Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion

Ganesh Sivaraman [1], *Vikramjit Mitra*[2], *Hosung Nam*[3], *Mark Tiede*[4], *Carol Espy-Wilson*[1]

[1] University of Maryland College Park, MD, [2]SRI International, Menlo Park, CA,
[3]Department of English Language and Literature, Korea University, Seoul, South Korea,
[4]Haskins Laboratories, New Haven, CT

{ganesa90, espy}@umd.edu, vmitra@speech.sri.com, nam@haskins.yale.edu,
tiede@haskins.yale.edu

Abstract

Speech inversion is a well-known ill-posed problem and addition of speaker differences typically makes it even harder. This paper investigates a vocal tract length normalization (VTLN) technique to transform the acoustic space of different speakers to a target speaker space such that speaker specific details are minimized. The speaker normalized features are then used to train a feed-forward neural network based acoustic-to-articulatory speech inversion system. The acoustic features are parameterized as time-contextualized mel-frequency cepstral coefficients and the articulatory features are represented by six tract-variable (TV) trajectories. Experiments are performed with ten speakers from the U. Wisc. X-ray microbeam database. Speaker dependent speech inversion systems are trained for each speaker as baselines to compare the performance of the speaker independent approach. For each target speaker, data from the remaining nine speakers are transformed using the proposed approach and the transformed features are used to train a speech inversion system. The performances of the individual systems are compared using the correlation between the estimated and the actual TVs on the target speaker's test set. Results show that the proposed speaker normalization approach provides a 7% absolute improvement in correlation as compared to the system where speaker normalization was not performed.

Index Terms: Acoustic to articulatory speech inversion, speaker normalization, Vocal Tract Length Normalization

1. Introduction

Speech inversion or acoustic-to-articulatory inversion of speech has been a widely researched topic in the last 40 years. Speech Inversion is the process of mapping the acoustic signal into articulatory parameters. If estimated accurately, articulatory information can be applied to speech accent conversion [1], speech therapy, language learning, and Automatic Speech Recognition (ASR) [2][3][4].

Real articulatory data is obtained from subjects using techniques like Electromagnetic Articulometry (EMA), X-ray microbeam, and real-time Magnetic Resonance Imaging (rt-MRI). However these techniques require sophisticated devices and are expensive and time consuming. Obtaining real articulatory data is not practically feasible for real world applications like ASR. Only the acoustic data is available from the speaker. Hence, it is essential to develop speech inversion systems that are speaker independent and can accurately estimate articulatory features for any unseen test speaker.

The mapping from acoustics to articulations is known to be highly non-linear and non-unique [5]. Adding speaker variability to the already challenging problem makes it even more difficult. Most research in speech inversion has been focused on developing accurate speaker dependent systems. Approaches like codebook search, feedforward neural networks, and Mixture Density Networks have been found to work well for speaker dependent speech inversion. There have been a few attempts to perform speaker independent speech inversion [6][7] which have been limited to two speakers from the MOCHA TIMIT dataset [8]. Hueber et al. [9] presents a Gaussian mixture regression based speaker adaptation scheme for a Gaussian Mixture Model (GMM) based speech inversion system. However, there has not to date been any effort in performing speaker adaptation for artificial neural network based speech inversion systems. This paper presents a Vocal Tract Length Normalization (VTLN) based approach to speaker adaptation for speech inversion. VTLN is a popular speaker adaptation technique in ASR which has so far not been applied to speech inversion.

The objective of this paper is to normalize acoustic data from multiple speakers towards the acoustic space of a target speaker. Unlike the usual adaptation of the test utterances towards the acoustic space of the training speaker, this approach aims to transform data from multiple speakers towards the acoustic space of a target test speaker to train a speech inversion system for the target speaker using the speaker normalized data. Diagonal covariance GMMs are trained for each speaker and a piecewise linear frequency warping similar to VTLN is performed to adapt the acoustic space of the training speakers towards that of the test speaker. More details of this adaptation procedure are provided in section 3.

The experiments in this paper are performed on a set of 10 speakers from the U. Wisconsin X-ray Microbeam (XRMB) database [10]. The articulatory features are represented by six tract-variable (TV) trajectories (described below). Using a leave-one-out methodology, separate experiments were performed for each speaker in which the acoustic features from the other 9 speakers were transformed using the VTLN approach. The transformed acoustic features were then used to train a speech inversion system. The performance of the system trained on VTLN adapted acoustic features was compared to the performance of speaker dependent systems. The performances of the individual systems were compared using the correlation between the estimated and the actual TVs on the target

speaker's test set. More details of the speech inversion system training and the experiments are provided in sections 4 and 5.

The results of the experiments showed that the VTLN based acoustic feature normalization improved the correlation score by 7% absolute over the case when no adaptation was performed. A detailed analysis of the results is presented in section 6.

This novel approach to speaker normalization for speech inversion provides interesting insights to this problem and opens several avenues for future work that are discussed in section 7 and 8 of the paper.

2. Dataset Description

The XRMB dataset was used for the experiments performed in this paper. The dataset consists of flesh point pellet trajectories of along with simultaneous audio recordings of continuous speech utterances for 46 different speakers. We converted the pellet trajectories to six Tract Variables (TVs) [11] using a geometric transformation procedure outlined in [12]. The six TVs were – Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL) and Tongue Tip Constriction Degree (TTCD). After discarding recordings containing mistracking and other errors due to conversion to TVs, the dataset consisted of 4 hours of speech and data from 46 speakers with unequal amounts of speech from each speaker. We selected 10 speakers from the XRMB dataset (5 males and 5 females) such that the amount of data from each speaker was roughly the same (around 6.5 to 8mins). For each speaker, we split the dataset into three sets – 80% for training, 10% for cross validation, and 10% for testing. All our experiments were performed using the data from these ten speakers.

3. Vocal tract length normalization approach

Any speaker adaptation scheme requires an acoustic model that approximates the acoustic space of the target speaker. We modeled the acoustic space of each speaker using unsupervised Gaussian Mixture Models. The acoustic features for the GMMs were 13 dimensional MFCCs computed with an analysis window of 20ms and a frame rate of 10ms, along with slope and accelerations. We trained 64 Gaussian components in an unsupervised manner on the 39 dimensional MFCC features. The diagonal covariance GMMs were trained iteratively by increasing the number of Gaussians from 2 to 64 by doubling the number of components in each stage. The GMM training routines were obtained from the MSR Identity Toolbox v1.0 [13]. Thus, such GMMs were trained for each of the 10 speakers chosen for the experiment.

3.1. Maximum likelihood approach to VTLN frequency warping

Vocal Tract Length Normalization (VTLN) aims to compensate for the effects of different vocal tract lengths by warping the frequency spectrum in the filterbank analysis before the computation of the cepstral coefficients. This warping was implemented by a simple piecewise linear warping function as shown in Figure 1. The warping factor α determines the nature of the warping function. The warping is implemented between the lower boundary of frequency analysis (LOFREQ) and the upper boundary of frequency analysis (HIFREQ). In order to adapt the acoustic features of speaker S_i to speaker S_j, a single warping factor α_{ij} is used for all utterances from speaker S_i. The warping factor α_{ij} is determined by a maximum likelihood approach as outlined below.

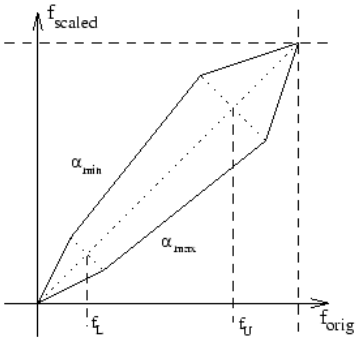

Figure 1: Frequency warping function implemented in HTK toolkit [14]

Let the GMM acoustic model for speaker S_j be λ_j, and the warped acoustic features for a signal frame of speaker S_i to the target speaker S_j be x_{ijt}. Then, the most likely warping factor α_{ij} is given by-

$$\alpha_{ij} = \underset{\alpha}{argmax} \sum_{t=1}^{N} \log(P(x_{ijt}|\lambda_j, \alpha)) \qquad (1)$$

In the above equation, $\sum_{t=1}^{N} \log(P(x_{ijt}|\lambda_j, \alpha))$ is the log likelihood of the transformed features of speaker S_i with respect to speaker S_j's acoustic model. The optimal α_{ij} is obtained by sweeping the value of α from 0.8 to 1.2 in steps of 0.025. Using the optimal α_{ij}, we compute the speaker adapted acoustic features for speaker S_i adapted to speaker S_j.

4. Speech Inversion system

Previous studies have demonstrated that Artificial Neural Networks (ANNs) can be used to reliably estimate the TV trajectories [14] from the speech signal. Once trained, ANNs require low computational resources compared to other methods in terms of both memory requirements and execution speed. In this paper, we trained speech inversion systems using a single hidden layer feed-forward neural network. Since only small amounts of data were available for each speaker, single hidden layer networks were chosen as the architecture. The inputs to the neural network were the 13 dimensional MFCCs contextualized with MFCC features from 8 frames on either side. Thus, the input dimension was 13x17 = 221. The outputs of the network were six dimensional TVs. We trained networks with 100, 200, 300, 400 and 500 nodes in the hidden layer and selected the best performing network based on performance on the test set. The outputs of the trained neural network were found to be noisy. The outputs were smoothed using a Kalman smoothing technique to obtain smooth TV estimates. Figure 2 shows the block diagram of our speech inversion system.

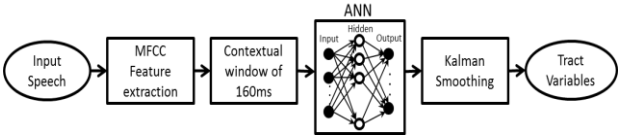

Figure 2: Block diagram of speech inversion system

5. Experiments

5.1. Speaker transformed datasets

Using the VTLN method described in Section 3, each speaker's data was transformed to each of the other 9 speakers' data. Thus, for each speaker, we have 10 sets of data – 1 from the speaker and other 9 transformed to the target speaker from the other 9 speakers using VTLN. The following figure shows the schematic of the transformation procedure for transforming data from speakers $S_b \ldots S_j$ to speaker S_a's acoustic space to create the transformed datasets $S_{ba} \ldots S_{ja}$. In this way, we created 90 transformed datasets tailored to each of the 10 speakers' acoustic spaces. Figure 3 shows the schematic of the procedure adopted to create the speaker transformed datasets.

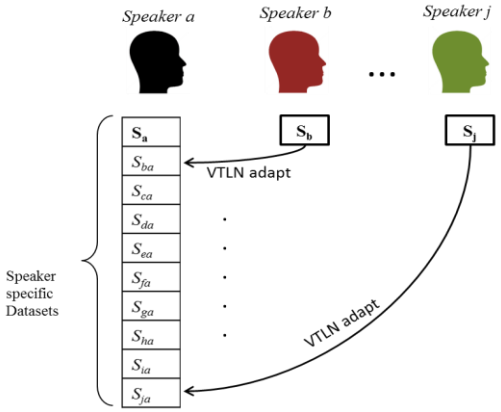

Figure 3: Schematic of speaker transformed datasets creation

5.2. Speech inversion systems trained

We trained four types of speech inversion systems for each speaker as described in Section 4. The following are the descriptions of the different inversion systems trained.

- **SD:** 10 Speaker Dependent (SD) speech inversion systems.
- **Sys1:** For each speaker, data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from $S_b \ldots S_j$ was randomly sampled to match the amount of data in S_a
- **Sys2:** For each speaker, VTLN transformed data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from $S_{ba} \ldots S_{ja}$ was randomly sampled to match the amount of data in S_a
- **Sys3:** For each speaker, data from the target speaker and the VTLN transformed data from other 9 speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, 10 such systems were trained. For example, for speaker 'a', data from S_a, $S_{ba} \ldots S_{ja}$ was randomly sampled to match the amount of data in Sa. The difference between System3 and System2 is that System3 has some of the target speaker's data in the training set.

In total, 40 speech inversion systems were trained. In the above described systems, the amount of training data for each system was kept the same in order to have a fair comparison with the SD system. However the transformed data available for each target speaker was about 10 times more because of the other 9 speakers' data put together. We created versions of Systems 1, 2, and 3 using all the transformed data. We call these systems Sys1_alldata, Sys2_ alldata, and Sys3_ alldata.

6. Results and Discussion

For each speaker, a test set containing 10% of the speaker's data was created which was kept separate from all the speech inversion training and VTLN procedure. Each of the systems SD, System1, 2, and 3 were evaluated on each speaker's test set. The Pearson product Moment Correlation (PPMC) was computed between the actual and estimated TVs. Table 1 shows the correlation results of all the speech inversion systems across all speakers. The numbers show correlation values averaged across all 6 TVs.

The correlation for LP tract variable is the least and that for TBCL is the highest. The performance of Sys1 is very poor compared to SD because the training dataset for this system consists of a small number of utterances from multiple speakers. Transforming the data from the other 9 speakers to the target speaker's acoustic space using the proposed VTLN approach provides an average of 7% absolute improvement in correlation over Sys1. The amount of improvement in correlation varies across all speakers. Some speakers like JW14 and JW24 show marginal or no improvement in the performance, whereas for JW31 we see a large 13% improvement. In order to see the influence of speaker specific training data on the performance, we created Sys3 which contained a part of the target speaker's training set data. The overall amount of training data for Sys3 was kept same as the amount of training data available for each target speaker. This provided an average of 3% improvement in correlation compared to Sys2. However, the correlations of Sys3 were still 13% below the average correlation of the SD systems. Figure 4 shows the plots of the estimated and actual TVs for a randomly selected test utterance from speaker JW26's test set.

Table 2 shows the correlation results for the speech inversion systems trained with all the available data from the other 9 speakers. These are the systems Sys1_alldata, Sys2_alldata and Sys3_alldata as described in section 5.2. We observe that the results are much better than those in table 1. The performance gain obtained by performing the VTLN adaptation is around 4% on an average above the correlation results of Sys1_alldata. It is interesting to observe that adding all the training data of the target speaker, as done in the training of Sys3_alldata provides a system that performs as well as the speaker dependent SD systems. This demonstrates that adding VTLN adapted data from multiple speakers does not degrade the performance of the speaker dependent systems.

Table 1: Correlation results of SD, Sys1, Sys2, and Sys3 for all speakers

Speech inversion System	Average amount of Training data (mins)	*Spk a* JW12 M	*Spk b* JW14 F	*Spk c* JW24 M	*Spk d* JW26 F	*Spk e* JW27 F	*Spk f* JW31 F	*Spk g* JW40 M	*Spk h* JW45 M	*Spk i* JW54 F	*Spk j* JW59 M	*Average*
SD	5.68	0.80	0.78	0.75	0.80	0.74	0.82	0.77	0.78	0.74	0.80	**0.78**
Sys1	5.68	0.61	0.62	0.58	0.58	0.50	0.45	0.51	0.57	0.52	0.56	**0.55**
Sys2	5.68	**0.66**	**0.61**	**0.59**	**0.69**	**0.58**	**0.58**	**0.61**	**0.60**	**0.64**	**0.65**	**0.62**
Sys3	5.68	0.68	0.64	0.66	0.70	0.61	0.62	0.64	0.65	0.65	0.66	**0.65**

Table 2: Correlation results of SD, Sys1_alldata, Sys2_alldata, and Sys3_alldata for all speakers

Speech Inversion System	Average amount of Training data (mins)	*Spk a* JW12 M	*Spk b* JW14 F	*Spk c* JW24 M	*Spk d* JW26 F	*Spk e* JW27 F	*Spk f* JW31 F	*Spk g* JW40 M	*Spk h* JW45 M	*Spk i* JW54 F	*Spk j* JW59 M	*Average*
SD	5.68	0.80	0.78	0.75	0.80	0.74	0.82	0.77	0.78	0.74	0.80	**0.78**
Sys1_alldata	51.13	0.71	0.71	0.70	0.69	0.65	0.60	0.63	0.69	0.69	0.68	**0.68**
Sys2_alldata	51.13	**0.74**	**0.73**	**0.73**	**0.78**	**0.69**	**0.69**	**0.70**	**0.72**	**0.70**	**0.74**	**0.72**
Sys3_alldata	56.81	0.80	0.78	0.77	0.81	0.74	0.79	0.76	0.77	0.76	0.79	**0.78**

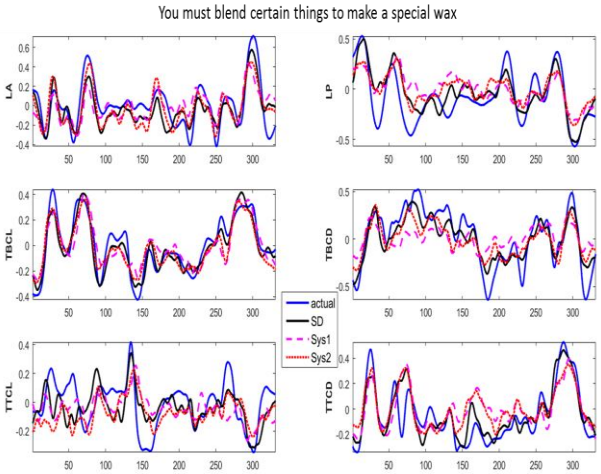

Figure 4: Plot of estimated and actual TVs for a test utterance from JW26's test set

7. Conclusions

Based on the results shown in tables 1 and 2, we can conclude that the amount of training data plays a great role in the accuracy of the speech inversion system. Even if the data is from multiple speakers, more data is always good.

The VTLN speaker adaptation normalizes multiple speakers' acoustic data to match a target speaker. VTLN provides an average of 7% absolute improvement of correlation (Sys1 to Sys2) on the speech inversion system trained on the 9 speakers' dataset. Adding a small amount of the target speaker's data in the training set improves the correlation further by 3% over Sys2. In spite of performing VTLN, the correlation performance of Sys2 trained on the transformed data is 16% poorer than the Speaker dependent system.

The systems trained with all data shows that having more training data from multiple speakers can make the speech inversion system better. The accuracy of Sys1_alldata is 10% poorer than SD due to the mismatch between the acoustic spaces of the training speakers and the test speakers. With the VTLN based transformation of the training data, the accuracy improves by 4%. This means our proposed adaptation technique helps reduce the mismatch between the acoustic spaces.

Adding all of the target speakers' training data along with the transformed data of the other 9 speakers' does not degrade the speaker dependent performance.

This approach of transforming the training data from multiple speakers to create multiple speaker adapted versions can be used to create a model selection based approach to speech inversion. In such a system we will have multiple speaker tuned models and then select the best matching model for a test utterance based on a maximum likelihood speaker matching approach.

The frequency warping technique used in this paper is a piecewise linear transformation. In future, we plan to explore some non-linear warping techniques. We also plan to explore other feature space acoustic model adaptation techniques like f-MLLR which are popular in the ASR community. The experiments in this paper show that data from multiple speakers can be normalized and combined to create better speech inversion systems. In future, we will try to combine data from different articulatory datasets.

8. Acknowledgements

This research was supported by NSF Grant # IIS-1162046.

9. References

[1] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7694–7698.

[2] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech*

Commun., vol. 37, no. 3–4, p. 319, 2002.

[3] V. Mitra, "Articulatory Information For Robust Speech Recognition." Ph.D. dissertation, University of Maryland, College Park, 2010.

[4] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3017–3021.

[5] C. Qin and M. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping.," *INTERSPEECH*, 2007.

[6] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Commun.*, vol. 66, pp. 1–16, Feb. 2015.

[7] A. Ji, "Speaker Independent Acoustic-to-Articulatory Inversion," Marquette University, 2014.

[8] A. A. Wrench, "A Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research.," *Phonus*. 18-Aug-2000.

[9] T. Hueber, L. Girin, X. Alameda-Pineda, and G. Bailly, "Speaker-Adaptive Acoustic-Articulatory Inversion Using Cascaded Gaussian Mixture Regression," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2246–2259, Dec. 2015.

[10] J. R. Westbury, "Microbeam Speech Production Database User''s Handbook," *IEEE Pers. Commun. - IEEE Pers. Commun.*, 1994.

[11] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *J. Acoust. Soc. Am.*, vol. 115, no. 5, p. 2430, May 2004.

[12] H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "A procedure for estimating gestural scores from speech acoustics.," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3980–9, Dec. 2012.

[13] S. Omid Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research," *Speech and Language Processing Technical Committee Newsletter*, Nov-2013.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book, version 3.4," 2006.