Representation learning using convolution neural network for acoustic-to-articulatory inversion

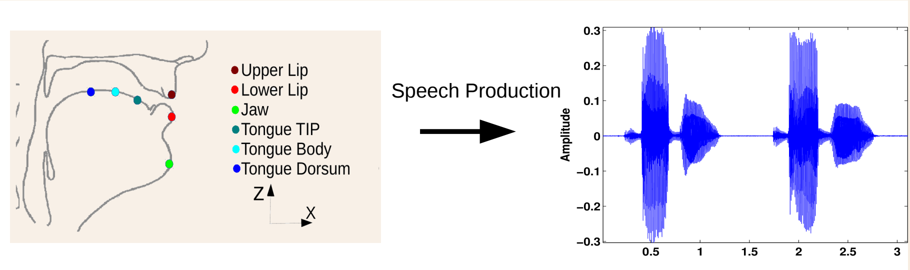
Aravind Illa, Prasanta Kumar Ghosh

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

ICASSP 2019,12 - 17 May.
Brighton, UK.

Section 1

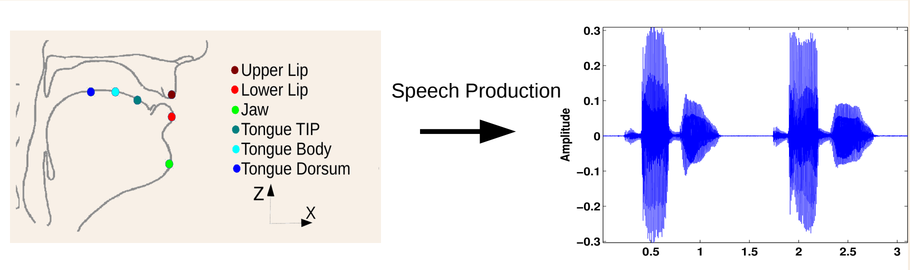
Speech Production



- Speech can be seen as the product of temporally overlapping gestures of articulators, each of which regulates the formation of constriction in vocal tract [1]

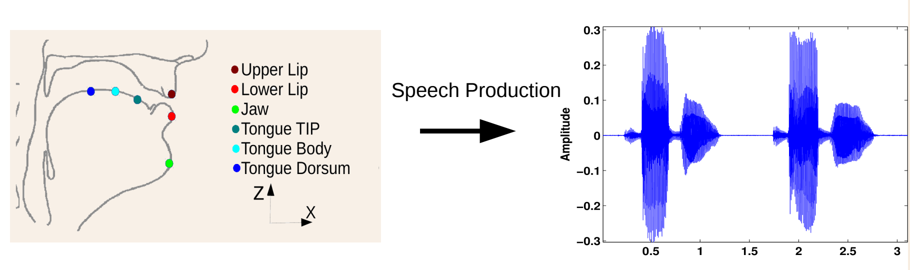
[1] Browman, C. P., and Goldstein, L. (1990).
[2] Livescu et.al. (2016).

Speech Production

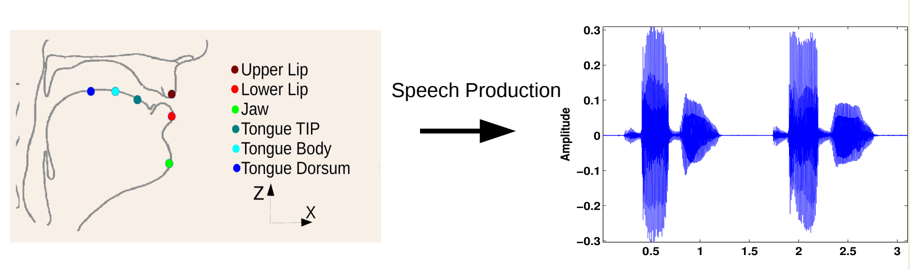


- Speech can be seen as the product of temporally overlapping gestures of articulators, each of which regulates the formation of constriction in vocal tract [1]
- Applications: ASR, Accent Conversion, Speaker Identification [2]

[1] Browman, C. P., and Goldstein, L. (1990).
[2] Livescu et.al. (2016).

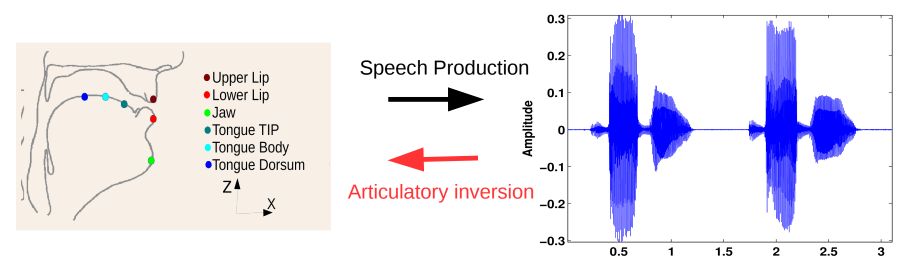


- Measurement Device: Electromagnetic articulography (EMA)



- Measurement Device: Electromagnetic articulography (EMA)
- Key articulators: lips, jaw, tongue and velum in the mid-sagittal plane.

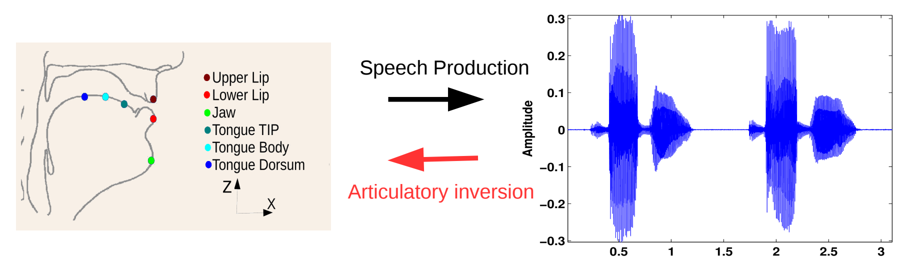
Acoustic to Articulatory Inversion (AAI)



Acoustic to Articulatory Inversion

- Estimating articulatory movements from speech acoustic features.

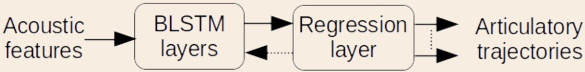
Acoustic to Articulatory Inversion (AAI)



Acoustic to Articulatory Inversion

- Estimating articulatory movements from speech acoustic features.
- Inverse mapping function is known to be **non-linear** and **non-unique**.

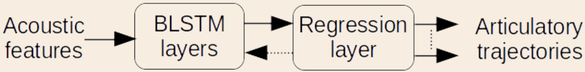
State-of-the-art model for AAI



Bidirectional LSTM

- RNNs are known to model the **temporal dynamics** by processing the sequence of input samples and maintaining a state information relative to history.

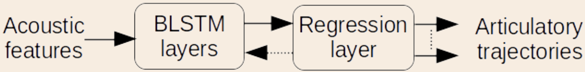
State-of-the-art model for AAI



Bidirectional LSTM

- RNNs are known to model the **temporal dynamics** by processing the sequence of input samples and maintaining a state information relative to history.
  - Preserves smoothing characteristics of articulatory trajectories

State-of-the-art model for AAI



Bidirectional LSTM

- RNNs are known to model the **temporal dynamics** by processing the sequence of input samples and maintaining a state information relative to history.
    - Preserves smoothing characteristics of articulatory trajectories
- Requires **adequate amount of data** from the **target subject**.

Choice of acoustic feautres for AAI

- Criterion: Maximize **Mutual Information** between acoustic and articulatory features.

[3] Prasanta Kumar Ghosh and Shrikanth Narayanan, (2010).

Choice of acoustic feautres for AAI

- Criterion: Maximize **Mutual Information** between acoustic and articulatory features.
- **Mel frequency cepstral coefficients (MFCCs)**[3] have been shown to be the best choice among the knowledge driven features (linear pre-diction coefficients (LPCs), cepstral representation of LPC (LPCC),and variants of LPC (line spectral frequency (LSF), reflection co-efficient (RC), log area ratio (LAR))

[3] Prasanta Kumar Ghosh and Shrikanth Narayanan, (2010).

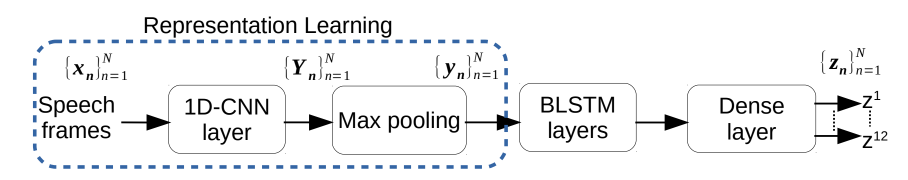
Choice of acoustic feautres for AAI

- Criterion: Maximize **Mutual Information** between acoustic and articulatory features.
- **Mel frequency cepstral coefficients (MFCCs)**[3] have been shown to be the best choice among the knowledge driven features (linear pre-diction coefficients (LPCs), cepstral representation of LPC (LPCC),and variants of LPC (line spectral frequency (LSF), reflection co-efficient (RC), log area ratio (LAR))
- Can we **learn** the representation of acoustic features directly from the raw waveform in a data driven manner?

[3] Prasanta Kumar Ghosh and Shrikanth Narayanan, (2010).

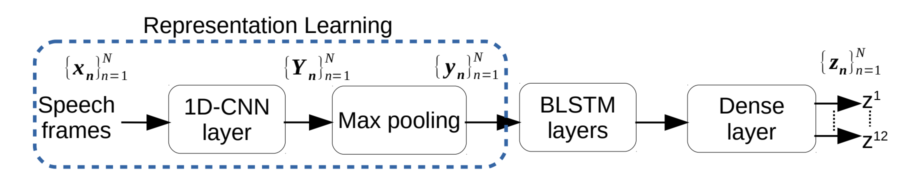
Section 2

End-to-End AAI



1. To extract the features from the speech frames, we consider a 1D-CNN layer as first layer.

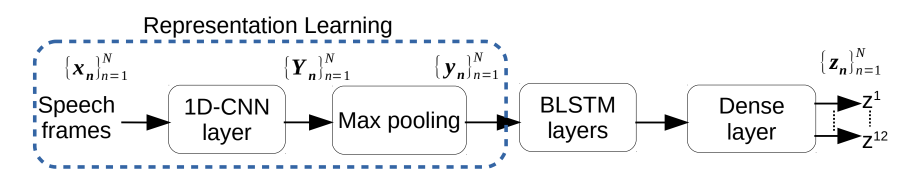
End-to-End AAI



1. To extract the features from the speech frames, we consider a 1D-CNN layer as first layer.
2. We compute the output of the convolution filter by

$$\mathbf{Y}_n = \sigma(\log(|\mathbf{F} * \mathbf{x}_n + \mathbf{b}|)) \tag{1}$$

End-to-End AAI

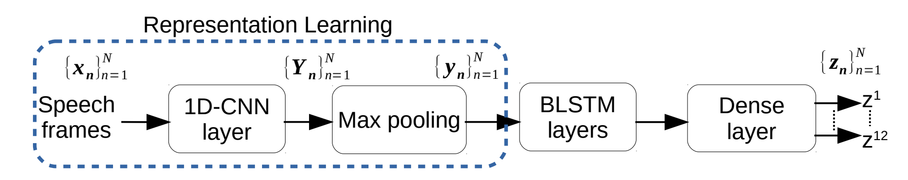


1. To extract the features from the speech frames, we consider a 1D-CNN layer as first layer.
2. We compute the output of the convolution filter by

$$\mathbf{Y}_n = \sigma(\log(|\mathbf{F} * \mathbf{x}_n + \mathbf{b}|)) \tag{1}$$

3. We propose an end-to-end network for AAI by cascading a CNN layer to the state-of-the-art BLSTM network.

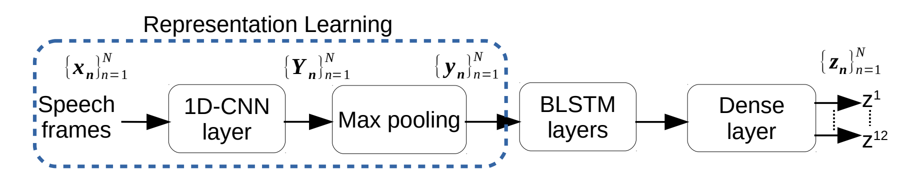
End-to-End AAI



Goal of Investigation

1. Can we **learn** the representation of acoustic features directly from the **raw waveform** using **1-D CNN**?

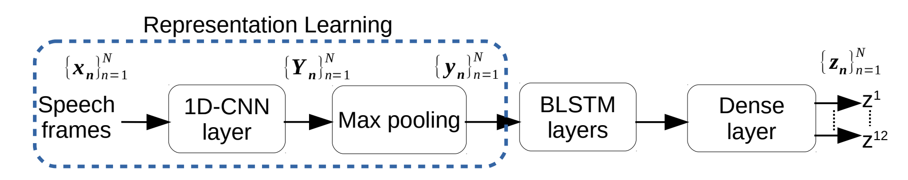
End-to-End AAI



Goal of Investigation

1. Can we **learn** the representation of acoustic features directly from the **raw waveform** using **1-D CNN**?
2. What kind of **representations** are learned by 1-D CNN?

End-to-End AAI



Goal of Investigation

1. Can we **learn** the representation of acoustic features directly from the **raw waveform** using **1-D CNN**?
2. What kind of **representations** are learned by 1-D CNN?
3. Is the **performance** of learnt features from 1-D CNN are competitive with knowledge based features (MFCC)?

Section 3

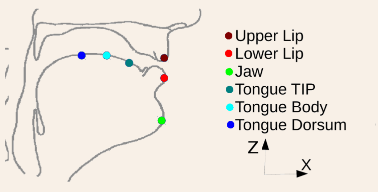
Data Collection: EMA

1. Electromagnetic articulography (EMA) AG501 was used to record the articulatory movement data.
   1. It has 24 channels to measure the horizontal, vertical and lateral displacements and angular orientations of a maximum of 24 sensors.
   2. Available sampling rate: 250 Hz and 1250 Hz. [4]

[4] 3d electromagnetic articulograph, available http://www.articulograph.de/

Data Collection

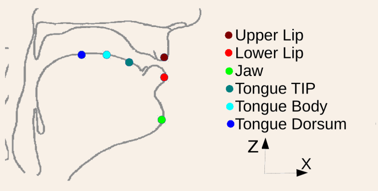
1. Six sensors are connected to obtain twelve articulatory features denoted by UL_x, UL_z, LL_x, LL_z, Jaw_x, Jaw_z, TT_x, TT_z, TB_x, TB_z, TD_x, TD_z.



[5] A. Wrench, MOCHA-TIMIT, speech database, Department of Speech and Language Sciences, Queen Margaret University College,Edinburgh, 1999.

Data Collection

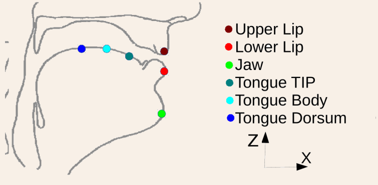
1. Six sensors are connected to obtain twelve articulatory features denoted by UL_x, UL_z, LL_x, LL_z, Jaw_x, Jaw_z, TT_x, TT_z, TB_x, TB_z, TD_x, TD_z.
2. 460 phonetically balanced English sentences [5]



[5] A. Wrench, MOCHA-TIMIT, speech database, Department of Speech and Language Sciences, Queen Margaret University College,Edinburgh, 1999.

Data Collection

1. Six sensors are connected to obtain twelve articulatory features denoted by UL_x, UL_z, LL_x, LL_z, Jaw_x, Jaw_z, TT_x, TT_z, TB_x, TB_z, TD_x, TD_z.
2. 460 phonetically balanced English sentences [5]
3. acoustic-articulatory data are recorded from 8 subjects (4 male and 4 female)
   –Total: 3.19 hours
   –Average duration/subject: 23.97 ($\pm$ 2.43) minutes.



[5] A. Wrench, MOCHA-TIMIT, speech database, Department of Speech and Language Sciences, Queen Margaret University College,Edinburgh, 1999.

Section 4

Experimental Setup

- Total 460 sentences:

  –368 for Train set (80%)

  –46 for validation (10%) and test (10%) sets.

Experimental Setup

- Total 460 sentences:
  –368 for Train set (80%)
  –46 for validation (10%) and test (10%) sets.
- Proposed AAI model details:
  –1-D CNN as First layer followed by three BLSTM layers with 150 units
  –Linear regression layer at last.

Experimental Setup

- Total 460 sentences:
  - –368 for Train set (80%)
  - –46 for validation (10%) and test (10%) sets.
- Proposed AAI model details:
  - –1-D CNN as First layer followed by three BLSTM layers with 150 units
  - –Linear regression layer at last.
- Baseline AAI model details:
  - –First three are BLSTM layers with 150 units
  - –Linear regression layer at last.

Experimental Setup

- Total 460 sentences:
  –368 for Train set (80%)
  –46 for validation (10%) and test (10%) sets.
- Proposed AAI model details:
  –1-D CNN as First layer followed by three BLSTM layers with 150 units
  –Linear regression layer at last.
- Baseline AAI model details:
  –First three are BLSTM layers with 150 units
  –Linear regression layer at last.
- Evaluation metrics:
  –Root Mean Square Error (RMSE)
  –Correlation Coefficient (CC).

Experimental Conditions

- Analysis on pre-emphasis:

  –Without pre-emphasis

  –With pre-emphasis=0.97

Experimental Conditions

- Analysis on pre-emphasis:
  –Without pre-emphasis
  –With pre-emphasis=0.97
- Data pooling for training:
  –Independent training
  –Joint training
  –Adaptation.

Experimental Conditions

- Analysis on pre-emphasis:
  - –Without pre-emphasis
  - –With pre-emphasis=0.97
- Data pooling for training:
  - –Independent training
  - –Joint training
  - –Adaptation.
- Comparison with Baseline approach:
  - –End-to-End AAI
  - –MFCC based BLSTM AAI .

Experimental Conditions

- Analysis on pre-emphasis:
  - –Without pre-emphasis
  - --With pre-emphasis=0.97
- Data pooling for training:
  - –Independent training
  - –Joint training
  - –Adaptation
- Comparison with Baseline approach:
  - –End-to-End AAI
  - –MFCC based BLSTM AAI

Analysis on pre-emphasis

Table: Performance of AAI with and without pre-emphasis.

	N_{cf}	$RMSE_{avg}$	CC_{avg}
Without Pre-emphasis	40	1.81	0.78
	100	1.82	0.78
	256	1.86	0.77
Pre-emphasis = 0.97	40	1.68	0.81
	100	1.66	0.81
	256	1.66	0.81

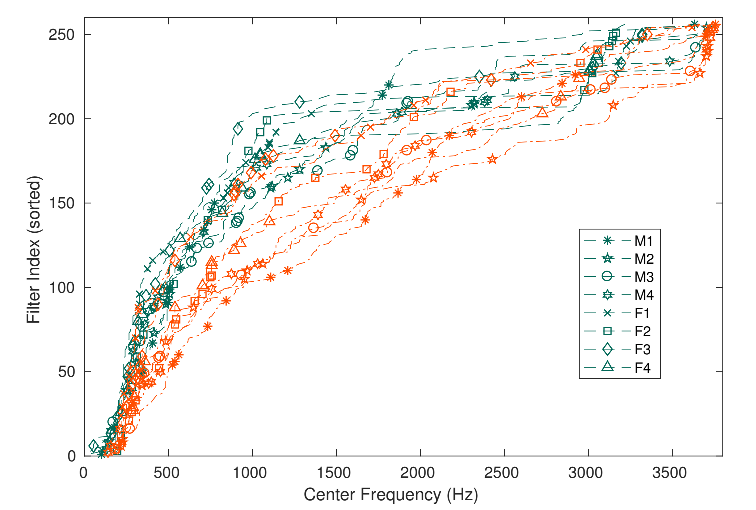


Figure: With (-·-·) and without (- - -) pre-emphasis operation.

Filters with center frequency ≤ 1000Hz

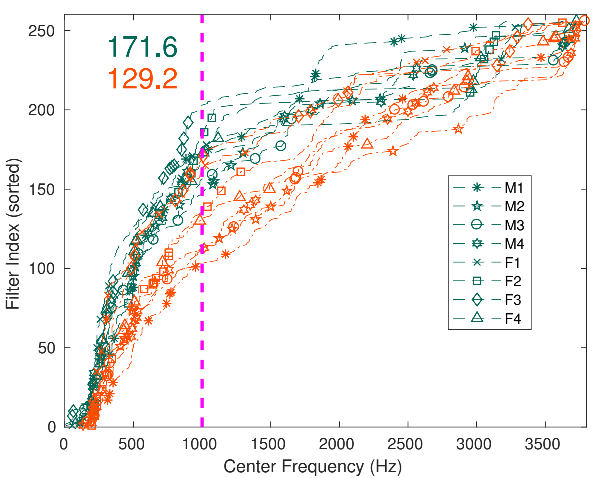


Figure: With (-·-) and without (- - -) pre-emphasis operation.

Joint training and adaptation

Table: Performance of AAI in terms of $RMSE_{avg}$ (mm) with different training approaches.

Training	N_{cf} =40	N_{cf} =100	N_{cf} =256
Independent	1.68	1.66	1.66
Joint	1.56	1.63	1.60
Adaptation	1.47	1.50	1.49

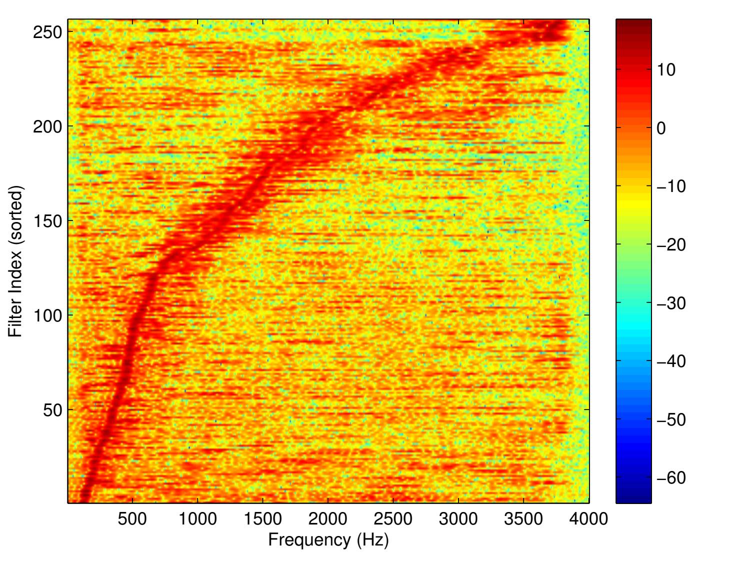


Figure: Magnitude response of learned filters after joint training

Comparison with MFCC

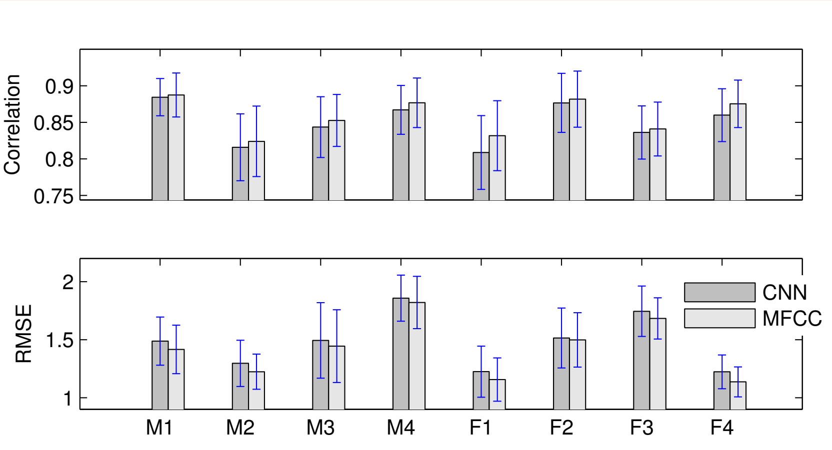


Figure: MFCC vs CNN features.

Comparison with MFCC

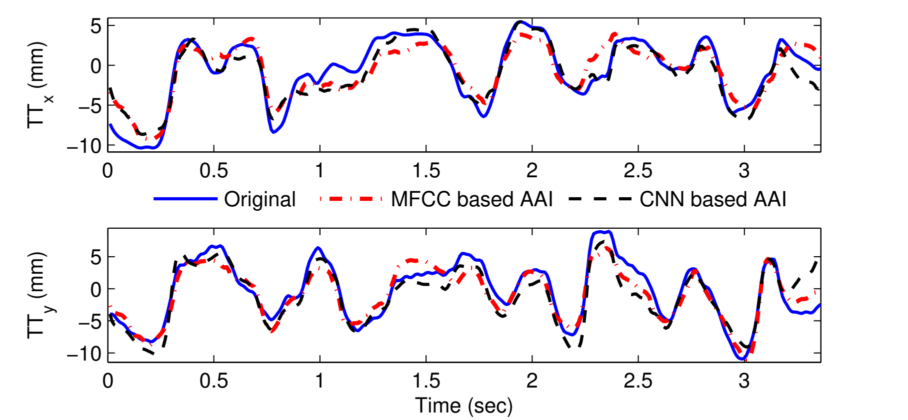


Figure: Tongue Tip trajectories.

Section 5

Conclusion

- Experiments performed with 8 subjects revealed that the proposed CNN based approach performs on par with MFCC.

Conclusion

- Experiments performed with 8 subjects revealed that the proposed CNN based approach performs on par with MFCC.
- Pre-emphasis helps to boost the high frequency components, thereby higher formant regions and plays an important role in improving the performance of AAI.

Conclusion

- Experiments performed with 8 subjects revealed that the proposed CNN based approach performs on par with MFCC.
- Pre-emphasis helps to boost the high frequency components, thereby higher formant regions and plays an important role in improving the performance of AAI.
- Interestingly, the frequency response is band-pass in nature and center frequencies are found to be similar to those of mel-scale.

Conclusion

- Experiments performed with 8 subjects revealed that the proposed CNN based approach performs on par with MFCC.
- Pre-emphasis helps to boost the high frequency components, thereby higher formant regions and plays an important role in improving the performance of AAI.
- Interestingly, the frequency response is band-pass in nature and center frequencies are found to be similar to those of mel-scale.
- This could be due to the fact that the speech gestural information is maximally preserved when speech signal is processed by auditory filters such as mel-scale or bark-scale [6].

[6] Prasanta Kumar Ghosh, Louis M Goldstein, and Shrikanth Narayanan (2011).

Acknowledgment

- All the subjects for their participation in the EMA data collection.
- Nisha, Kaustubha for helping in recordings.
- Pratiksha Trust for their support.
- We thank IEEE Signal Processing Society and IEEE Bangalore section for supporting conference travel.

–Thanks!!

Thanks for your attention!

References

1. Browman, C. & Goldstein, L. (draft). Articulatory Phonology (1990)
2. Livescu, K., Rudzicz, F., Fosler-Lussier, E., Hasegawa-Johnson, M., & Bilmes, J. (2016). Speech Production in Speech Technologies: Introduction to the CSL Special Issue. Computer Speech & Language, 36, 165172.
3. Prasanta Kumar Ghosh and Shrikanth Narayanan, A generalized smoothness criterion for acoustic-to-articulatory inversion, The Journal of the Acoustical Society of America, vol. 128, no. 4, pp. 21622172, 2010.
4. EMA AG501: 3d electromagnetic articulograph, available http://www.articulograph.de/
5. A. Wrench, MOCHA-TIMIT, speech database, Department of Speech and Language Sciences, Queen Margaret University College,Edinburgh, 1999
6. Prasanta Kumar Ghosh, Louis M Goldstein, and Shrikanth S Narayanan, Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures, The Journal of the Acoustical Society of America, vol. 129, no. 6, pp. 40144022, 2011.
7. K. Richmond, Estimating articulatory parameters from the acoustic speech signal, Ph.D. dissertation, University of Edinburgh, 2002.
8. Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, L. C. (2015). A DEEP RECURRENT APPROACH FOR ACOUSTIC-TO-ARTICULATORY INVERSION (pp. 4450–4454).
9. Li, M., Kim, J., Lammert, A., Ghosh, P. K., Ramanarayanan, V., & Narayanan, S. (2016). Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer Speech & Language, 36, 196211.