Multitask learning based multi-corpus acoustic-to-articulatory speech inversion

Ganesh Sivaraman[2] , Nadee Seneviratne[1],, Carol Espy-Wilson[1]
[1]University of Maryland College Park, MD, USA
[2]Pindrop Security Inc, Atlanta, GA, USA

1. INTRODUCTION

- Speech inversion (SI): a highly non-linear and non-unique mapping of the acoustic signal to the **articulatory dynamics**[1]
- Articulatory measurements are sensitive to
  - Measurement method and equipment
  - Anatomy of speakers
  - Sensor placement
- Most previous studies limited to single corpus studies.
- We propose to **generalize the SI system by using a multi-task learning model to develop a multi-corpus SI system.**
- All articulatory data are represented as Tract Variable trajectories which are reasonably speaker invariant.

2. DATASETS DESCRIPTION

2.1 X-Ray Microbeam (XRMB) dataset[2]

- Naturally spoken utterances and XRMB cinematography of the mid-sagittal plane of the vocal tract using pellets placed at points along the vocal tract.

2.2 Electromagnetic Articulometry (EMA)-IEEE dataset[4]

- Recordings of subjects reciting 720 phonetically balanced IEEE sentences at normal and fast production rates (using a 5-D EMA system).
- 9 TVs - LA, LP, Jaw Angle (JA), TTCL, TTCD, Tongue Middle Constriction Location (TMCL), Tongue Middle Constriction Degree (TMCD), TBCL and TBCD.

2.3 Multichannel Articulatory (MOCHA) - TIMIT dataset[5]

- Speech data and EMA data recorded simultaneously for subjects speaking British English.

Table 1: Articulatory datasets description

Dataset	# Subjects	Hours of Data	# TVs	TVs
XRMB	21 M, 25 F	4	6	LA, LP, TBCL, TBCD, TTCL, TTCD
EMA-IEEE	4 M, 4 F	7.05	9	LA, LP, JA, TTCL, TTCD, TMCL, TMCD, TBCL, TBCD
TIMIT	1 M, 1 F	1.01	9	LA, LP, JA, TTCL, TTCD, TMCL, TMCD, TBCL, TBCD

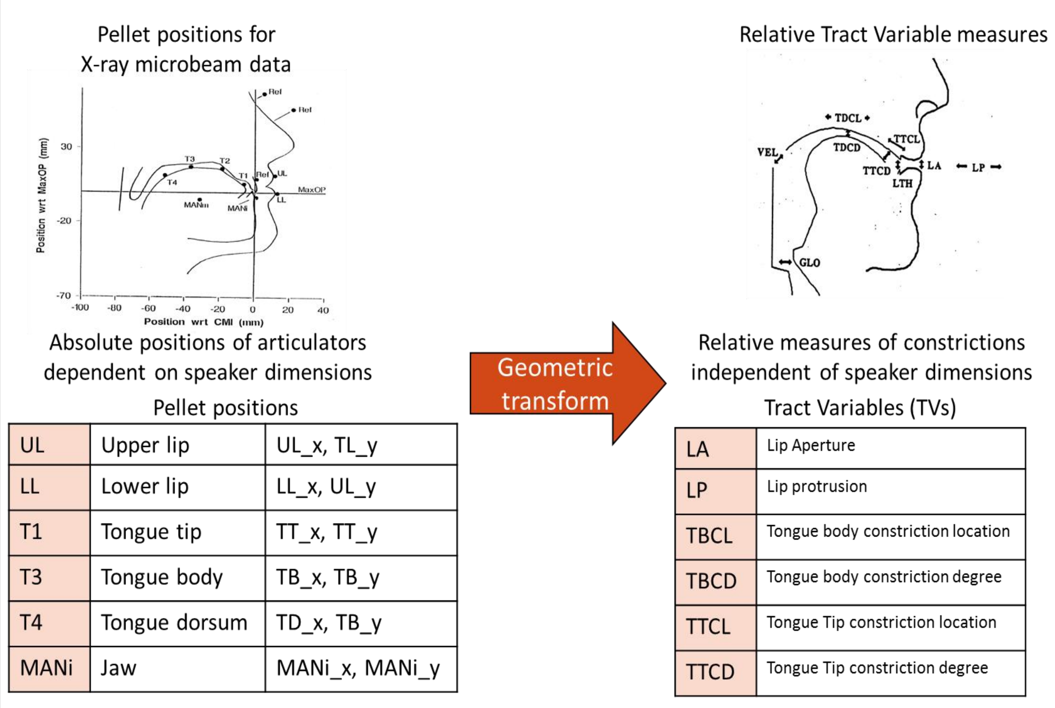

Figure 1: Schematic of transformation of XRMB database from pellets to TV trajectories [3]

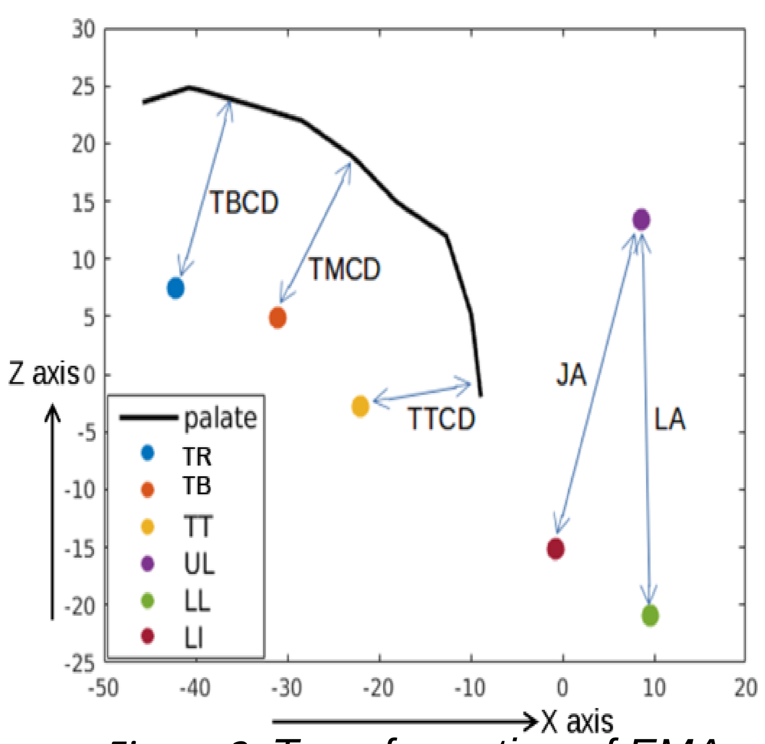

Figure 2: Transformation of EMA sensor positions to TVs

3. METHODOLOGY

- Input Feature Vector: **Contextualized MFCCs** (17 frames x 13 coefficients), z-normalized per speaker.
- **Feedforward Neural Network** was trained to learn three different sets of TVs corresponding to speech samples in the three databases (three tasks).
- The **hidden layers (5) of the model are shared** by these three output tasks.
- The three tasks of estimating TVs for XRMB, EMA-IEEE, and MOCHA-TIMIT speech utterances had **6, 9, and 9 output nodes** respectively.
- Single corpus SI systems were trained for all the 3 datasets for comparison.
- Pearson correlation of cross-corpus TV estimates was computed to evaluate the cross-corpus performance and generalization of the system.

Acknowledgements: *This work was made possible by National Science Foundation grants numbered IIS1162525, IIS1162046, and BCS1436600 and a hardware grant from NVIDIA.*

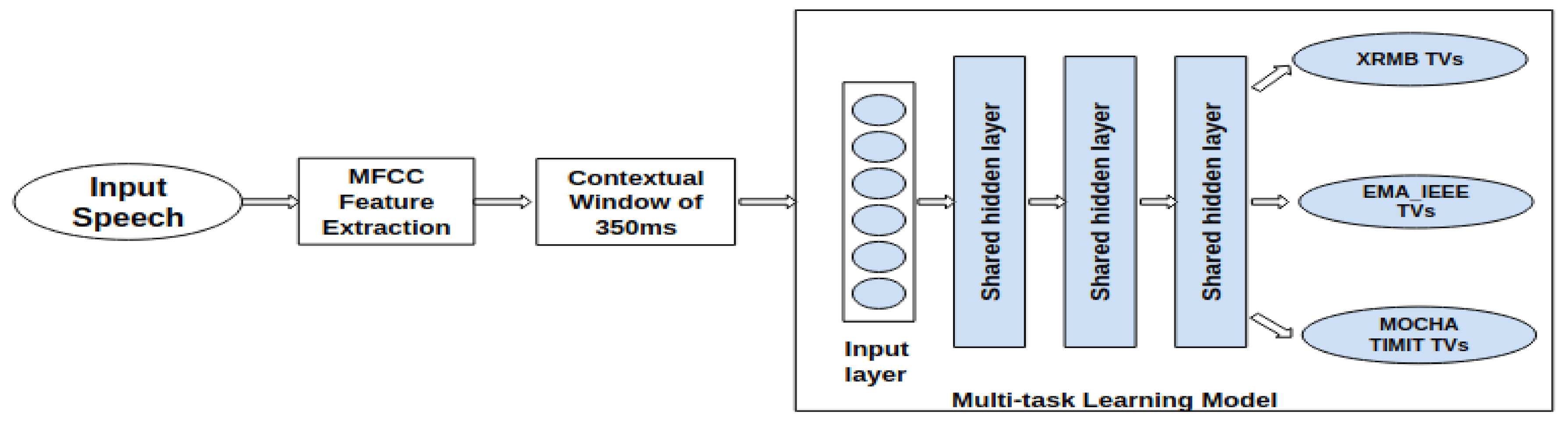

Figure 3: Block diagram of the multi-corpus speech inversion system

4. EXPERIMENTS & RESULTS

Table 2: Average correlations of TVs estimated by single-corpus SI systems for the best performing models (baseline)

Dataset	Model Architecture	Validation Set Average Corr.
XRMB	5 hidden layers, 512 nodes each	**0.789**
EMA-IEEE	5 hidden layers, 1024 nodes each	**0.826**
MOCHA-TIMIT	5 hidden layers, 1024 nodes each	**0.730**

Table 3: Cross correlations of TVs of test samples evaluated on best performing single-corpus models

Test set	Best Model - XRMB	Best Model - EMA-IEEE	Best Model - TIMIT
XRMB	**0.779**	**0.543**	**0.460**
EMA-IEEE	**0.453**	**0.821**	**0.540**
TIMIT	**0.475**	**0.608**	**0.735**

Table 4: Cross correlations of TVs of test samples evaluated on multi-corpus model

Test set	XRMB Output	EMA-IEEE Output	TIMIT Output
XRMB	**0.761 (-2.3%)**	**0.581 (6.9%)**	**0.596 (29.5%)**
EMA-IEEE	**0.576 (27.1%)**	**0.812 (-1.1%)**	**0.724 (34.2%)**
TIMIT	**0.576 (21.3%)**	**0.692 (13.9%)**	**0.781 (6.3%)**

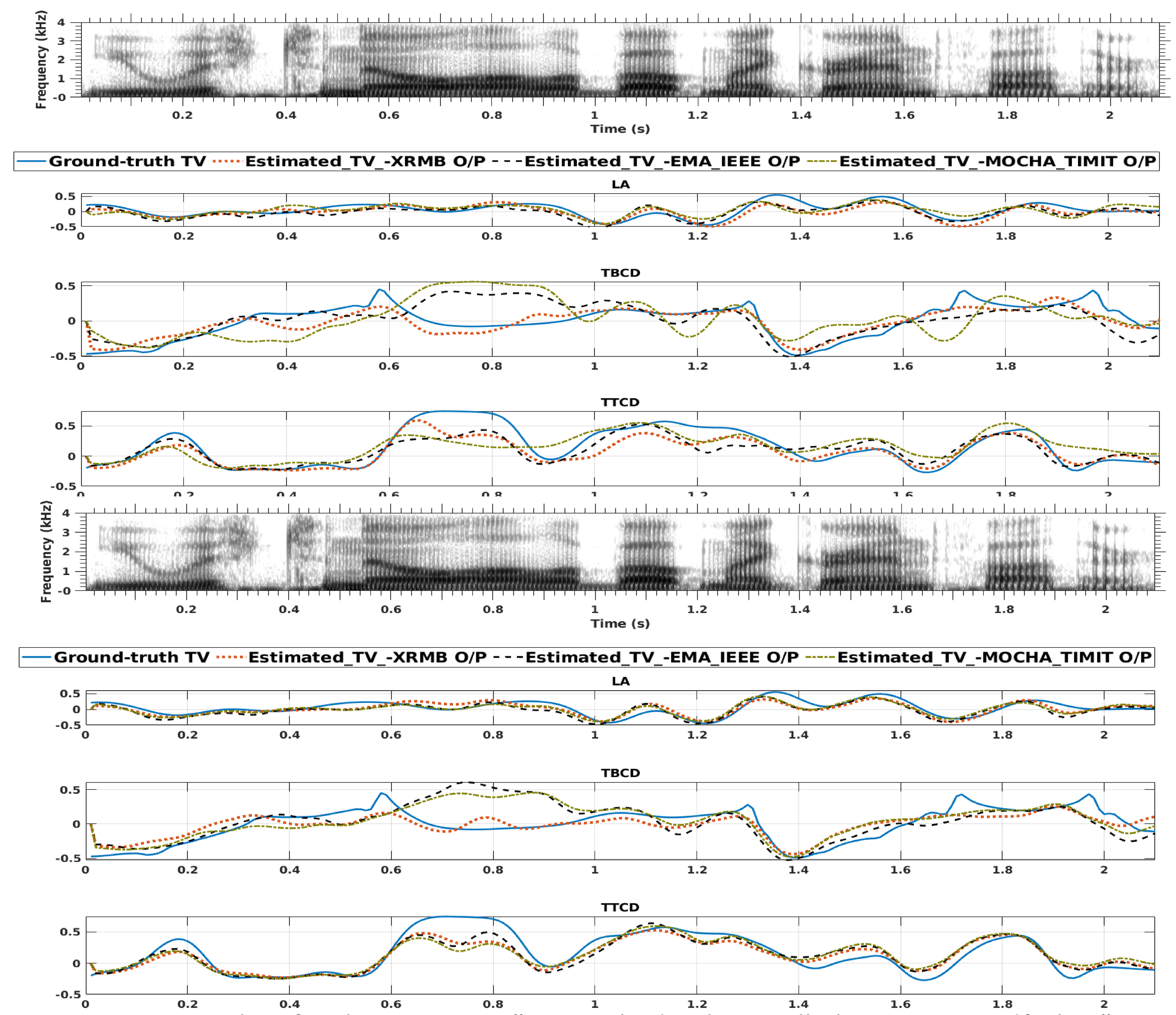

Figure 5: TV plots for the utterance "You wished to know all about my grandfather" estimated multi-corpus joint model.

5. CONCLUSION

- Cross corpus correlations of estimated TVs increased when using multi-corpus SI system.
- Minimal degradation in performance for the matched corpus test case.
- Proposed multi-corpus SI system perform better in generalizing articulatory dynamics of speech samples in multiple databases.

6. REFERENCES

[1] C. Qin and M. Carreira-Perpiñán, "An empirical investigation of the non-uniqueness in the acoustic-to-articulatory mapping." INTERSPEECH, 2007.
[2] J. R. Westbury, "Speech Production Database User ' S Handbook," IEEE Personal Communications -, vol. 0, no. June, 1994.
[3] V. Mitra, "Articulatory Information For Robust Speech Recognition." Ph.D. dissertation, University of Maryland, College Park, 2010.
[4] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," The Journal of the Acoustical Society of America, vol. 141, no. 5, pp. 3580–3580, 2017.
[5] A. A. Wrench, "A Multichannel Articulatory Database and its Application for Automatic Speech Recognition," Proceedings of 5th Seminar of Speech Production, pp. 305–308, 2000.