A Neural-Network-Based Formant Tracker Trained on Unlabeled Data

Jason Lilley and H. Timothy Bunnell

Nemours Biomedical Research

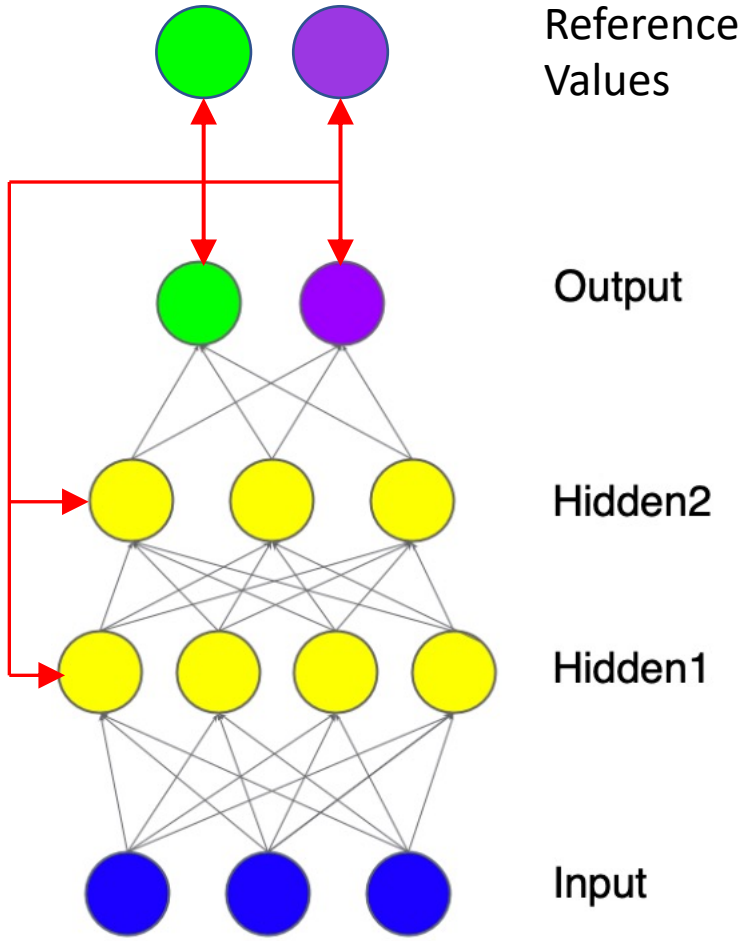
Wilmington, DE, USA

The Problem: Formant measurements for large corpora

- Formant measurement is often a vital step in phonetic analysis
- Manual formant measurement can be difficult, tedious
- Popular off-the shelf, LPC-based formant trackers are prone to large errors, at least with default settings (Deng et al. 2006; Schiel & Zitzelberger, 2018; Dissen et al. 2019)
- Perhaps better measurements can be obtained by tuning tracker settings to individual speakers
- But this would be quite tedious for a large corpus of many speakers

Deep Neural Networks (DNNs)

- Consist of layers of **units** (circles)
- Units compute a nonlinear function of the weighted sum of its inputs: $y = f(WX + B)$
- Outputs are compared to target reference values to compute a **loss** (error)
- Loss is used to adjust model weights in a process called backpropagation (red arrows)
- Special types of neural networks:
  - Convolutional neural network (CNN): Used for image input (e.g. spectrograms)
  - Recurrent neural network (RNN): Used for sequences (e.g. sound waves)
    - **LSTM** (Long Short-Term Memory): one type of RNN



“DeepFormants”: a DNN Formant Tracker

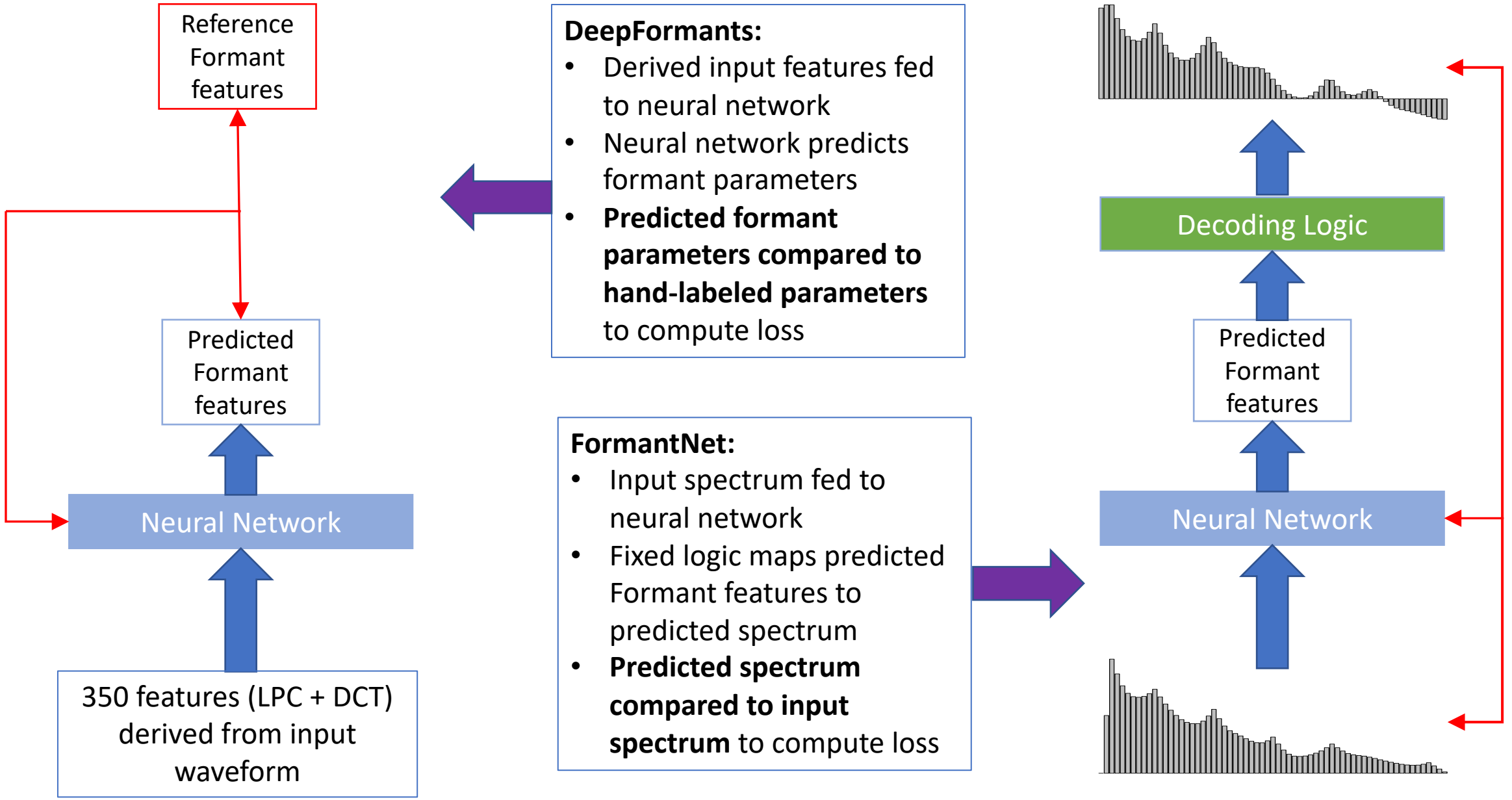
- Dissen et al. (2016, 2019) trained a neural network (“DeepFormants”) to estimate formant frequencies F1-F3
  - Trained on the hand-measured F1-F3 frequencies of a 538-file subset (Deng et al. 2004) of the 6300-file TIMIT corpus (Garofolo et al. 1993)
    - 346 files used for training; 192 withheld for testing
  - Two models: one LSTM; one a combination RNN-CNN (“DF-RCNN”)
- Showed results generally superior to other methods on the TIMIT dataset (Praat, WaveSurfer, MSR, KARMA; see also Schiel & Zitzelberger, 2018)

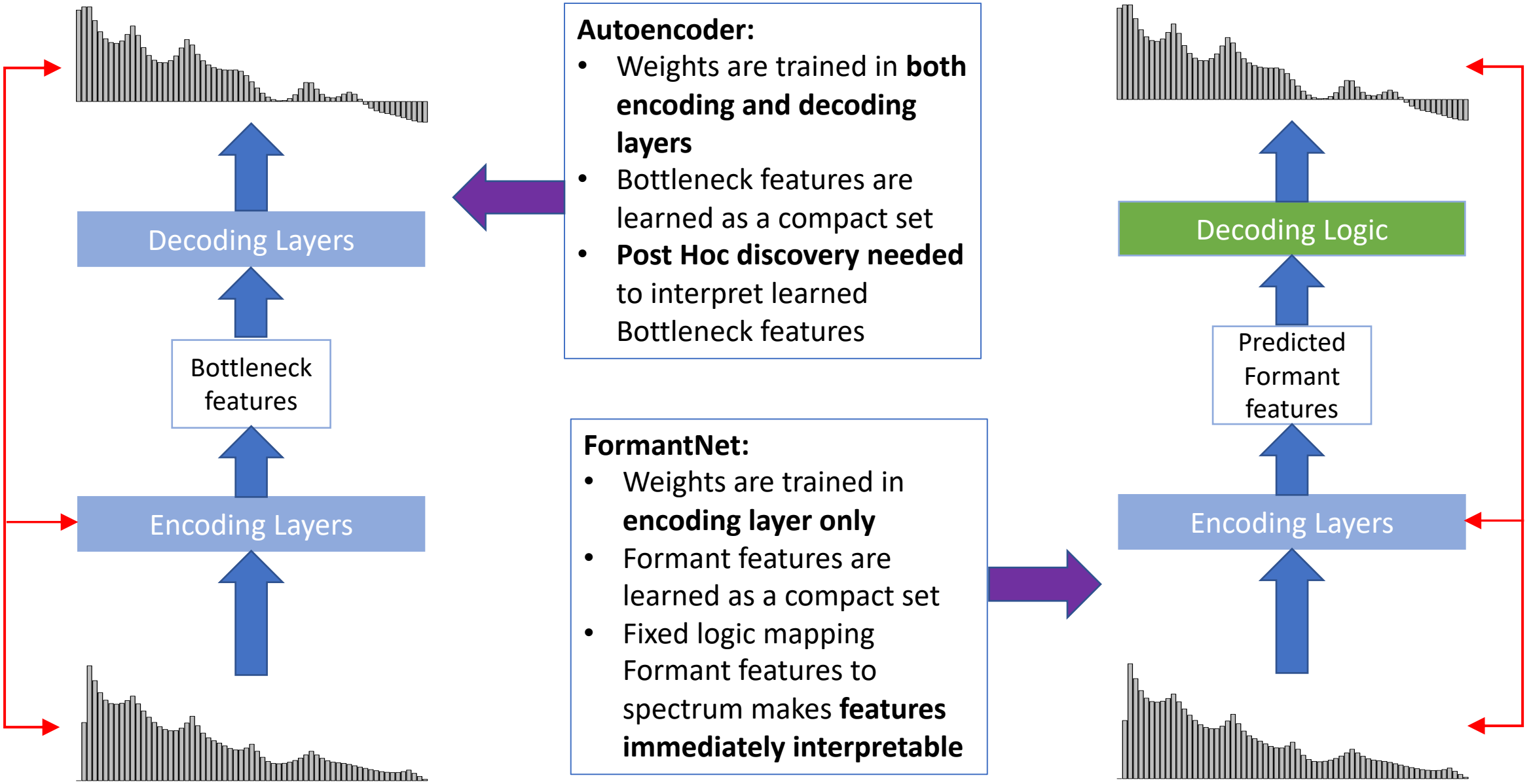
“DeepFormants” uses supervised training

- TIMIT-trained “DeepFormants” models show higher errors on other corpora, e.g. Hillenbrand and Clopper datasets (Dissen et al. 2019)
  - They demonstrate improved performance by **adapting** their models to the new datasets
  - But the adaptation process also requires **prior formant measurements** from the new corpora
- Bottom line: The DeepFormants approach requires **prior formant measurements** from any corpus it is trained on or adapted to
- Can we train a DNN formant tracker without prior measurements? YES!

Our Solution: “FormantNet”

- Unsupervised training: Requires no prior formant measurements
- Model input: 257-point smoothed spectral envelope computed from a 32-msec window of speech collected every 5 msec
- Model output: Formant frequencies, bandwidths, and amplitudes for **all** formants in the input frequency range (0-8 KHz)
  - A model of the **entire** vocal-tract signal
- Formant parameters used to calculate a model spectral envelope
- Loss is calculated as difference between input and model envelopes
- Loss is back-propagated to amend the model weights


Reference Formant features
Predicted Formant features
Neural Network
350 features (LPC + DCT) derived from input waveform
DeepFormants:
• Derived input features fed to neural network
• Neural network predicts formant parameters
• Predicted formant parameters compared to hand-labeled parameters to compute loss
FormantNet:
• Input spectrum fed to neural network
• Fixed logic maps predicted Formant features to predicted spectrum
• Predicted spectrum compared to input spectrum to compute loss
Decoding Logic
Predicted Formant features
Neural Network


Decoding Layers
Bottleneck features
Encoding Layers
Autoencoder:
• Weights are trained in both encoding and decoding layers
• Bottleneck features are learned as a compact set
• Post Hoc discovery needed to interpret learned Bottleneck features
FormantNet:
• Weights are trained in encoding layer only
• Formant features are learned as a compact set
• Fixed logic mapping Formant features to spectrum makes features immediately interpretable
Decoding Logic
Predicted Formant features
Encoding Layers

Approach

- Spectrum level *h* at frequency *f*, for formant parameters *F*, *B*, *A* (from Fant 1970):

$$h_{F,B,A}(f) = \frac{A \times \left(F^2 + {B^2}/{4}\right)}{\sqrt{\left((f-F)^2 + {B^2}/{4}\right) \times \left((f+F)^2 + {B^2}/{4}\right)}}$$

Approach (continued)

- Total spectrum level given formants *1 ... I* and antiformants *1 ... J* :

$$h_{total}(f) = \sum_{i=1}^{I} h_{F_i,B_i,A_i}(f) \div \sum_{j=1}^{J} h_{F_j,B_j,A_j}(f)$$

- Spectrum is converted to decibels:

$$h_{Dec}(f) = 20.0 \times log_{10}\big(0.001 + h_{total}(f)\big)$$

- Loss is mean squared difference between input and calculated spectra, over all spectral frequencies and input frames:

$$Loss = \frac{1}{T \times F} \sum_{t=0}^{T} \sum_{f=0}^{F} \Big(h_{Dec}(f) - h_{Inp}(f)\Big)^2$$

Approach (continued)

- FormantNet requires that the user choose the number of formants and antiformants to model, prior to model training
  - Should be based on the average number expected within the entire input frequency range (e.g. 0-8000 Hz), for the talker population in the corpus
  - Too few: The model spreads out the formants to cover the frequency range
  - Too many: The model will insert extra formants where they don’t belong

- For TIMIT (0-8 KHz): best models had 6 formants, 1 antiformant
  - Generates a model of the *entire* vocal-tract signal

Implementation Details

- “Train” portion of the TIMIT dataset split in two:
  - 4140-file training set
  - 480-file validation set
- Hand-labeled formant measurements not used for training
- Trained for max. 200 epochs; model w/best validation loss chosen
- Evaluated on the 192-file hand-labeled “test” set
- Different architectures evaluated:
  1) Convolutional neural network (CNN)
  2) LSTM, a type of Recurrent neural network (RNN)
  3) Bidirectional LSTM (BLSTM)

InterSpeech 2021 results: Comparison of model architectures

MAE, all segments				
Model	Mean	F1	F2	F3
LSTM1	**114**	**100**	**115**	**126**
BLSTM1	114	102	115	126
LSTM3	131	102	146	146
CNN3	129	105	117	165

MAE, vowels				
Model	Mean	F1	F2	F3
LSTM1	**76**	**64**	**75**	**90**
BLSTM1	77	65	77	90
LSTM3	81	65	81	96
CNN3	86	65	81	111

LSTM1: 1 LSTM layer (512 units) + 1 Dense output layer (20 units)

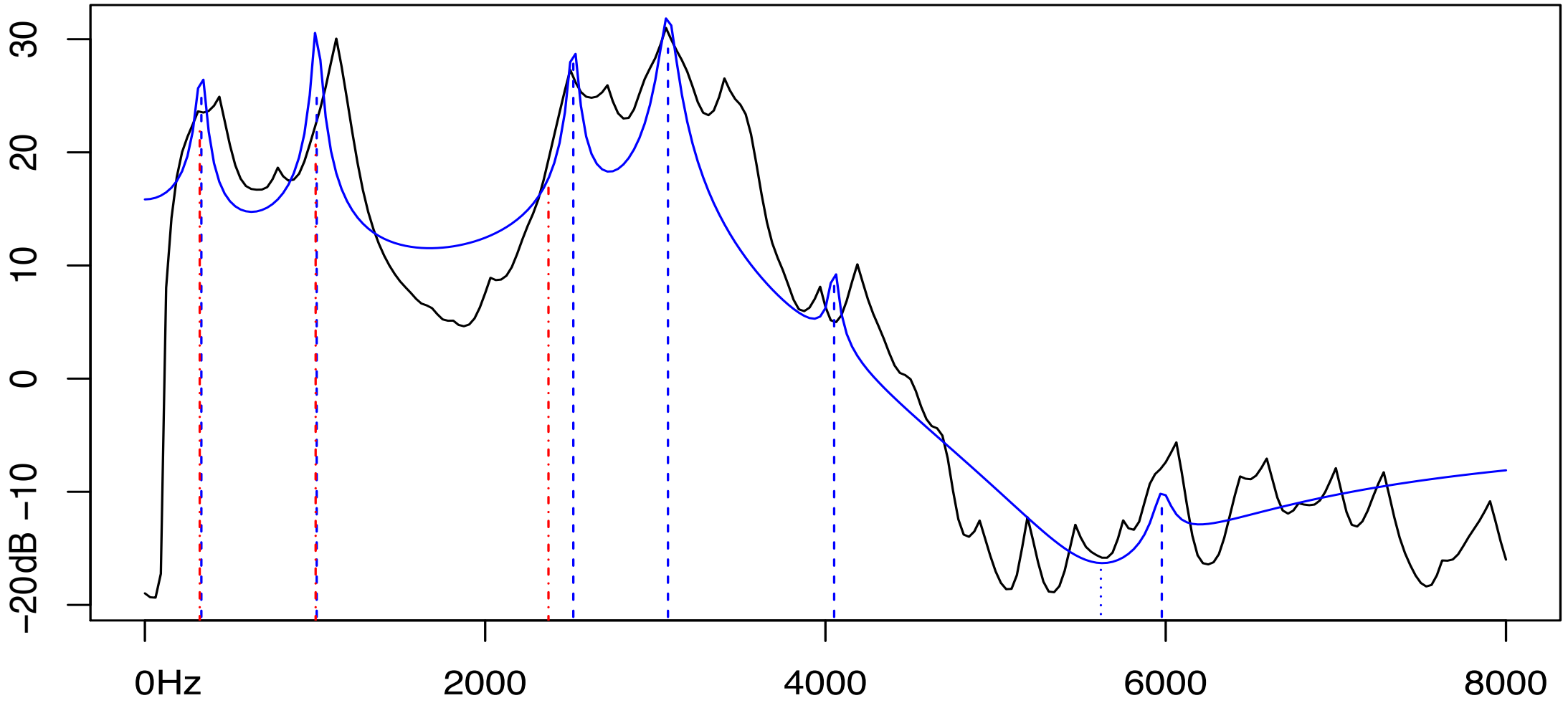
- Performed just as well as larger LSTM, BLSTM, and CNN models

IS2021 results: Comparison with other trackers

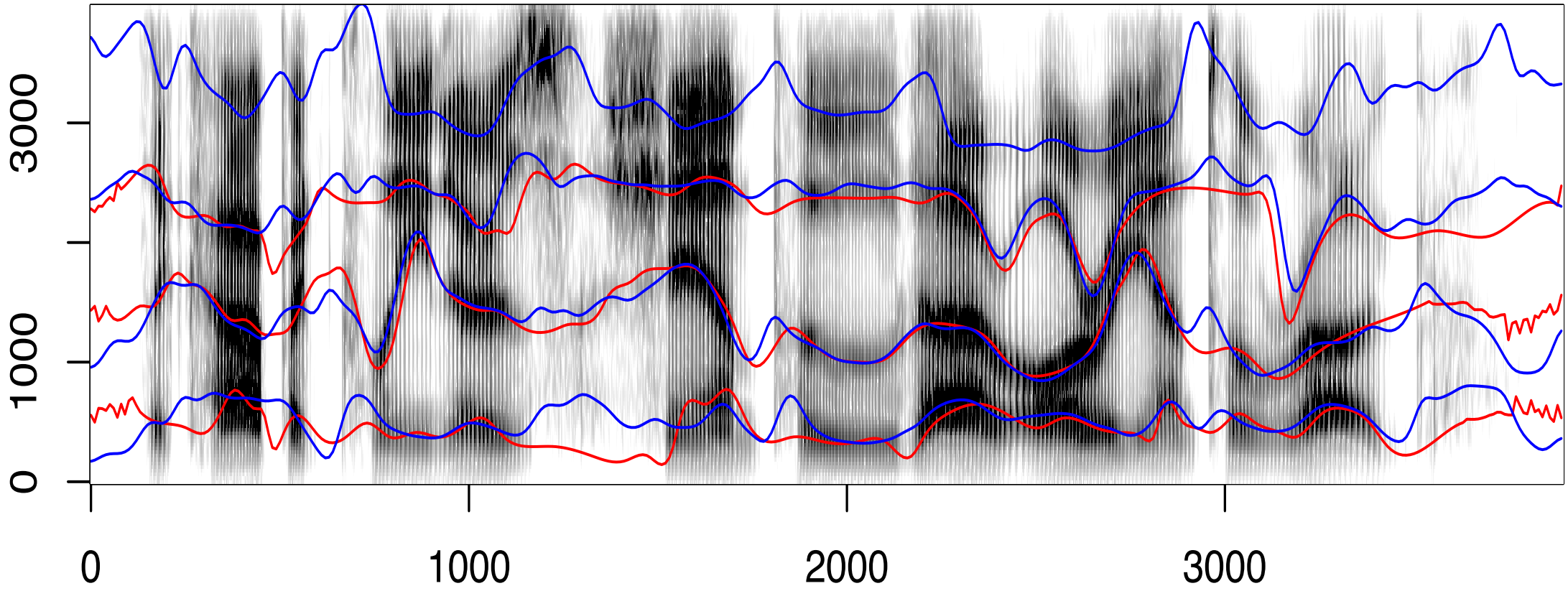
RMSE, all segments				
Model	Mean	F1	F2	F3
LSTM1	173	143	177	**195**
Deep Formants	**163**	118	**169**	204
DF-RCNN	173	127	180	213
KARMA	220	**114**	226	320

MAE, vowels				
Model	Mean	F1	F2	F3
LSTM1	**76**	64	75	**90**
Deep Formants	82	54	81	112
DF-RCNN	78	**53**	**72**	108
MSR 2006	98	64	105	125
WaveSurfer	106	70	94	154
Praat	209	130	230	267

Example input vs. model spectra: /u/



Example spectrogram



Red: Hand-measured frequencies (F1-F3)

Blue: Model frequencies (F1-F4)

New advances

1) Delta-frequency (Δf) loss
2) Source function modeling

Delta-frequency (Δf) Loss

- Added loss for change in predicted frequency of each formant from one frame to the next (N is number of resonances):

$$\Delta f Loss = \frac{1}{T \times N} \sum_{t=1}^{T} \sum_{n=1}^{N} f_{n,t} - f_{n,t-1}$$

- Total loss is regular loss plus **weighted** sum of Δf loss:

$$Loss = SpectrumLoss + \boldsymbol{W} \times \Delta f Loss$$

- Best value of $\boldsymbol{W}$: 0.05 to 0.15

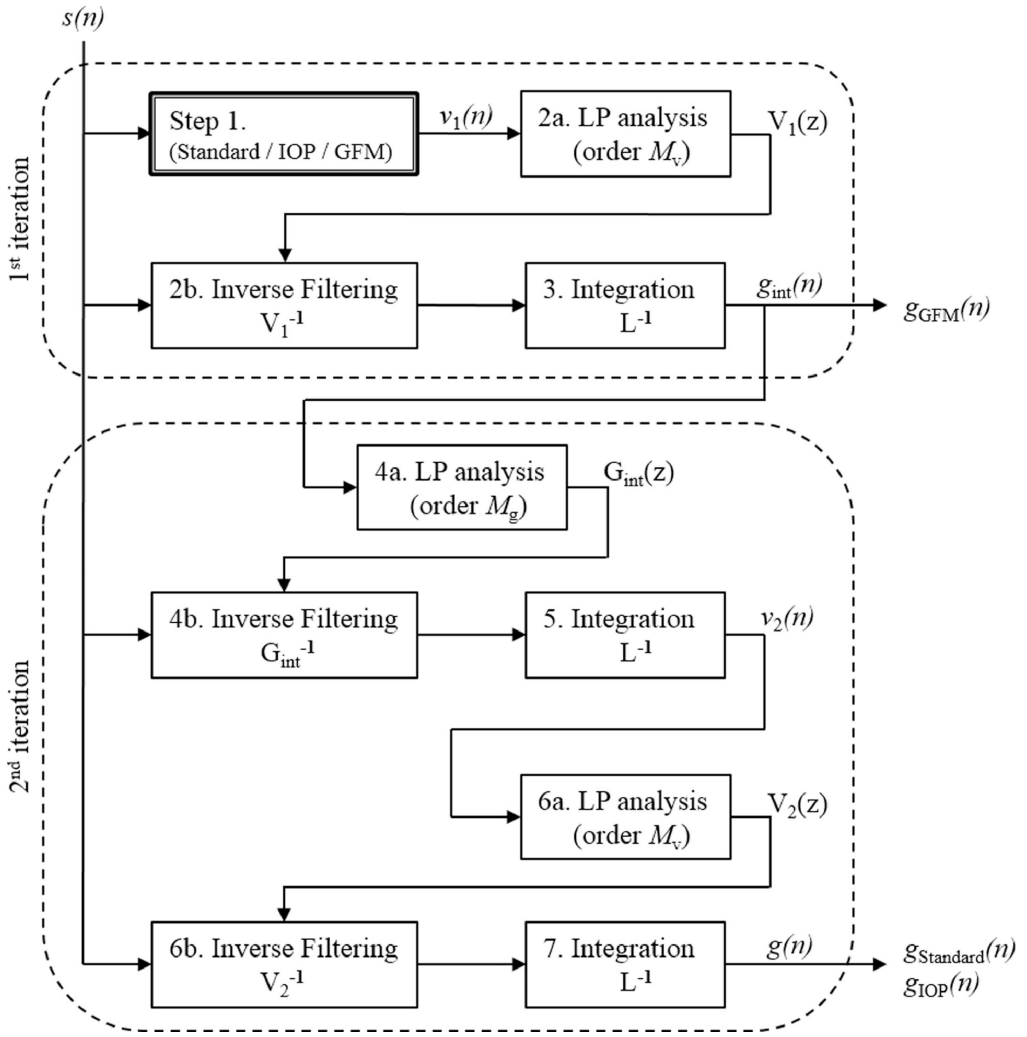
Source function modeling

Iterative Adaptive Inverse Filtering (IAIF) (Alku 1991, 1992): complex algorithm to separate and remove the glottal excitation function from the vocal tract transfer function

Is the IAIF interfering with measurement of F1?

We compared IAIF to 2 source modeling alternatives:

1) Pre-emphasis ($S_t = s_t - 0.98\, s_{t-1}$)
2) No source model



(from Mokhtari et al. 2018)

Summary: Δf Loss and Source Modeling

MAE, all segments					
Source	*W*	Mean	F1	F2	F3
IAIF	0	114	100	115	126
PE	0	118	92	123	140
None	0	115	94	121	130
IAIF	.05	105	97	99	118
PE	.15	**98**	**88**	**98**	**109**
None	.10	99	89	98	111

MAE, vowels					
Source	*W*	Mean	F1	F2	F3
IAIF	0	76	64	75	90
PE	0	76	62	76	94
None	0	76	61	80	92
IAIF	.05	75	62	72	92
PE	.15	71	58	**70**	**86**
None	.10	**70**	**51**	72	87

Comparison with other trackers

RMSE, all segments				
Model	Mean	F1	F2	F3
FNet PE15	**150**	125	146	**174**
FNet NS10	152	133	**144**	176
LSTM1	173	143	177	195
Deep Formants	163	118	169	204
DF-RCNN	173	127	180	213
KARMA	220	**114**	226	320

MAE, vowels				
Model	Mean	F1	F2	F3
FNet PE15	71	58	**70**	**86**
FNet NS10	**70**	**51**	72	87
LSTM1	76	64	75	90
Deep Formants	82	54	81	112
DF-RCNN	78	53	72	108
MSR	98	64	105	125
WaveSurfer	106	70	94	154
Praat	209	130	230	267

Further experiments

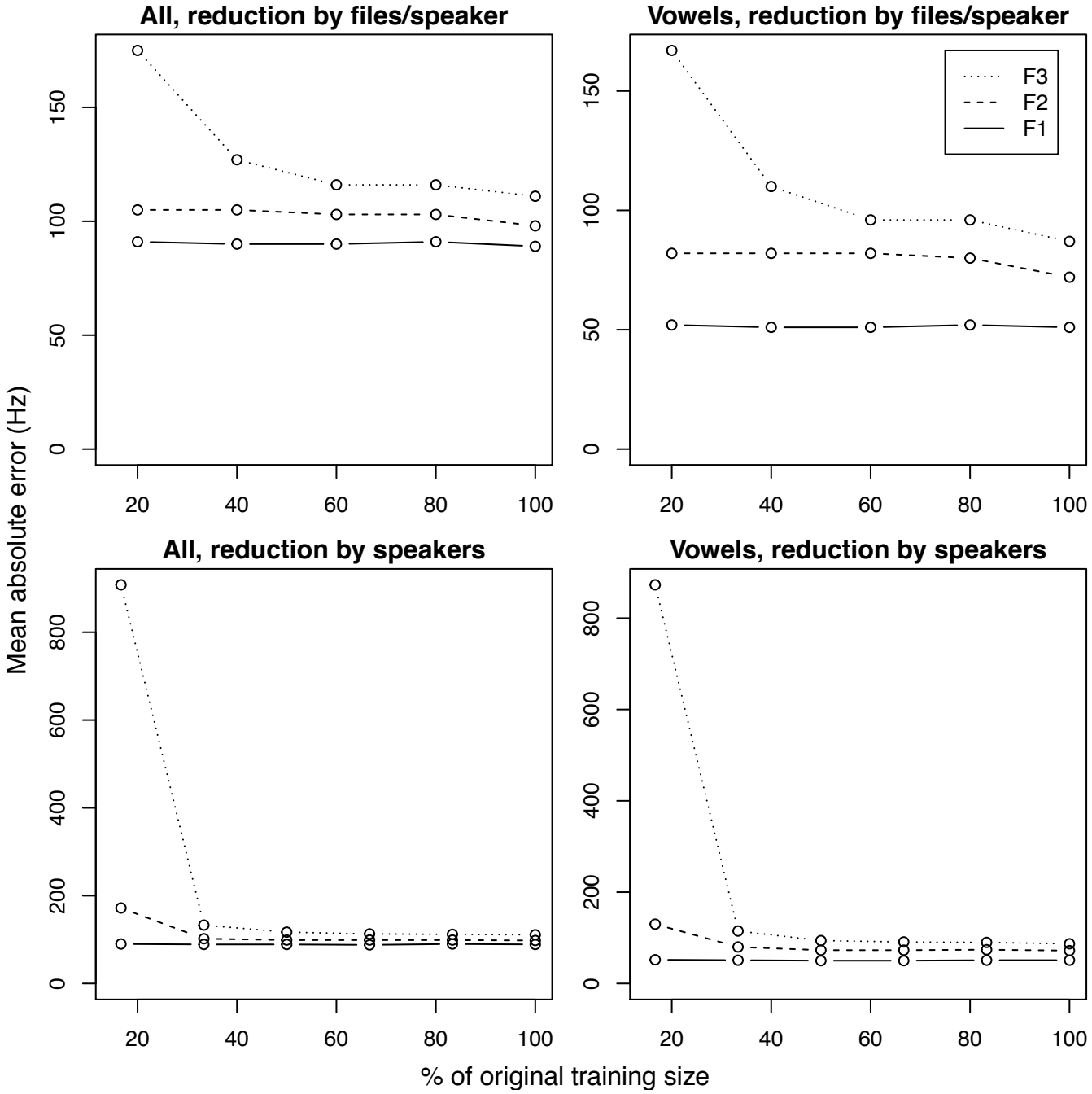
- Effect of training set size
- Testing the models on a different corpus

Effect of Training Size

1. Reduction by # of files/speaker:
   - From 10 files/speaker (4140 total)
   - To 2 files/speaker (828 total)

2. Reduction by # of speakers
   - From 414 speakers (4140 files)
   - To 69 speakers (690 files)

Validation set also reduced proportionally

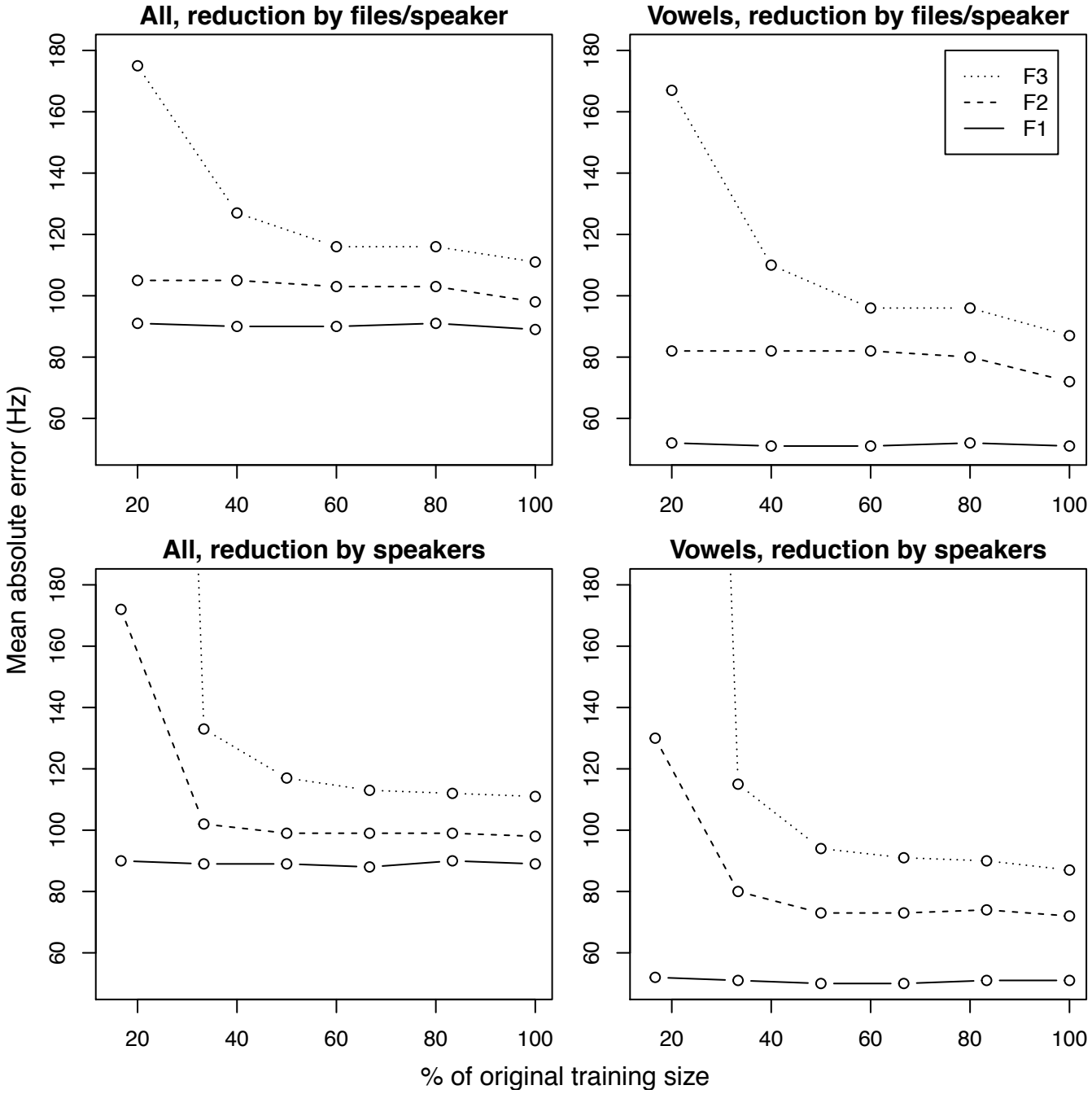


Effect of Training Size

1. Reduction by # of files/speaker:
   - From 10 files/speaker (4140 total)
   - To 2 files/speaker (828 total)

2. Reduction by # of speakers
   - From 414 speakers (4140 files)
   - To 69 speakers (690 files)

Validation set also reduced proportionally



Testing and Adapting the TIMIT model on the Hillenbrand corpus

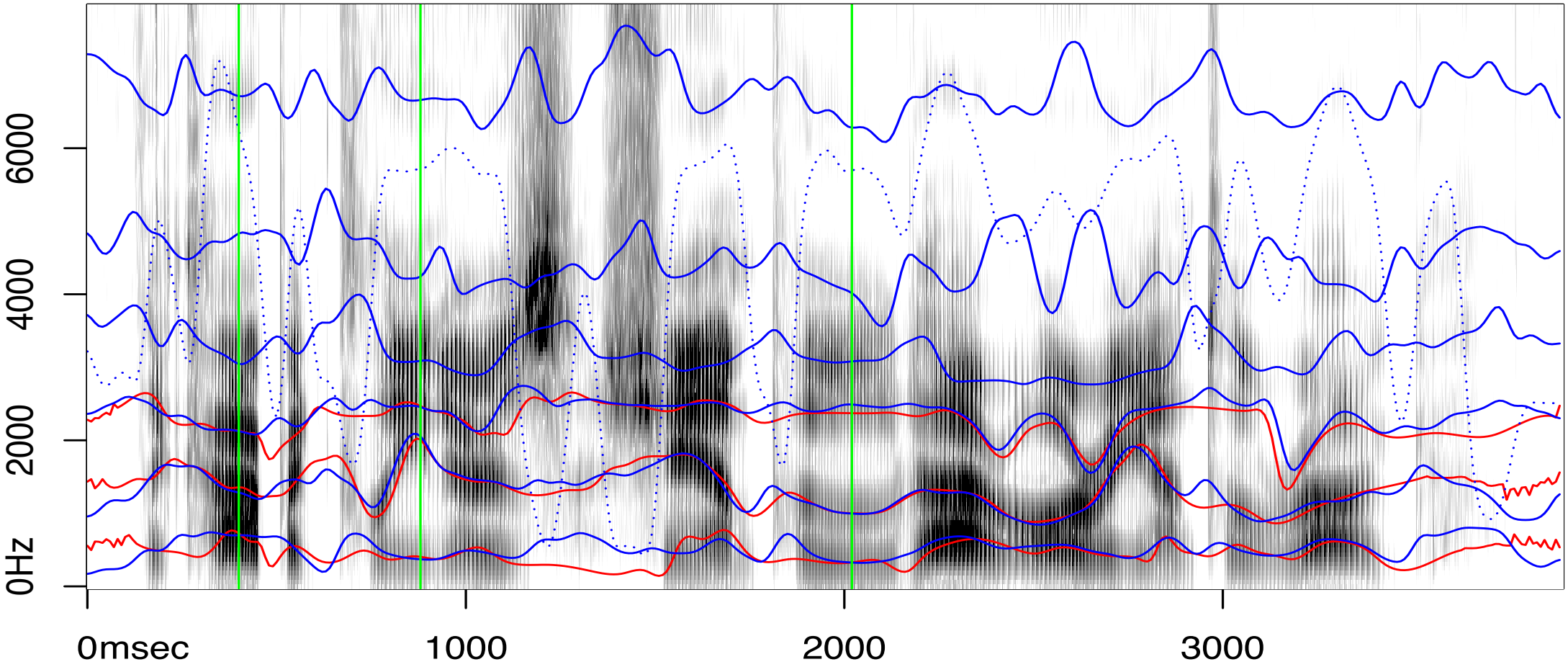
- Hillenbrand dataset:
  - 1668 vowels
  - Mix of adults and 10-12yo kids
- Like DeepFormants, our TIMIT models do worse on other corpora (though adaptation helps)
- With a large enough dataset, it's probably better to train from scratch – no hand-labels needed!

MAE, vowels				
Model	Mean	F1	F2	F3
FormantNet Hillenbrand	237	47	291	376
FN TIMIT	109	33	133	161
FN TIMIT adapted	**73**	**29**	**85**	**107**
DeepFormants	143	71	160	131
DeepFormants adapted	83	36	100	116

Conclusions

- “FormantNet” algorithm can be used to train a tracker on a corpus without prior hand-obtained formant measures for training
  - Works best on a reasonably large corpus
- Lower error than popular LPC-based trackers and supervised DNNs
- Source modeling is a continued avenue of research
- Produces a model of the **entire** vocal-tract signal: Bandwidths, amplitudes, and higher formants
  - Useful for e.g. formant speech synthesis
- Available now: github.com/NemoursResearch/FormantNet

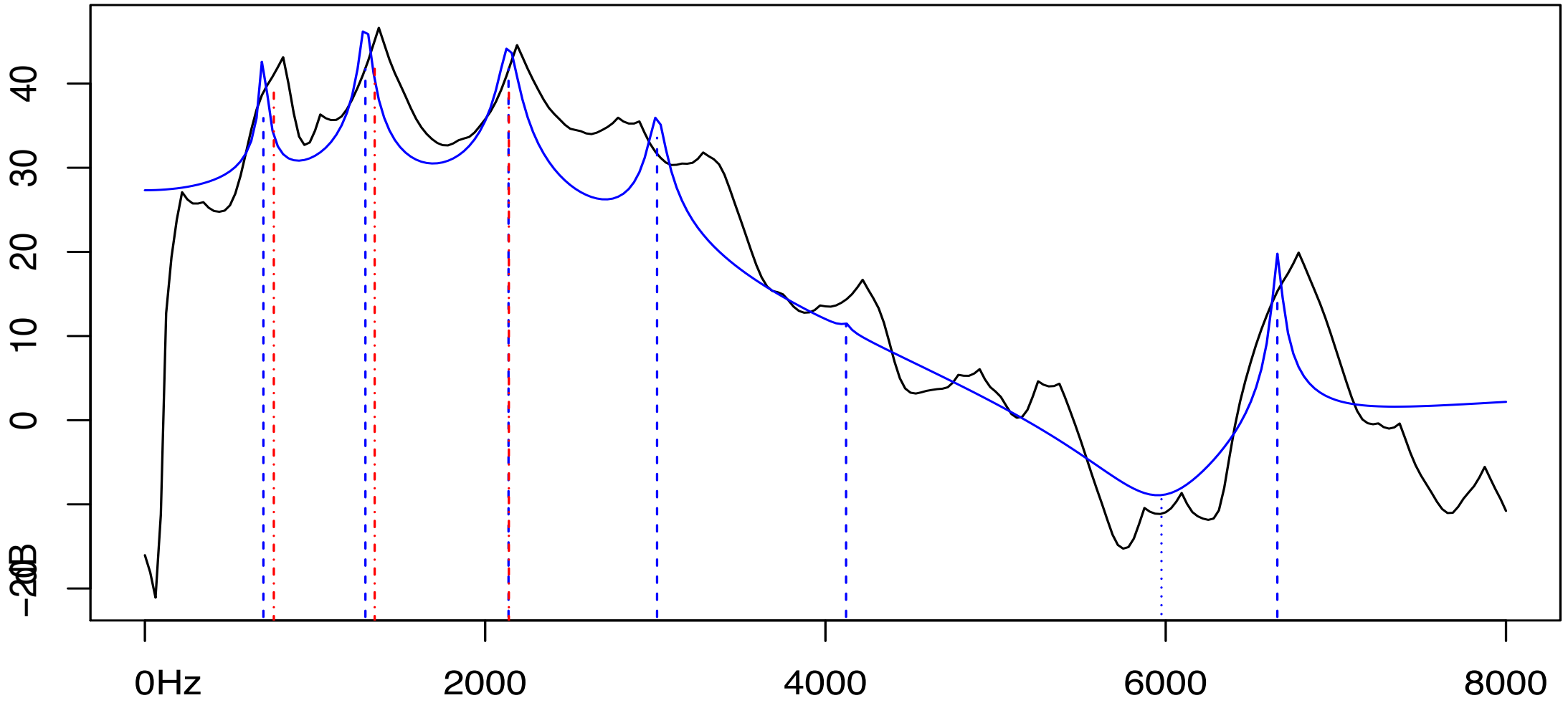
Example spectrogram



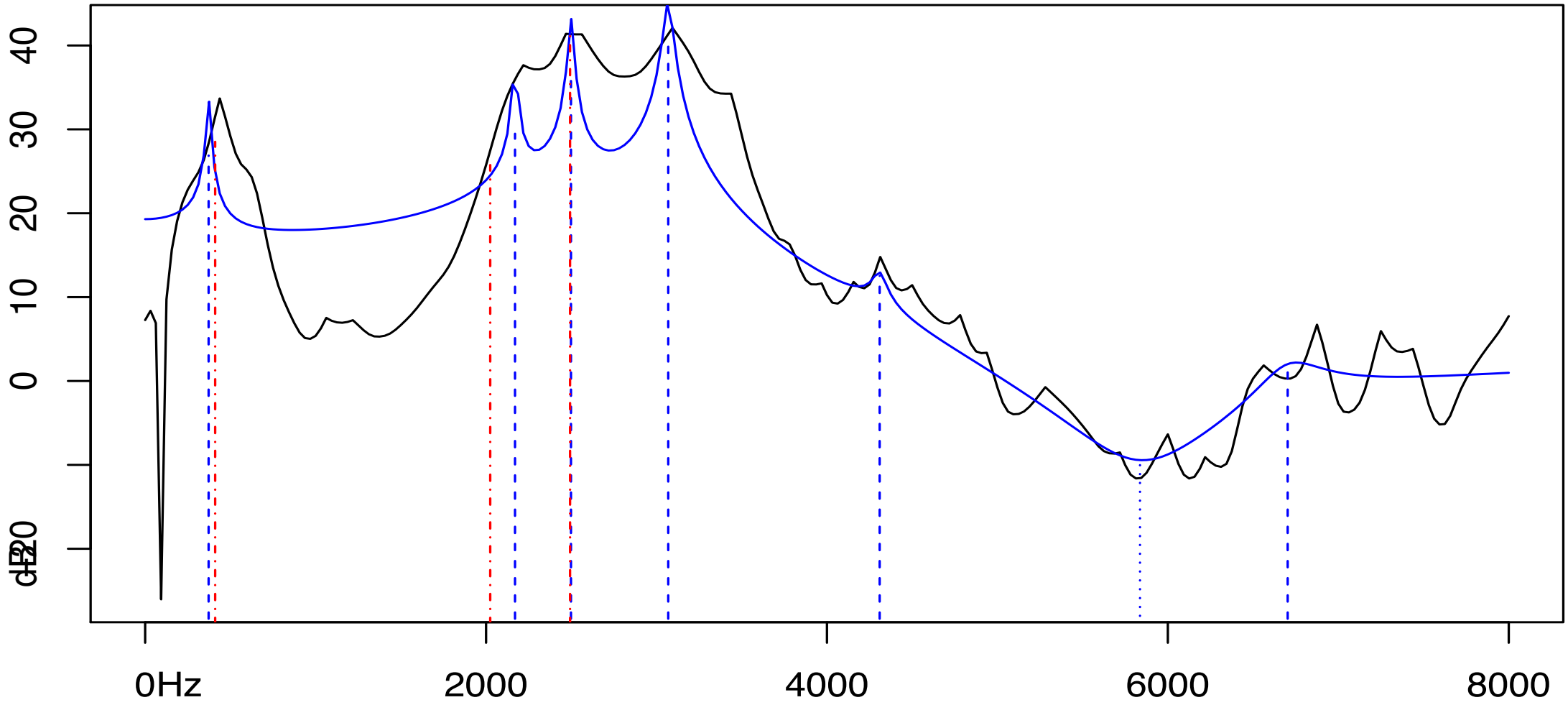
Red: Hand-measured frequencies (F1-F3)

Blue: Model frequencies (F1-F4)

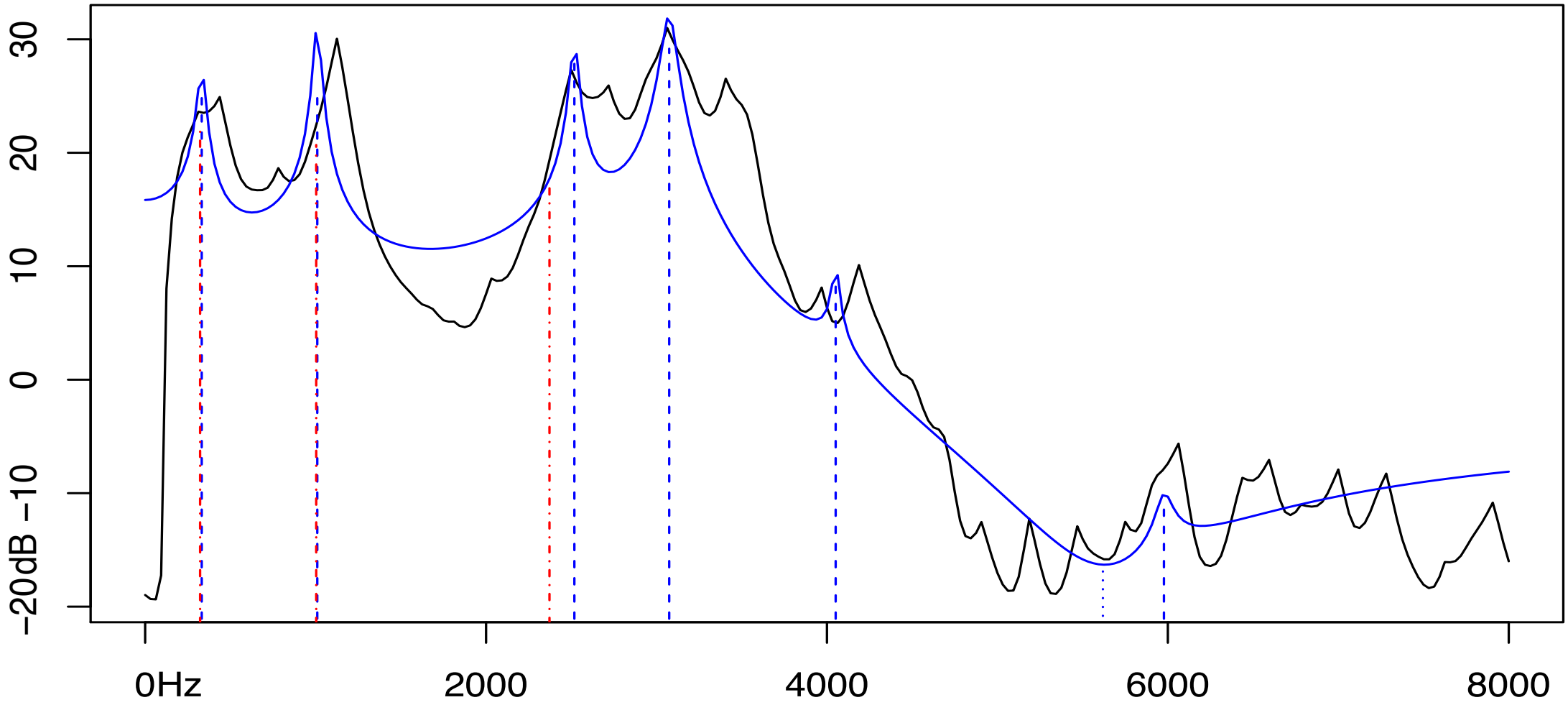
Example spectra: /a/



Example spectra: /i/



Example spectra: /u/



Cited References

P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive filtering,” *Speech Communication*, vol. 19, pp. 459–476, 1992.

P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341-345, 2002.

C. G. Clopper and T. N. Tamati, “Effects of local lexical competition and regional dialect on vowel production,” *Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 1-4, 2014.

L. Deng, L. J. Lee, H. Attias, and A. Acero, “A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances,” in *Proceedings of ICASSP 2004—IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. I-557, 2004.

L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, “A database of vocal tract resonance trajectories for research in speech processing,” in *Proceedings of ICASSP 2006—IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. I-I, 2006.

Y. Dissen, J. Goldberger, and J. Keshet, “Formant estimation and tracking: A deep learning approach,” *Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 642-653, Feb. 2019.

G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1970.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *The DARPA TIMIT acoustic-phonetic continuous speech corpus.* Linguistic Data Consortium, 1993.

J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099- 3111, 1995.

J. Lilley and H. T. Bunnell, “Unsupervised Training of a DNN-based Formant Tracker,” to appear in *Proceedings INTERSPEECH 2021*.

D. D. Mehta, D. Rudoy, and P. J. Wolfe, “Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking,” *Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732-1746, 2012.

P. Mokhtari, B. Story, P. Alku, and H. Ando, “Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production”, *Speech Communication*, vol. 104, pp. 24-38, Nov. 2018.

F. Schiel and T. Zitzelberger, “Evaluation of automatic formant trackers,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan, May 7-12, 2018, pp. 2843-2848.

K. Sjölander and J. Beskow, “WaveSurfer – an open source speech tool,” in *Proceedings INTERSPEECH 2000 – 1st Annual Conference of the International Speech Communication Association*, Beijing, China, Oct. 16-20, 2000, pp. 464–467.

More references

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org. Version 2.3.

S. Fulop and C. Shadle, "Automated formant tracking using reassigned spectrograms," *Journal of the Acoustical Society of America*, vol. 143, no. 3, p. 1870, 2018. Doi:10.1121/1.5036138

D. Gowda, M. Airaksinen, and P. Alku, "Quasi closed phase analysis of speech signals using time varying weighted linear prediction for accurate formant tracking," in *Proceedings of ICASSP 2016—IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 4980-4984, 2016.

D. Gowda, M. Airaksinen, and P. Alku, "Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation," *Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1542-1553, 2017. Doi: 10.1121/1.5001512

D. Gowda and P. Alku, "Time-varying quasi-closed-phase weighted linear prediction analysis of speech for accurate formant detection and tracking," in *Proceedings INTERSPEECH 2016 –17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 8-12, 2016, pp. 1760–1764.

G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," *Advances in Neural Information Processing Systems* 6, pp. 3-10, 1994.

J. N. Holmes, "Research report - formant synthesizers - cascade or parallel," *Speech Communication*, vol. 2, pp. 251-273, 1983.

M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp, 233-243, 1991. Doi: 10.1002/aic.690370209

J. P. Olive, "Automatic Formant Tracking by a Newton-Raphson Technique," *Journal of the Acoustical Society of America*, vol. 50, pp. 661-670, 1971. Doi: 10.1121/1.1912681

M. A. Ramírez, "Hybrid autoregressive resonance estimation and density mixture formant tracking model," *IEEE Access*, vol. 6, pp. 30217-30224, 2018.

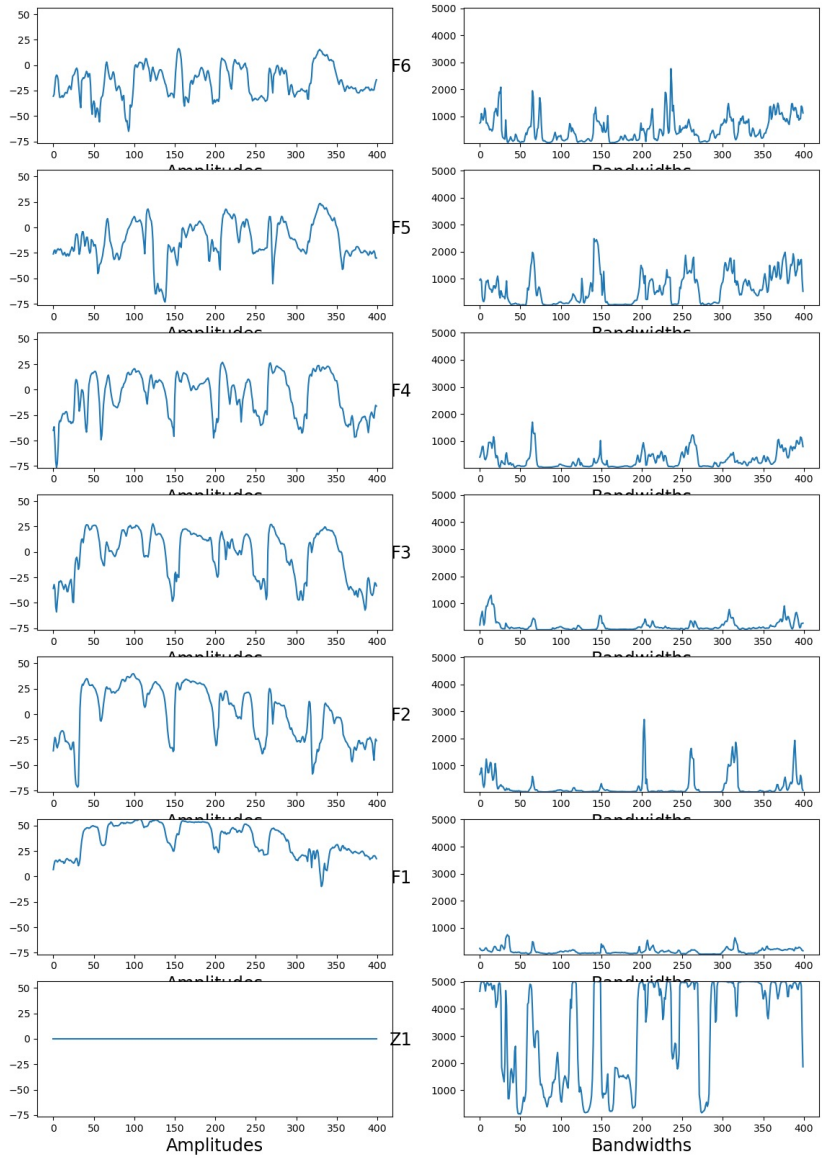
M. Scheffer, *Advanced Speech Signal Processor (libassp).* http://www.sourceforge.net/projects/libassp

K. Sjölander, *Snack-Sound-Toolkit.* http://www.speech.kth.se/snack

R. Sharma, L. V. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, and S. R. M. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review," *Speech Communication*, vol. 88, pp. 39-64, 2017. Doi: 10.1016/j.specom.2016.12.004

B. H. Story and K. Bunton, "Formant measurement in children's speech based on spectral filtering," *Speech Communication*, vol. 76, pp. 93-111, 2016.

Amplitudes and Bandwidths



Including test set in training material
(no source model, *W*=.10)

MAE, all segments					
Training set	n	Mean	F1	F2	F3
All train	4140	99	89	98	111
Partial train + test	4140	99	88	98	111
All train + test	5820	99	88	98	111
Train + test + val	6300	99	88	98	110

MAE, vowels					
Training set	n	Mean	F1	F2	F3
All train	4140	70	51	72	87
Partial train + test	4140	71	51	73	88
All train + test	5820	72	51	75	90
Train + test + val	6300	71	51	74	87

Testing and Adapting the TIMIT model on the Hillenbrand corpus

- Hillenbrand dataset:
  - 1668 single-syllable words
  - Only vowel formants measured
  - Mix of adults and 10-12yo kids
- Small dataset: adaptation helps!
- With a large enough dataset, it's probably better to train from scratch

MAE, vowels				
Model	Mean	F1	F2	F3
Hillenbrand	237	47	291	376
TIMIT NS10	109	33	133	161
TIMIT NS10 adapted	73	**29**	85	107
TIMIT PE15	123	43	140	188
TIMIT PE15 adapted	**59**	47	**48**	**84**
DeepFormants	143	71	160	131
DeepFormants adapted	83	36	100	116

Application: Sensitivity vs. Praat

- Ran t-tests comparing each formant (F1-F3) of each pair of dialects (8) in TIMIT (84 total comparisons)
- Compared hand labels (ground truth), FormantNet, and Praat
- FormantNet correctly found 8 more significant differences than Praat
- Accuracy: FN 83%, Praat 73%

	FormantNet		Praat	
Hand Labels	sig.	not sig.	sig.	not sig.
sig.	68	9	60	17
not sig.	5	2	6	1

(Thanks to Maxwell Hope)

Δf Loss: Effect of weight W (IAIF model)

MAE, all segments				
W	Mean	F1	F2	F3
0	114	100	115	126
0.01	111	102	107	123
0.05	**105**	97	**99**	**118**
0.10	114	**91**	103	149
0.15	152	97	157	204
0.20	321	114	218	631
0.25	192	112	223	242

MAE, vowels				
W	Mean	F1	F2	F3
0	76	64	75	90
0.01	**75**	65	74	**87**
0.05	**75**	62	**72**	92
0.10	88	**58**	74	133
0.15	132	60	134	202
0.20	295	77	182	624
0.25	177	77	205	250

∆f loss (Pre-emphasis model)

MAE, all segments				
w	Mean	F1	F2	F3
0	118	92	123	140
.05	99	91	99	108
.10	98	90	97	108
.15	98	88	98	109
.20	105	90	98	128
.25	121	92	109	164

MAE, vowels				
w	Mean	F1	F2	F3
0	76	61	76	92
.05	72	62	71	84
.10	71	60	70	84
.15	71	58	70	86
.20	78	58	70	106
.25	98	59	83	151

∆f loss (No source model)

MAE, all segments				
w	Mean	F1	F2	F3
0	115	94	121	130
.05	102	93	105	109
.10	99	89	98	111
.15	113	89	102	149
.20	123	89	105	176
.25	131	89	112	193

MAE, vowels				
w	Mean	F1	F2	F3
0	76	52	80	94
.05	71	51	76	85
.10	70	51	72	87
.15	87	51	78	133
.20	99	51	81	166
.25	109	51	89	186

Source Modeling (no Δf Loss)

MAE, all segments				
Source	Mean	F1	F2	F3
IAIF	**113**	101	**114**	**126**
Pre-emphasis	115	94	121	130
None	118	**92**	123	140

MAE, vowels				
Source	Mean	F1	F2	F3
IAIF	77	65	**75**	**90**
Pre-emphasis	**76**	**52**	80	94
None	76	61	76	92

Source Modeling (w=5)

MAE, all segments				
Model	Mean	F1	F2	F3
IAIF	105	97	99	118
Pre-emphasis	102	93	105	109
None	99	91	99	108

MAE, vowels				
Model	Mean	F1	F2	F3
IAIF	75	62	72	92
Pre-emphasis	71	51	76	85
None	72	62	71	84

Source Modeling (w=10)

MAE, all segments				
Model	Mean	F1	F2	F3
IAIF	105	90	100	123
Pre-emphasis	99	89	98	111
None	98	90	97	108

MAE, vowels				
Model	Mean	F1	F2	F3
IAIF	77	57	71	101
Pre-emphasis	70	51	72	87
None	71	60	70	84

Source Modeling (w=15)

MAE, all segments				
Model	Mean	F1	F2	F3
lAIF	152	97	157	204
Pre-emphasis	113	89	102	149
None	98	88	98	109

MAE, vowels				
Model	Mean	F1	F2	F3
lAIF	132	60	134	202
Pre-emphasis	87	51	78	133
None	71	58	70	86