaluation of the noise robustness of the TV estimation procedure. In the second task, the estimated TVs were used in conjunction with the MFCCs to observe if they aid in improving ASR noise robustness. It was observed that a decrease in SNR resulted in an increase in the RMSE of the estimated TVs and a simultaneous decrease in correlation. A speech enhancement algorithm was used to clean the noisy speech and the cleaned speech in turn was used for the estimation of the TVs. It was observed that using such a speech enhancement technique helped to reduce the RMSE and simultaneously improve the correlation at a given SNR.

In this paper we have used a direct inverse model to estimate the TVs; moreover, the inverse model was trained with a significantly smaller amount of data than that available in the Aurora-2 training database. The results in this paper are a preliminary study showing that tract variables can contribute more positively toward improving ASR noise robustness.

5. Conclusion and Discussion

Two main tasks were performed in this research. The first task involved evaluation of the noise robustness of the TV estimation procedure. In the second task, the estimated TVs were used in conjunction with the MFCCs to observe if they aid in improving ASR noise robustness. It was observed that a decrease in SNR resulted in an increase in the RMSE of the estimated TVs and a simultaneous decrease in correlation. A speech enhancement algorithm was used to clean the noisy speech and the cleaned speech in turn was used for the estimation of the TVs. It was observed that using such a speech enhancement technique helped to reduce the RMSE and simultaneously improve the correlation at a given SNR. This trend was fairly consistent as observed from Figures 1 and 2. The ASR experiments show that the use of the TVs helped to improve the noise robustness. The TVs obtained from the MPOAPP enhanced speech helped to obtain better ASR performance than the one without any processing. These results suggest that the TVs, if estimated properly, can help to improve the noise robustness of ASR systems.

In this paper we have used a direct inverse model to estimate the TVs; moreover, the inverse model was trained with a significantly smaller amount of data than that available in the Aurora-2 training database. The results in this paper are a preliminary study showing that tract variables can contribute more positively toward improving ASR noise robustness. Future research should address using more training data for the TV estimator. Also using articulatory data to define a ground-truth for spontaneous speech TV estimation would greatly help the process of spontaneous speech TV estimation and would contribute more positively toward improving ASR noise robustness.

Acknowledgements
This research was supported by NSF Grant # IIS0703859, IIS-0703048 and IIS0703782.

6. References