A comparative study of estimating articulatory movements from phoneme sequences and acoustic features

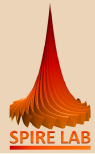
Abhayjeet Singh, Aravind Illa, Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering Department,
Indian Institute of Science (IISc), Bengaluru, India

Thursday, 07 May, 16:30 - 18:30

Poster Session: TH3.PB: Speech Production

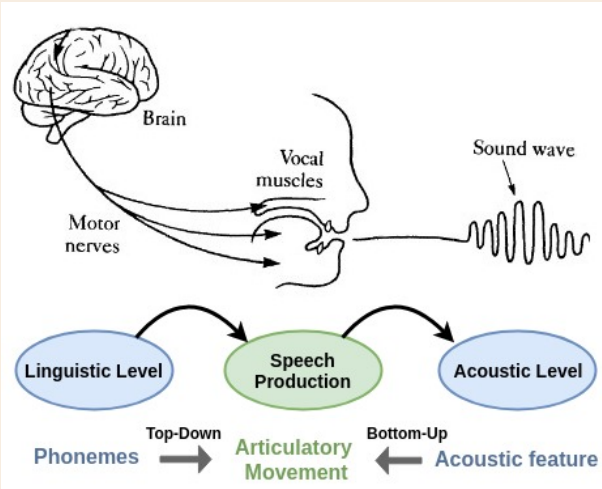
Overview



Overview

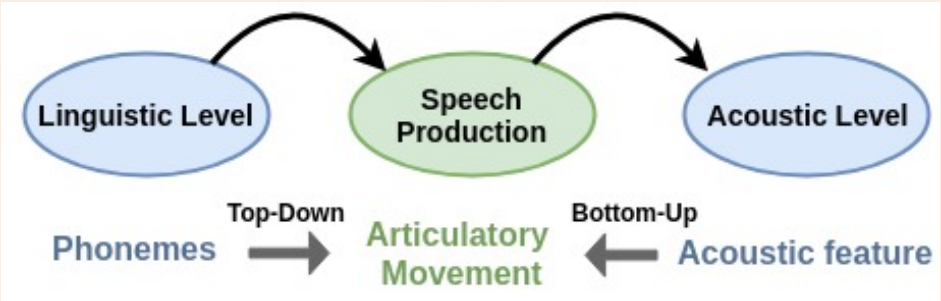
Top-down process in Speech Production

- In speech production, articulatory movements provide an intermediate representation between neuro-motor planning (high level) and speech acoustics (low level).
- Neuro-motor planning in brain aims to convey linguistic information as discrete abstract units which are passed via motor nerves to activate vocal muscles.



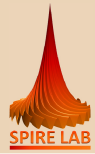
Motivation

- Estimating articulatory movement information from linguistic level or Phonemes (high level) → Top-Down approach.
- Estimating articulatory movement information from Acoustic features (Low level) → Bottom-up approach.



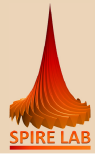
- To what extent articulatory motion can be extracted from the linguistic information (top-down) compared with that from acoustic to articulatory inversion (bottom-up).

Objectives of the work



- Prediction of articulatory motion from phoneme sequences.
- Comparison of performance of models predicting articulatory sequences from a phoneme sequence without timing information, with timing information and acoustic features.

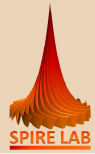
Objectives of the work



- To experimentally examine where production of articulatory sequences can be accurately determined from linguistic features without any timing information.

Input Features	**Encoded information**
Phoneme Sequences(PHN)	*linguistic*
Time Aligned Phonemes(TPHN)	*linguistic+timing*
MFCC	*linguistic+para-linguistic+timing*
MFCC+TPHN	*linguistic+para-linguistic+timing*

Overview



Data Collection

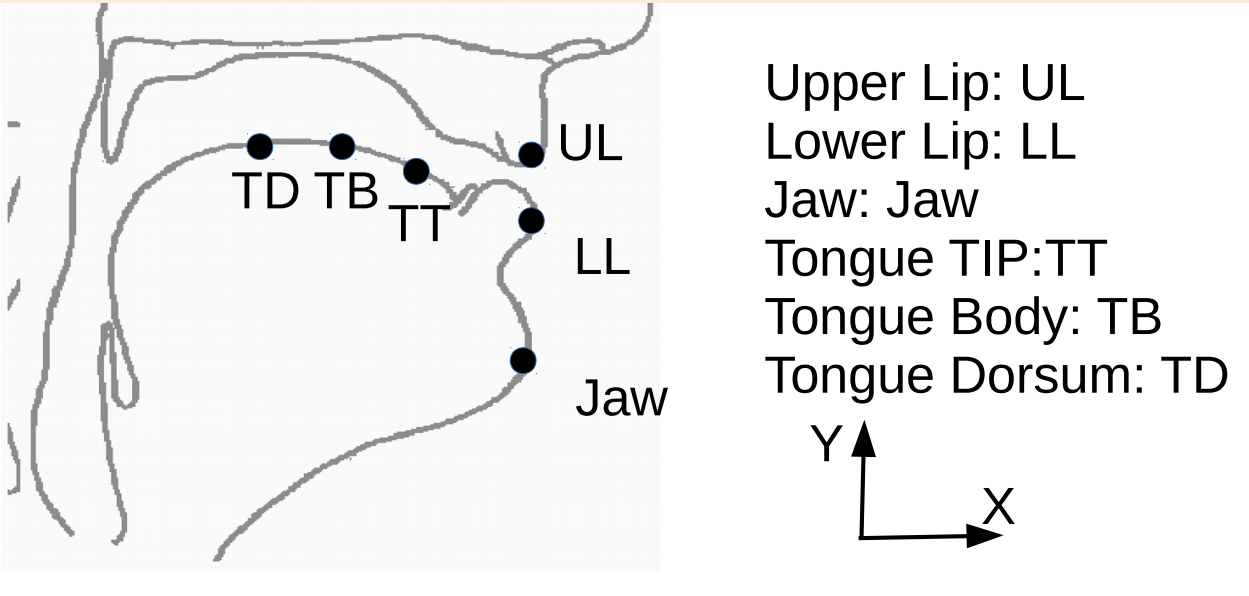
- **Dataset:** 460 phonetically balanced English sentences from MOCHA-TIMIT corpus[1].
- 10 subjects: 6 males (M1, M2, M3, M4, M5, M6) and 4 females (F1, F2, F3, F4) (20-28 years of age)
- Recorded audio using microphone and corresponding articulatoy movements using EMA AG501[2]

[1] A. Wrench, "Mocha-TIMIT speech database," The 18th International Conference on Pattern Recognition, 1999.

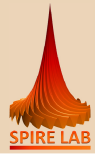
[2] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 4/2/2020

Data Collection

- Six sensors are connected: UL-upper lip, LL-lower lip, Jaw-jaw, TT-tongue tip, TB-tongue body, TD-tongue dorsum.



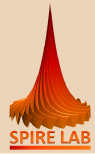
Input Features



- Force alignment is performed using Kaldi[3] speech recognition toolkit on recorded speech to obtain phonetic transcription which consists of 39 ARPABET symbols and an extra label for silence.
- Phonetic features: Input phoneme labels are represented as 40 dimensional one-hot vectors (PHN, TPHN).
- Acoustic features: Mel frequency cepstral coefficients (MFCC) with window length (20ms) and shift (10ms).

[3] Daniel Povey et al., "The Kaldi speech recognition toolkit," in IEEE workshop on automatic speech recognition and understanding, 2011.

Summary of Input Features



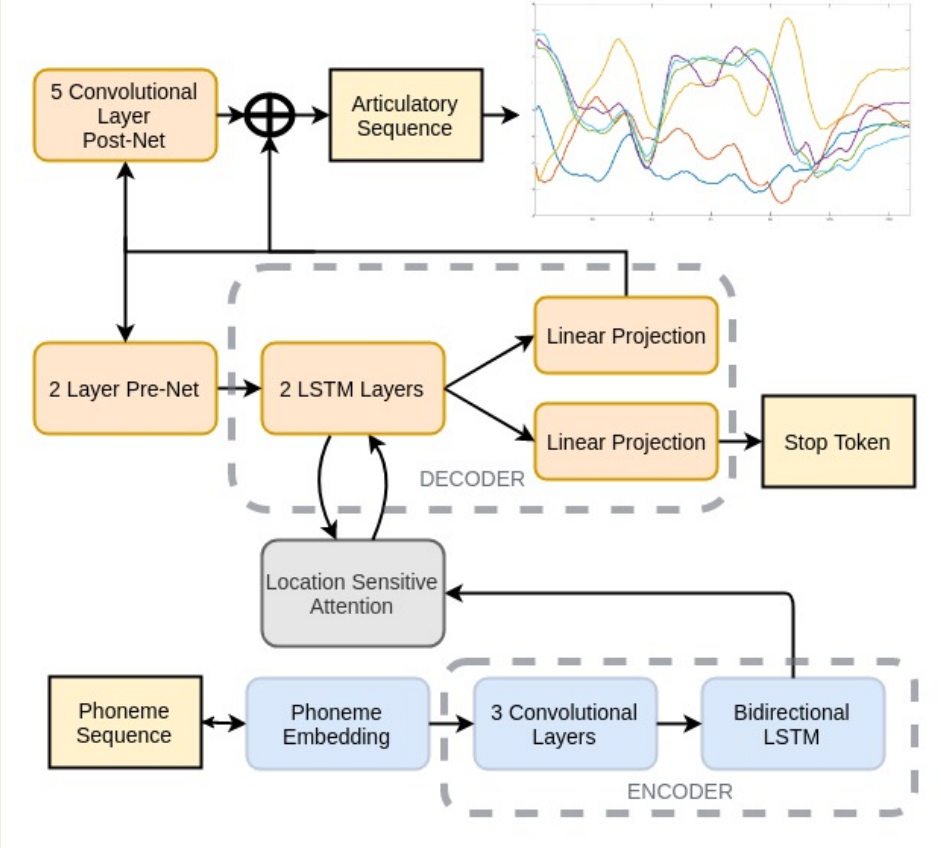
- Summary of input features, corresponding encoded information and models used for articulatory movement estimation

Input Features	Dimension	Encoded information	Model
Phoneme Sequences(PHN)	*40*	*linguistic*	*Attention*
Time Aligned Phonemes(TPHN)	*40*	*linguistic+timing*	*BLSTM*
MFCC	*13*	*linguistic+para-linguistic+timing*	*BLSTM*
MFCC+TPHN	*53*	*linguistic+para-linguistic+timing*	*BLSTM*

Overview

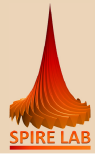
Tacotron architecture

- Tacotron[4] architecture models duration information for articulatory movement estimation from PHN.
- Three major components in tacotron model are: Encoder, Attention and Decoder.



[4] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 2018, pp. 4779–4783

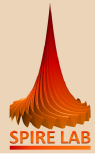
BLSTM model



- BLSTM model is used to learn mappings from TPHN, MFCC and MFCC+TPHN features to articulatory movements.
- This model includes 3 BLSTM layers with 256 units each and a dense layer at the output.

Overview

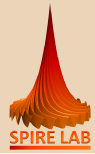
Experimental Setup



- Dataset split: Train(80%), Test(10%) and Validation(10%)*.
- Due to scarcity of training data for tacotron attention to learn alignments, we adopted Generic training and subject specific Fine-tuning on it.
- Evaluation metric: Correlation coefficient (CC) and RMSE.

*Subject-Dependent training: train and test sentences from same subject

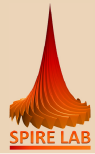
Training



- Three types of training done for all features:

Training Type	No. of Models	Training Initialization
Subject-Dependent	*10 per subject*	*Random*
Generic	*One for all subjects*	*Random*
Fine-Tuning	*10 per subject*	*Generic model*

Overview



Performance Comparison

- Performance comparison across different features:

Training	PHN		TPHN		MFCC		MFCC+TPHN	
	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
Subject-Dependent (SD)	2.04 (0.109)	0.33 (0.031)	1.243 (0.087)	0.808 (0.033)	1.116 (0.095)	0.844 (0.025)	1.05 (0.086)	0.87 (0.024)
Generic(G)	1.48 (0.098)	0.68 (0.043)	1.44 (0.108)	0.74 (0.046)	1.107 (0.091)	0.849 (0.023)	1.01 (0.083)	0.877 (0.022)
Fine-Tuning(FT)	**1.18** (0.113)	**0.806** (0.039)	**1.239** (0.084)	**0.815** (0.033)	**1.090** (0.088)	**0.854** (0.024)	**0.99** (0.085)	**0.884** (0.021)

Values between parentheses are average standard deviation taken across 10 subjects.

Performance Comparison

Training	PHN		TPHN		MFCC		MFCC+TPHN	
	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
Subject-Dependent (SD)	2.04 (0.109)	0.33 (0.031)	1.243 (0.087)	0.808 (0.033)	1.116 (0.095)	0.844 (0.025)	1.05 (0.086)	0.87 (0.024)
Generic(G)	1.48 (0.098)	0.68 (0.043)	1.44 (0.108)	0.74 (0.046)	1.107 (0.091)	0.849 (0.023)	1.01 (0.083)	0.877 (0.022)
Fine-Tuning(FT)	**1.18** (0.113)	**0.806** (0.039)	**1.239** (0.084)	**0.815** (0.033)	**1.090** (0.088)	**0.854** (0.024)	**0.99** (0.085)	**0.884** (0.021)

- In all cases **FT models perform better than SD models** → Pooling data from all subjects helps in learning generic specific mappings across multiple subjects.

Performance Comparison

Training	PHN		TPHN		MFCC		MFCC+TPHN	
	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
Subject-Dependent (SD)	2.04 (0.109)	0.33 (0.031)	1.243 (0.087)	0.808 (0.033)	1.116 (0.095)	0.844 (0.025)	1.05 (0.086)	0.87 (0.024)
Generic(G)	1.48 (0.098)	0.68 (0.043)	1.44 (0.108)	0.74 (0.046)	1.107 (0.091)	0.849 (0.023)	1.01 (0.083)	0.877 (0.022)
Fine-Tuning(FT)	**1.18** (0.113)	**0.806** (0.039)	**1.239** (0.084)	**0.815** (0.033)	**1.090** (0.088)	**0.854** (0.024)	**0.99** (0.085)	**0.884** (0.021)

- G models vs SD Models $\rightarrow$ performance for TPHN decreases due to lack of speaker specific para-linguistic features and in case PHN features G models perform better than SD due to lack of training data in SD.

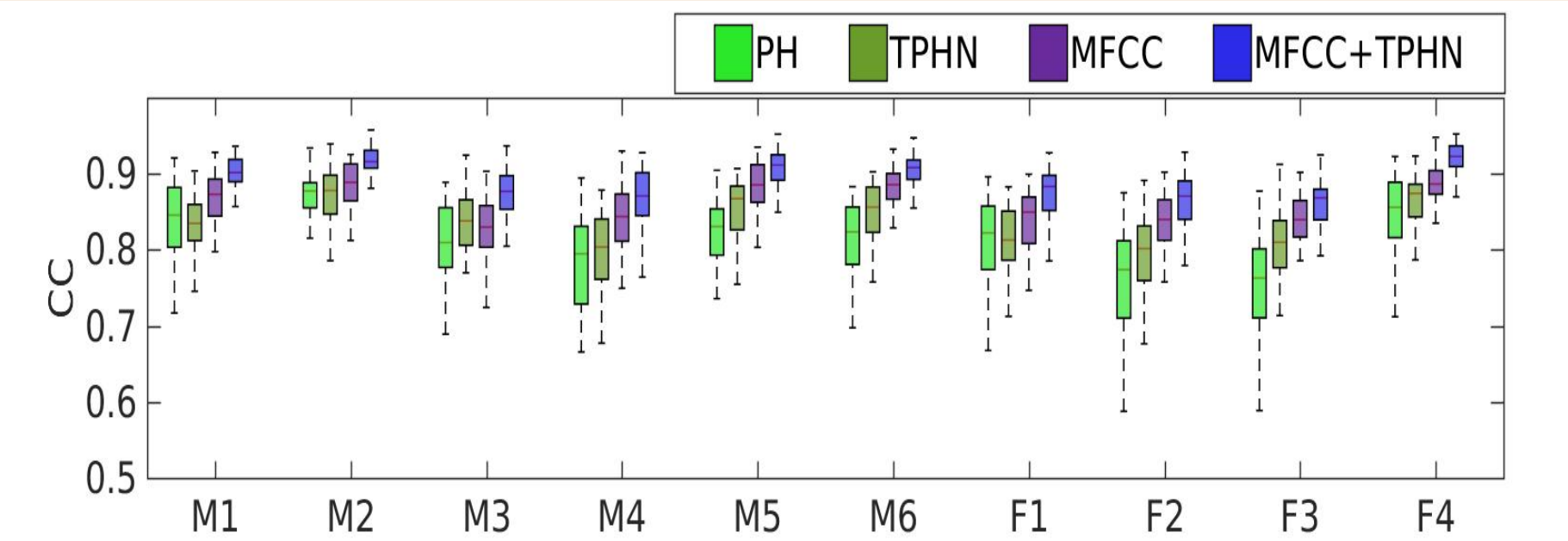
Performance Comparison

Training	PHN		TPHN		MFCC		MFCC+TPHN	
	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
Subject-Dependent (SD)	2.04 (0.109)	0.33 (0.031)	1.243 (0.087)	0.808 (0.033)	1.116 (0.095)	0.844 (0.025)	1.05 (0.086)	0.87 (0.024)
Generic(G)	1.48 (0.098)	0.68 (0.043)	1.44 (0.108)	0.74 (0.046)	1.107 (0.091)	0.849 (0.023)	1.01 (0.083)	0.877 (0.022)
Fine-Tuning(FT)	**1.18** (0.113)	**0.806** (0.039)	**1.239** (0.084)	**0.815** (0.033)	**1.090** (0.088)	**0.854** (0.024)	**0.99** (0.085)	**0.884** (0.021)

- Relative improvements from generic to fine-tune model across the PHN, TPHN, MFCC, and MFCC+TPHN are 18.53%, 9.3%, 0.5% and 0.8%, respectively.

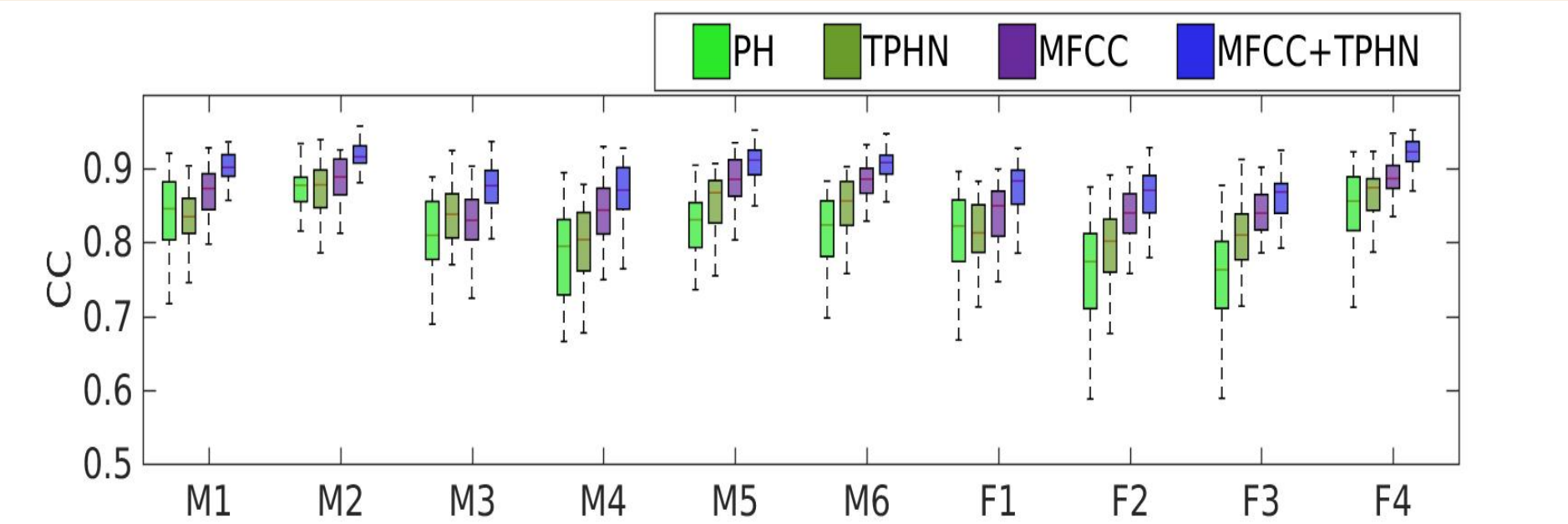
Discussion

- CC across all articulators for each speaker (M1, M2, M3, M4, M5, M6, F1, F2, F3 and F4):



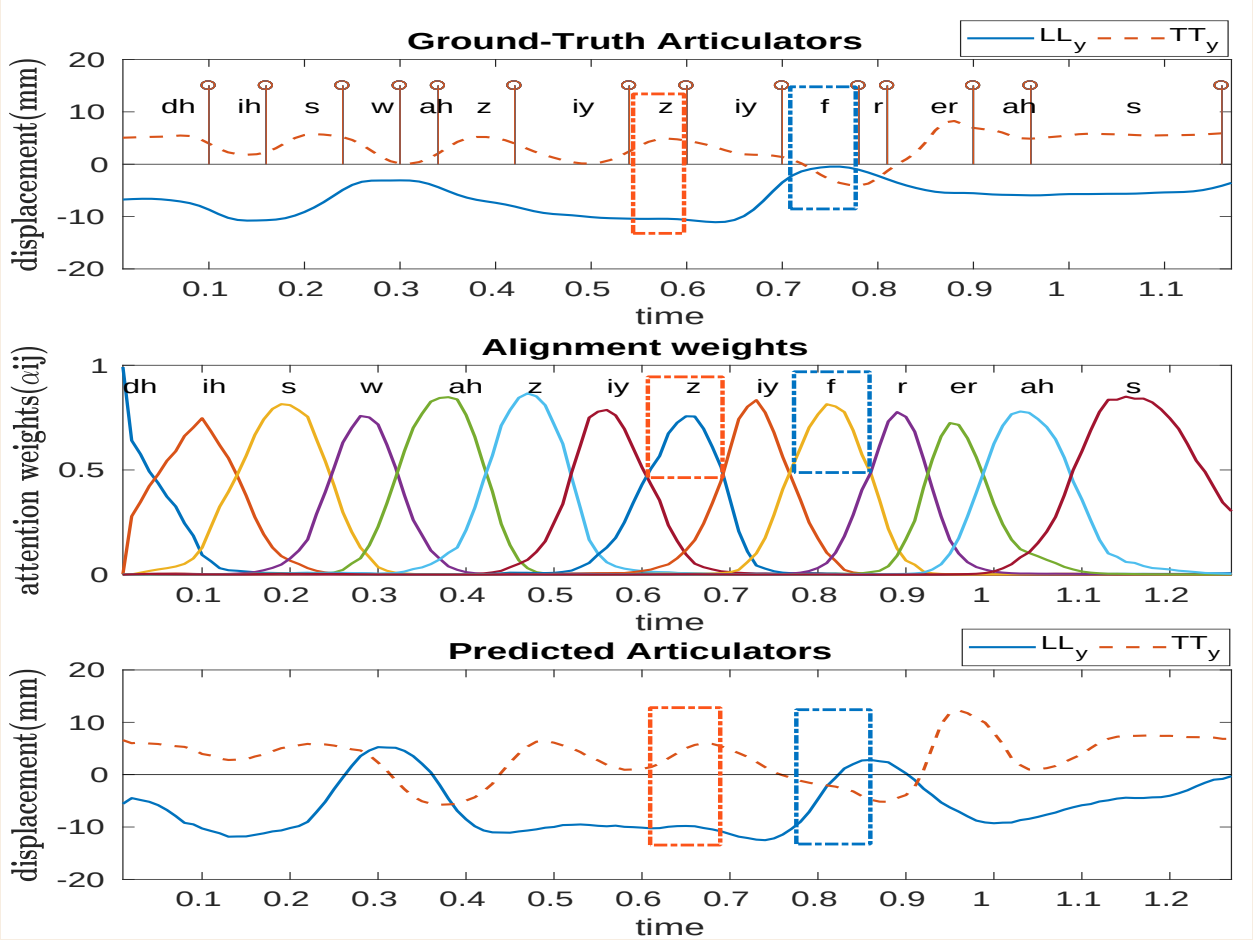
- PHN vs TPHN $\rightarrow$ performance is nearly same, indicating that timing information can be recovered from phoneme sequence to estimate articulatory trajectories using attention.

CC across all articulators for each speaker



- MFCC outperforms both PHN and TPHN because articulatory information is maximally preserved when speech acoustic signal is processed by auditory filters such as mel-scale.

Illustration of attention weights

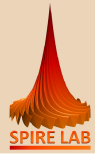


Comparison with [5] on MNGU dataset:

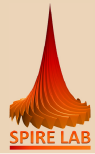
Model	Correlation
Random Initialized	*0.324*
Fine-tuned on top of Generic model	*0.778*
HMM model [5]	*0.600*

[5] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi, "An analysis of HMM-based prediction of articulatory movements," Speech Communication , vol. 52, no. 10, pp. 834–846, 2010.

Overview

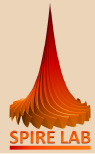


Conclusion



- Accuracy of estimating articulatory movements from PHN is comparable to features with timing information (TPHN).
- Attention networks are able to learn the timing information to estimate articulatory movements.

Future Work



- In future, we plan to utilize the estimated articulatory movements in speech synthesis task and in developing audio-visual speech synthesis systems.
- Analysis on the PHN and TPHN performance with respect to broad-phoneme classes, and place and manner of articulation.

Acknowledgement

- Authors thank all the subjects for their participation in the data collection and the **Department of Science and Technology, Govt. of India** for their support in this work.

THANK YOU

Have Questions/Suggestions?
Write to us at spirelab.ee@iisc.ac.in