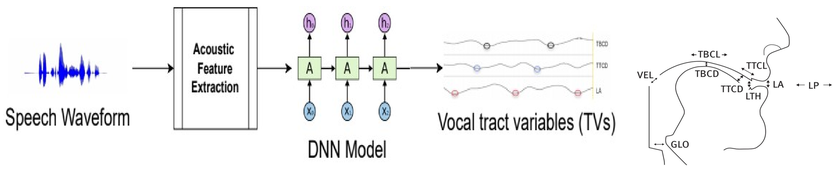
The Secret Source : Incorporating Source Features to Improve Acoustic-to-articulatory Speech Inversion

Yashish M. Siriwardena and Carol Espy-Wilson
Institute for Systems Research, Electrical and Computer Engineering,
University of Maryland College Park, USA

Paper #5096

Acoustic-to-articulatory Speech Inversion (SI)



- The SI system solves the inverse problem of determining the trajectories of the movement of speech articulators from speech
- The 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)

Constriction	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)

Motivation for our work

- Learning proxy source level features (Aperiodicity, Periodicity and Pitch) to leverage any source-filter interactions to improve SI task
- Effectiveness of Multi Task Learning (MTL) frameworks in learning parallel tasks or related additional targets to improve SI
- Exploring different deep neural network (DNN) based model architectures (eg. BiLSTMS, CNN-BiLSTMs, Temporal Convolutional Networks (TCN)) in developing speaker-independent SI systems

Articulatory Datasets

The X-ray microbeam (XRMB)

- Naturally spoken isolated sentences and short read paragraphs collected from 32 male and 25 female subjects
- X-ray microbeam cinematography of the midsagittal plane

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)
- 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)

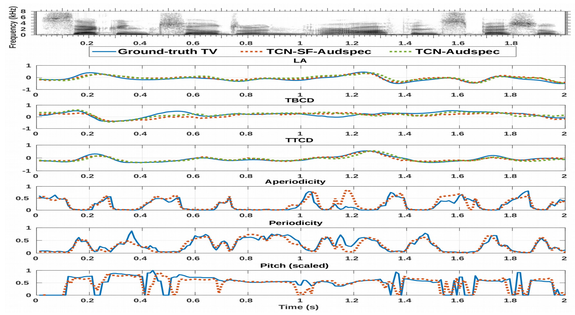
Results

Comparison with baseline SI systems: XRMB dataset

Model	LA	LP	TBCL	TBCD	TTCL	TTCD	Ap.	Per.	Pitch	AVG. TVs	Avg. all
TCN-Audspec	0.7977	0.7942	0.7883	0.7836	0.7743	0.7684	-	-	-	0.7844	-
TCN-SF-Audspec	**0.8448**	**0.8640**	**0.8604**	**0.8818**	**0.9029**	**0.9005**	**0.9082**	**0.8860**	**0.9021**	**0.8770 (9.3%)**	**0.8834**
TCN-Mspec	0.7432	0.7427	0.7366	0.7244	0.7244	0.6993	-	-	-	0.7273	-
TCN-SF-Mspec	0.8364	0.8639	0.8727	0.8607	0.8807	0.8917	0.8732	0.9005	0.8638	0.8677 (14%)	0.8715
BiGRNN-MFCC	0.8801	0.6200	0.8580	0.7382	0.6922	0.9206	-	-	-	0.7848	-
BiGRNN-SF-MFCC	0.8810	0.6211	0.8628	0.7365	0.7019	0.9191	0.8693	0.9163	0.7209	0.7871 (0.2%)	0.8032
CNN-BiGRNN-Mspec	0.8801	0.6165	0.8505	0.7355	0.7146	0.9171	-	-	-	0.7858	-
CNN-BiGRNN-SF-Mspec	0.8799	0.6246	0.8566	0.7302	0.7065	0.9175	0.8794	0.9296	0.7441	0.7859 (0.01%)	0.8076
CNN-BLSTM-Mspec	0.8770	0.6184	0.8463	0.7200	0.6915	0.9197	-	-	-	0.7788	-
CNN-BLSTM-SF-Mspec	0.8774	0.6202	0.8525	0.7172	0.6941	0.9180	0.8734	0.9263	0.7442	0.7799 (0.1%)	0.8026

Estimated TVs and Source Features

- LA and constriction degree TVs + source features for the utterance 'second children are often special' estimated by the proposed TCN-SF-Audspec model compared to the TCN-Audspec



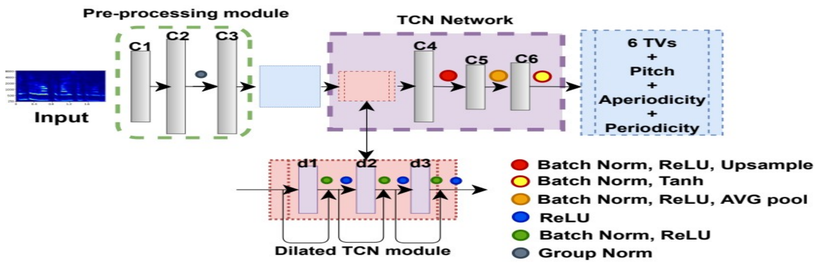
Comparison with baseline SI systems: HPRC dataset

Model	AVG. 9 TVs	Avg. all
TCN-Audspec	0.4805	-
TCN-SF-Audspec	**0.7573 (27.7%)**	**0.7636**
TCN-Mspec	0.4763	-
TCN-SF-Mspec	0.6503 (17.4%)	0.6621
BiGRNN-MFCC	0.7118	-
BiGRNN-SF-MFCC	0.7153 (0.3%)	0.7263
CNN-BiGRNN-Mspec	0.7277	-
CNN-BiGRNN-SF-Mspec	0.7290 (0.1%)	0.7461
CNN-BLSTM-Mspec	0.7245	-
CNN-BLSTM-SF-Mspec	0.7259 (0.1%)	0.7428

Input Acoustic Features

- Auditory Spectrograms (Audspecs)
  - Sound signals in the auditory pathway undergo a series of complex transformations and converts the acoustic spectrum of the stimulus into an internal representation, called the auditory spectrum
  - Enhanced and a noise-robust estimate of the Fourier-based spectrogram with roughly a logarithmic frequency scale (Wang et al., 1994)

Temporal Convolution Network (TCN)



- Takes in the Audspecs as input and estimates both TVs and source level features (aperiodicity, periodicity and pitch) as the output
- All models trained and evaluated in a 'speaker-independent' fashion

Conclusions and Future Work

- Incorporating source features into the mix of TVs is helping the estimation of articulatory variables and hence improving the performance of SI systems
- The proposed TCN model which uses Audspecs (or Mspecs) as inputs shows the best improvement in performance
- The improvement in performance is consistent across two publicly available articulatory datasets (XRMB and HPRC datasets)
- Both the input speech representation and the DNN model architecture play a role in learning complex dependencies between the source and articulatory targets
- Further analysis needs to be done to investigate the ways and instances by which the source features are actually interacting with the TVs and what the TCN models are actually capturing as source-filter interactions

Acknowledgments

This work was supported by the National Science Foundation grant IIS1764010

References

[1] Catherine P Browman and Louis Goldstein, "Articulatory Phonology : An Overview *," Phonetica, vol. 49, pp. 155–180, 1992
[2] John R Westbury, "Speech Production Database User ' S Handbook," IEEE Personal Communications - IEEE Pers. Commun., vol. 0, no. June, 1994
[3] Kuansan Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 3, pp. 421–435, 1994