Acoustic-to-Articulatory Speech Inversion with Multi-task Learning

Yashish M. Siriwardena[1], Ganesh Sivaraman[2], Carol Espy-Wilson[1]

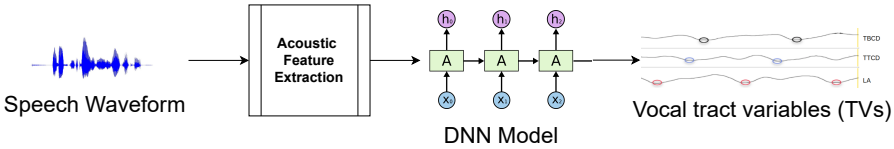
[1]University of Maryland College park, MD, USA
[2]Pindrop, GA, USA

September 3, 2022

Outline

Outline

Acoustic-to-Articulatory Speech Inversion (SI)



- The inverse problem of determining the trajectories of the movement of speech articulators from the speech signal
- Resulting time varying trajectories are called vocal tract variables (TVs)

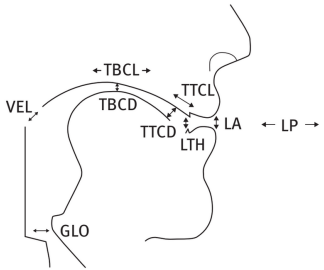
Why SI systems ?

- To better understand the speech production process
- To improve speech applications like ASR, speech synthesis, speech therapy and mental health assessment

Vocal Tract Variables (TVs)

- Based on Articulatory Phonology (AP) (Browman and Goldstein, 1992) which characterizes the kinematic state of articulatory constrictors by its corresponding degree and location coordinates

Constrictors	**Vocal tract variables (TVs)**
Lip	Lip Aperture (LA) Lip Protrusion(LP)
Tongue Tip	Tongue tip constriction degree (TTCD) Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD) Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)



Previous Work with Developing SI Systems

- Attempts at recovering articulatory movements from the continuous speech signal have a long history (Papcun et al., 1992)
- Similar acoustic consequences can result in from completely different articulatory configurations : **One-to-many mapping** (Maeda et al, 1990, Gunther et al. 1999)
- Previous work with codebook search (Ghosh and Narayanan, 2010), feed-forward neural networks and Mixture Density Networks (Richmond, 2006) focused on **Speaker-dependent SI systems**
- The work of Mitra et al.(2009) and Sivaraman et al. (2019) on developing **speaker-independent SI systems** uses simple feed-forward neural network architectures
- Recent work with BiLSTM models (Illa and Ghosh, 2018)(Illa and Ghosh, 2019b) and transformer models (Udupa et al., 2021) has focused on **Speaker-adaptation methods**

Motivation for our work

- Deep neural network (DNN) based models propelling the development of SI systems to new heights. (Bidirectional LSTMs (BiLSTMS), CNNBiLSTMs, Temporal Convolutional Networks (TCN) and transformer models)
- Leveraging on phoneme information to improve the SI task without explicitely using the phoneme transcriptions at inference
- Multi Task Learning (MTL) being a useful tool which can reuse the existing knowledge and reduce the cost of collecting larger challenging datasets (e.g. articulatory datasets)
- The success of MTL with the use of more data from different learning tasks compared to learning a single task (Zhang et al. 2021)

Outline

What is Multi-task Learning ?

- Formally presented by Caruana et al. (1997) as an inductive transfer mechanism with the principle goal of improving generalizability
- Improves generalizability by leveraging domain-specific information of training data which can be used in related tasks
- Solution for the data sparsity problem where one task has a limited number of labeled data

MTL in speech related applications

- Automatic Speech Recognition (ASR) tasks (Kim et al. 2017, Hori et al. 2017)
- Text-to-speech (TTS) (Chien et al. 2021)
- Speech emotion recognition (SER) (Li et al. 2019, Cai et al. 2021)
- Preliminary MTL based SI system (Xie et al. 2016)

Outline

Articulatory Dataset

Haskins Production Rate Comparison (HPRC)

- Recordings from 4 female and 4 male subjects reciting 720 phonetically balanced IEEE sentences (IEEE, 1969) at normal and fast production rates (Tiede et al., 2017)
- Every sentence produced at speaker's preferred 'normal' speaking rate and then a 'fast' repetition of the same
- Recordings done using 5-D electromagnetic articulometry (EMA) system
- Three additional TVs: Jaw Angle (JA), Tongue Middle Constriction Location (TMCL) and Tongue Middle Constriction Degree (TMCD)

Acoustic and Phoneme features

Mel-Frequency Cepstral Coefficients (MFCCs)

- 13 MFCCs extracted from fixed length segments (2 sec) of speech
- 20ms Hamming analysis window with a 10ms frame shift

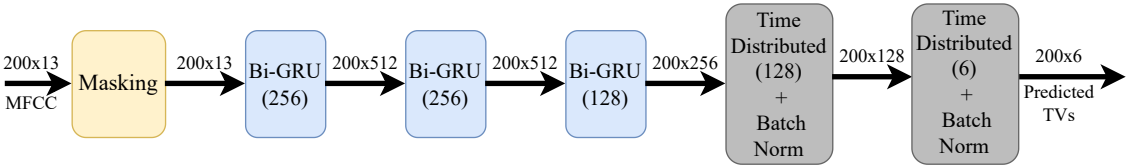
Phoneme Features

- Phone alignment extracted using the Penn Phonetics Lab Forced Aligner(P2FA)
- Removed allophonic variations of the monophones and retained only 40 monophone units
- One-hot encoded frame-wise monophone labels used as the phonetic features

Outline

Bidirectional Gated RNN (BiGRNN) SI model

- A novel SI system implemented with GRUs and trained in speaker-independent fashion
- MFCCs are utterance wise normalized (z-normalized) prior to model training
- No manual contextualization of acoustic features or post filtering of estimated TVs



Outline

Model training algorithms for MTL models

Algorithm 1 Iterative Loss Optimization

Require: : $x \epsilon R^{L \times d}, y_{ph}, y_{tv}, Epochs(\epsilon R)$
  while $i < Epochs$ **do**
    $\hat{y}_{tv} \leftarrow f_{\phi[i-1]}(x)$
    $L_{tv} \leftarrow MAE(\hat{y}_{tv}, y_{tv})$
    $\phi[i] \leftarrow \min_{\phi} L_{tv}$
    $\hat{y}_{ph} \leftarrow g_{\phi[i]}(x)$
    $L_{ph} \leftarrow CrossEntropy(\hat{y}_{ph}, y_{ph})$
    $\phi[i]^* \leftarrow \min_{\phi} L_{ph}$
    $i \leftarrow i + 1$
  end while=0

Model training algorithms for MTL models (Cont..)

Algorithm 2 Joint Loss Optimization

Require: : $x \epsilon R^{L \times d}, ValLoss, p \epsilon R, \alpha(0 < \alpha < 1), y_{ph}, y_{tv}$

while $ValLoss[i] < ValLoss[i-p]$ **do**

$\hat{y}_{ph} \leftarrow g_{\phi[i-1]}(x)$

$\hat{y}_{tv} \leftarrow f_{\phi[i-1]}(x)$

$L_{ph} \leftarrow CrossEntropy(\hat{y}_{ph}, y_{ph})$

$L_{tv} \leftarrow MAE(\hat{y}_{tv}, y_{tv})$

$L_{joint} \leftarrow L_{tv} + \alpha L_{ph}$

$\phi[i] \leftarrow \min_{\phi} L_{joint}$

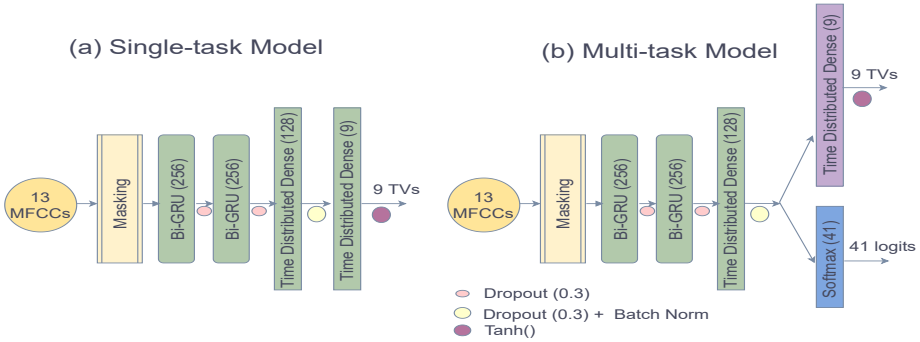
$i \leftarrow i + 1$

end while=0

Outline

Single-task vs Multi-task models for SI

- Learning the acoustic-to-phoneme mapping as a parallel task for speech inversion
- HPRC dataset (both 'normal' and 'fast' rate) with phonetic alignment for the speech utterances



Single-task vs Multi-task models for SI : Results

- Multi-task model with algorithm 2 training is faster to converge and better with TV predictions
- All models evaluated with the Pearson Product Moment Correlation (PPMC) scores

Model	LA	LP	JA	TTCL	TTCD	TMCL	TMCD	TBCL	TBCD	Average
Single-task	0.764	0.661	0.790	0.706	0.778	0.741	0.801	0.725	0.742	0.745
Multi-task (Algo 1)	0.792	0.681	0.796	0.747	0.793	0.775	0.799	0.760	0.764	0.767
Multi-task (Algo 2)	0.794	0.680	0.806	0.741	0.797	0.775	0.806	0.762	0.766	**0.770**

Table: Training Time : Single-task and Multi-task models

Model Type	No. of Trainable Parameters	Training Time
Single-task	2.19 M	10 $(\pm 2) min$
Multi-task (Algo 1)	2.20 M	61 $(\pm 5) min$
Multi-task (Algo 2)	2.20 M	15 $(\pm 2) min$

Outline

Baseline comparison with previous work

Model	Average PPMC score
Feed-forward model* [1]	0.705
CNN-BiLSTM model* [2]	0.755
Single-task BiGRNN model	0.745
Proposed Multi-task BiGRNN model	0.770

* The comparison is not perfect given that there can be differences in test splits used for evaluation

[1] A. S. Shahrebabaki, G. Salvi, T. Svendsen, and S. M. Siniscalchi, "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 135–147, 2022.

[2] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-Sequence Articulatory Inversion Through Time Convolution of Sub-Band Frequency Signals," in Proc.Interspeech 2020, 2020, pp. 2882–2886. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1140

Outline

Ablation Study

- Changed the weight α in the MTL model trained with algorithm 2
- α controls the amount of contribution from the phoneme prediction loss L_{ph} to the joint loss L_{joint}

	Average PPMC	**Phoneme Accuracy (%)**
$\alpha = 0.0$	0.743	2.25
$\alpha = 0.1$	0.762	70.60
$\alpha = 0.3$	0.766	72.53
$\alpha = 0.5$	**0.770**	72.88
$\alpha = 0.8$	0.759	72.90
$\alpha = 1.0$	0.758	73.60

Outline

Conclusions

- MTL model trained with algorithm 2 shows a relative improvement of 2.5% over the single-task model for TV prediction
- MTL with algorithm 2 training is faster to converge and better with TV predictions compared to algorithm 1
- MTL based SI system mostly improves in estimating location related TVs with respect to the single-task model
- MTL based SI system achieves the best PPMC scores over the existing SI systems on the HPRC dataset
- Ablation study shows the importance of multi-task learning (i.e. learning a related, shared task) on improving SI systems

Outline

Future work

- Tapping into larger existing datasets of audio, phonetic transcription (e.g. Librispeech (Panayotov et al., 2015) : 1000 hours of speech)
  - Transfer-learning and model adaptation paradigms to pre-train the MTL model to learn the acoustic-to-phoneme mapping with existing corpora of audio, phonetic transcriptions
- Exploring MTL based SI performance with other available articulatory datasets (eg. Wisconsin XRMB database (Westbury et al. 1994b))

Thank you !!

References

[1] R. Caruana, “Multitask Learning,” 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734
[2] Y. Zhang and Q. Yang, “A survey on multi-task learning,” IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2021.
[3] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, “12-in-1: Multi-task vision and language representation learning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recogni- tion (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 10 434–10 443.
[4] S. Kim, T. Hori, and S. Watanabe, “Joint ctc-attention based end- to-end speech recognition using multi-task learning,” 2017 IEEE International Conference on Acoustics, Speech and Signal Pro- cessing (ICASSP), pp. 4835–4839, 2017.
[5] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm,” in INTERSPEECH, 2017.

Auxiliary Slides !!

LA, TBCD, TTCD and TMCD TVs

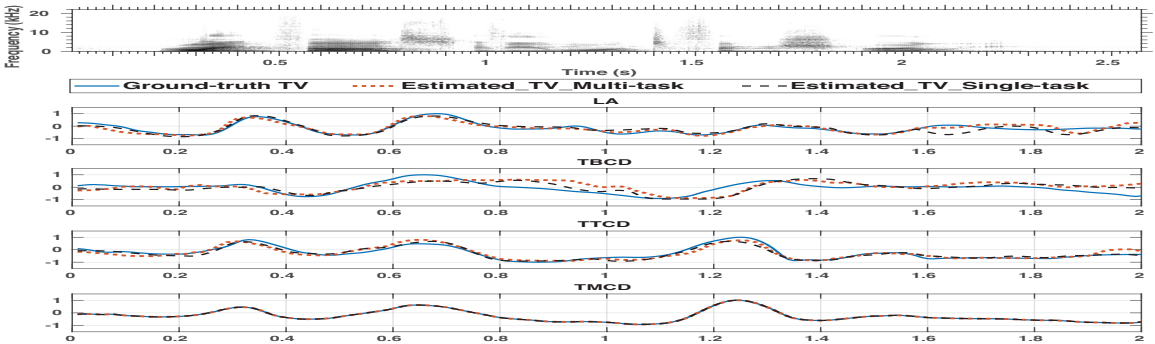


Figure: LA and constriction degree TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model

LP, TBCL, TTCL and TMCL TVs

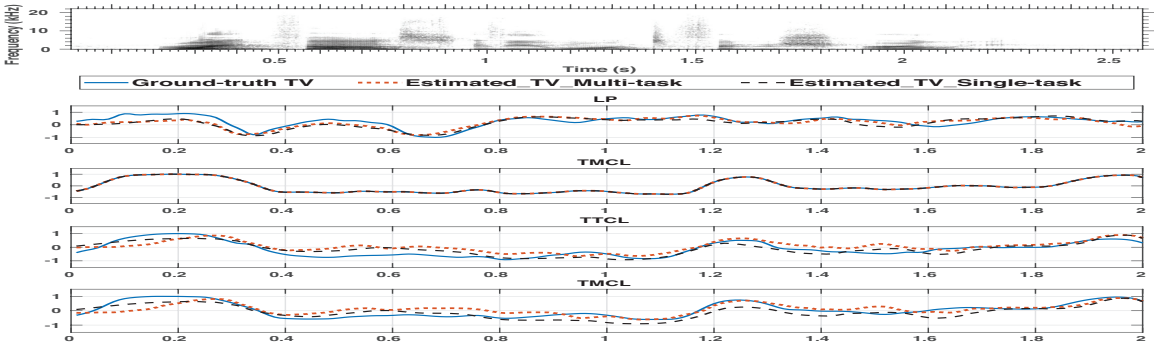


Figure: LP and constriction location TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model