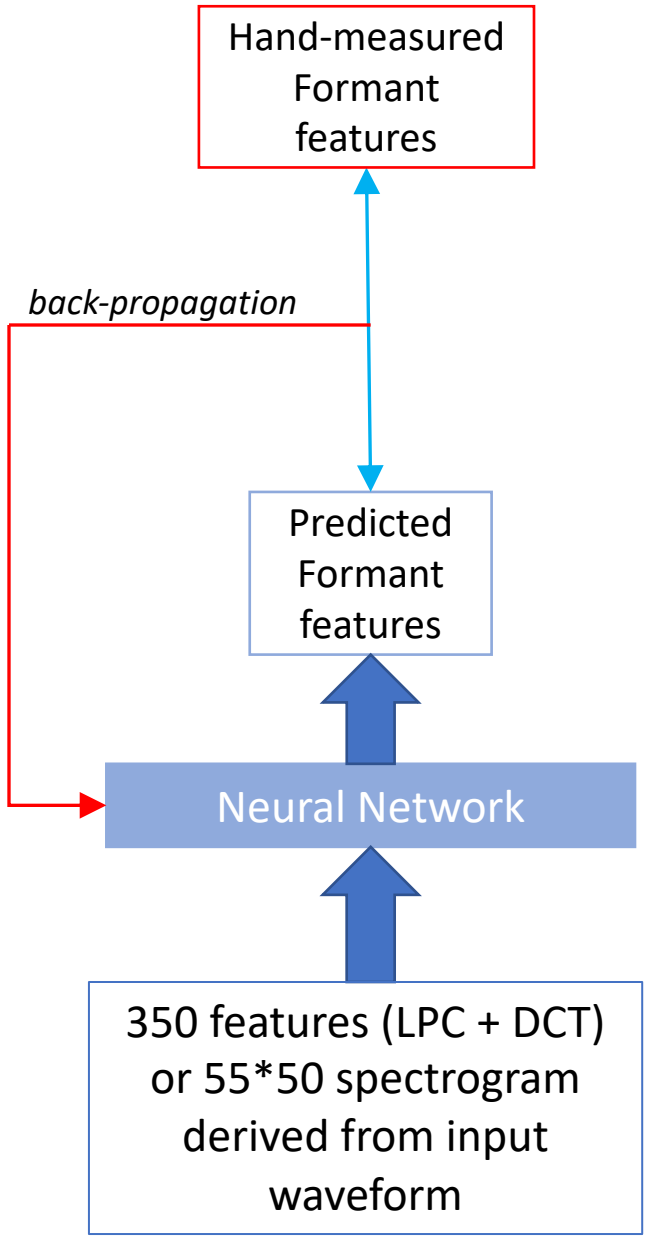
# Unsupervised Training of a DNN-based Formant Tracker

Jason Lilley and H. Timothy Bunnell
Nemours Biomedical Research
Wilmington, DE, USA

InterSpeech 2021 in Brno, Czechia
August 31, 2021

# 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 corpus of many speakers

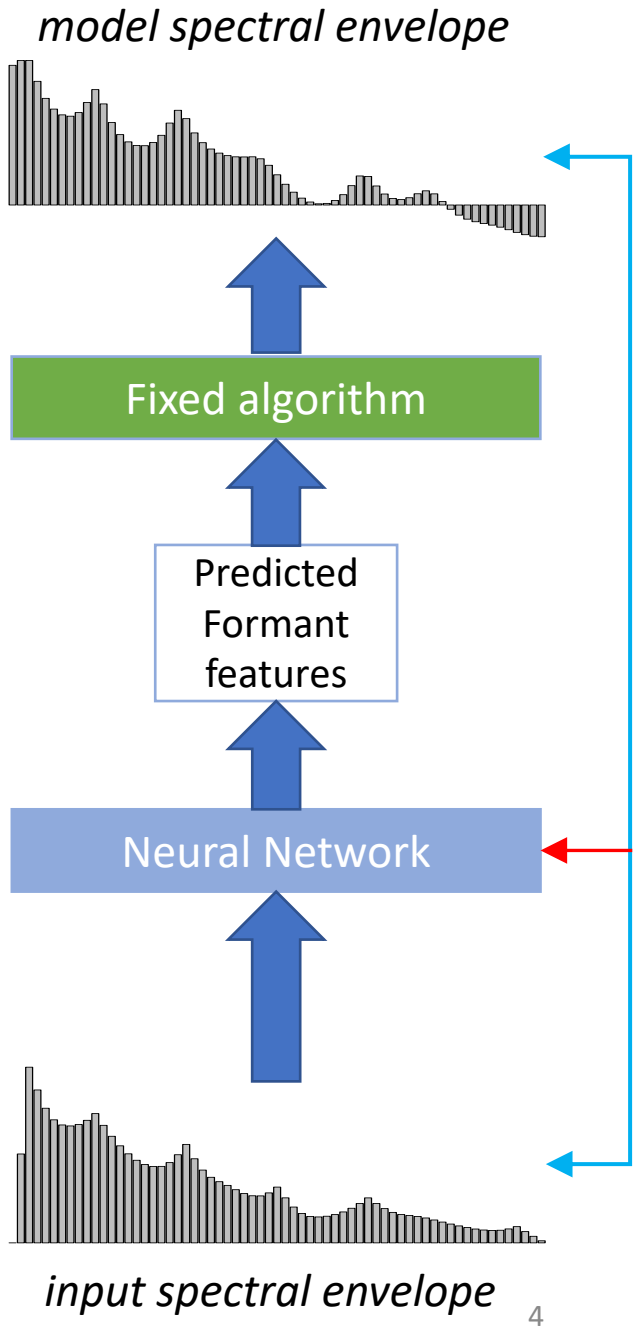


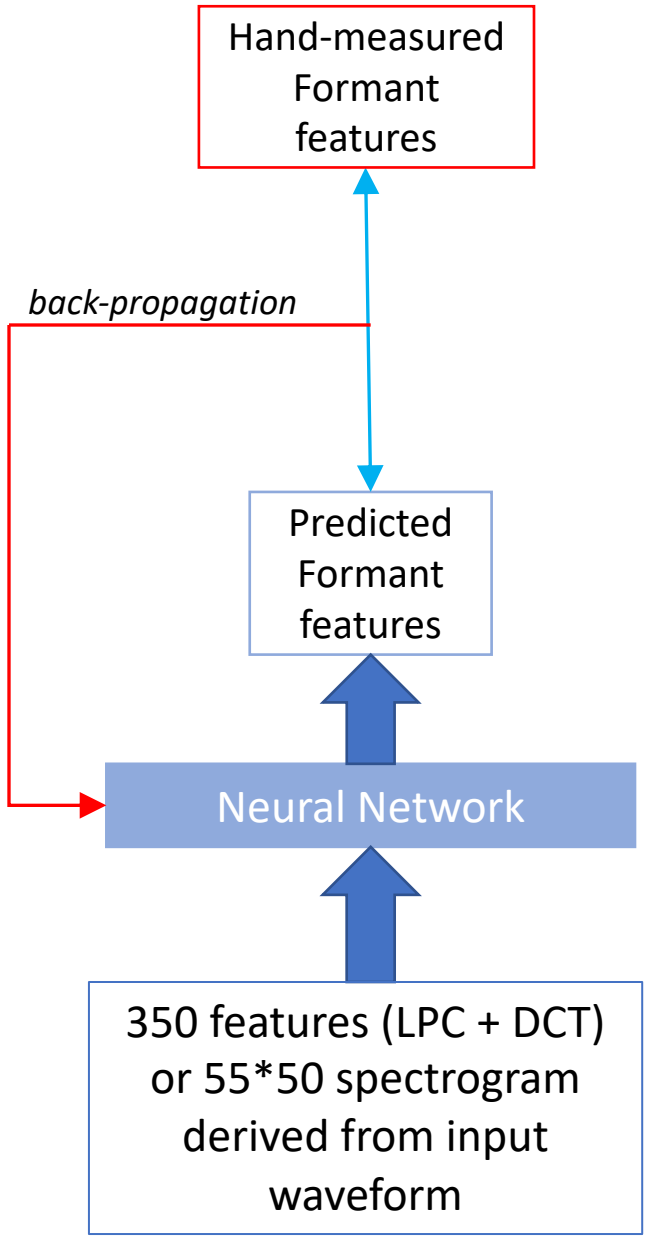

# Dissen et al. (2019): a supervised DNN formant tracker ("DeepFormants")

- Trained neural networks to estimate formant frequencies F1-F3 of the TIMIT corpus
  - Trained on the hand-measured F1-F3 frequencies of a 538-file subset (Deng et al. 2004) of the TIMIT corpus
  - 2 models: LSTM ("**DF-RNN**"), RNN-CNN ("**DF-RCNN**")
- Results generally superior to other methods on the TIMIT corpus (e.g. Praat, WaveSurfer, MSR, KARMA), but higher errors on other corpora w/o adaptation
- Training (or adaptation) requires **prior formant measurements** from the corpus to be tracked

# Our Solution: "FormantNet"

- **Unsupervised** training: Requires no prior formant measurements
- Model input: 257-point smoothed **spectral envelope** per frame (32-msec window)
- Model output: Frequencies, bandwidths, and amplitudes for **all** formants in the input frequency range (0-8 KHz)
  - Number of formants to model is selected before training
  - Produces 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



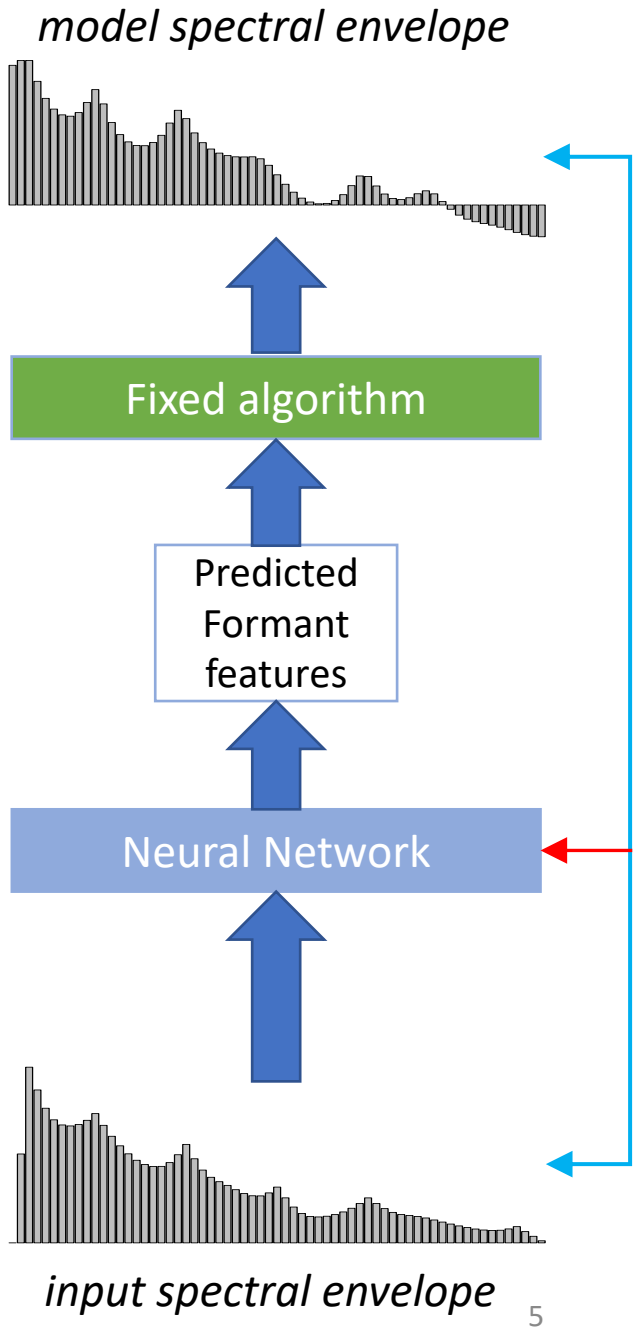


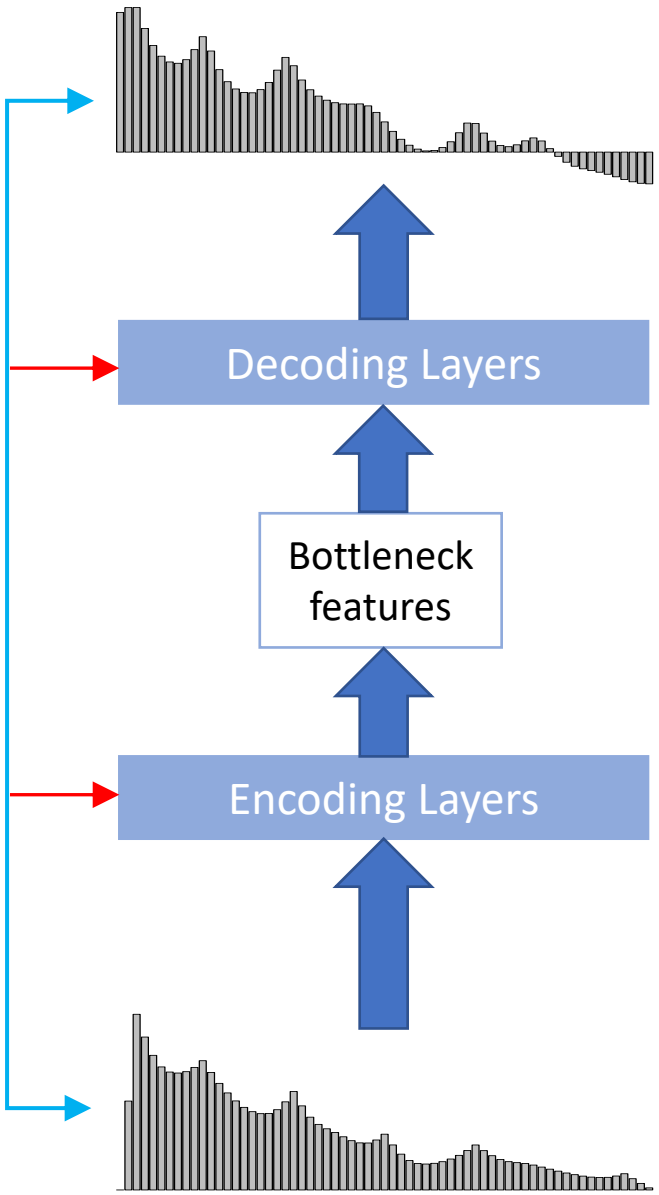

**Dissen et al. (2019):**

- 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



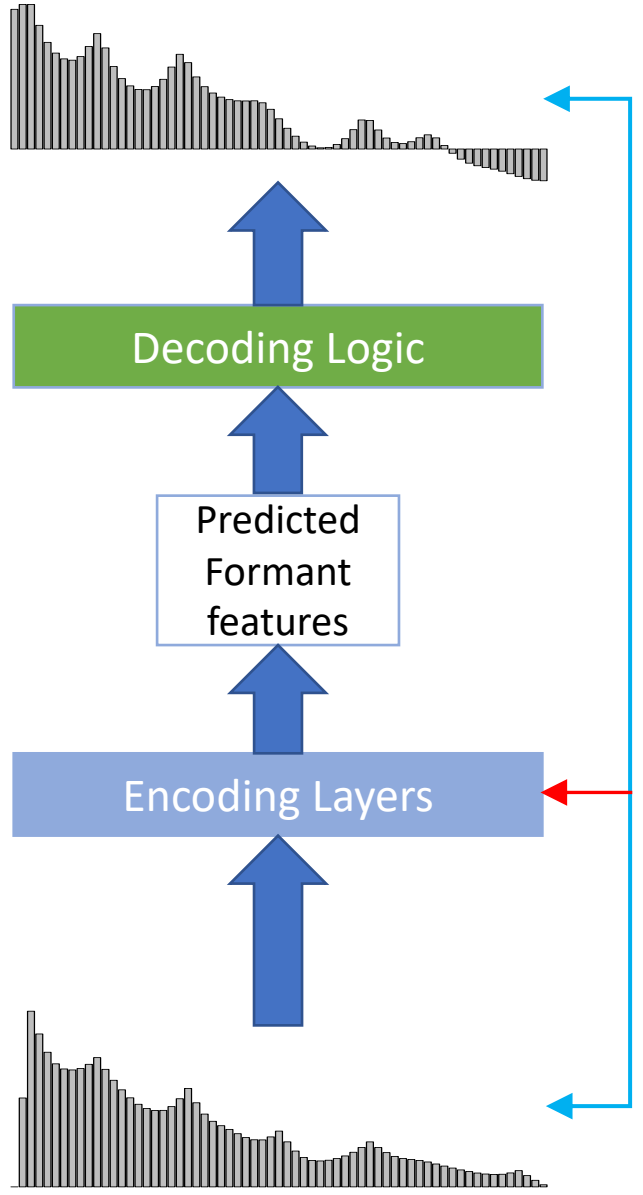



**Autoencoder:**

- 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**



# Approach

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

$$h_{F,B,A}(f) = \frac{A \times \left(F^2 + {B^2}/{4}\right)}{\sqrt{\left((f-F)^2 + {B^2}/{4}\right) \times \left((f+F)^2 + {B^2}/{4}\right)}} \qquad (1)$$

# Approach (continued)

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

$$h_{total}(f) = \sum_{i=1}^{I} h_{F_i,B_i,A_i}(f) \div \sum_{j=1}^{J} h_{F_j,B_j,A_j}(f) \qquad (2)$$

- Spectrum is converted to decibels:

$$h_{Dec}(f) = 20.0 \times log_{10}\big(0.001 + h_{total}(f)\big) \qquad (3)$$

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

$$Loss = \frac{1}{T \times F} \sum_{t=0}^{T} \sum_{f=0}^{F} \Big(h_{Dec}(f) - h_{Inp}(f)\Big)^2 \qquad (4)$$

# 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)

# Results: Comparison of model architectures

- Test set: Hand-labeled VTR-TIMIT (Deng et al. 2004)

| 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)

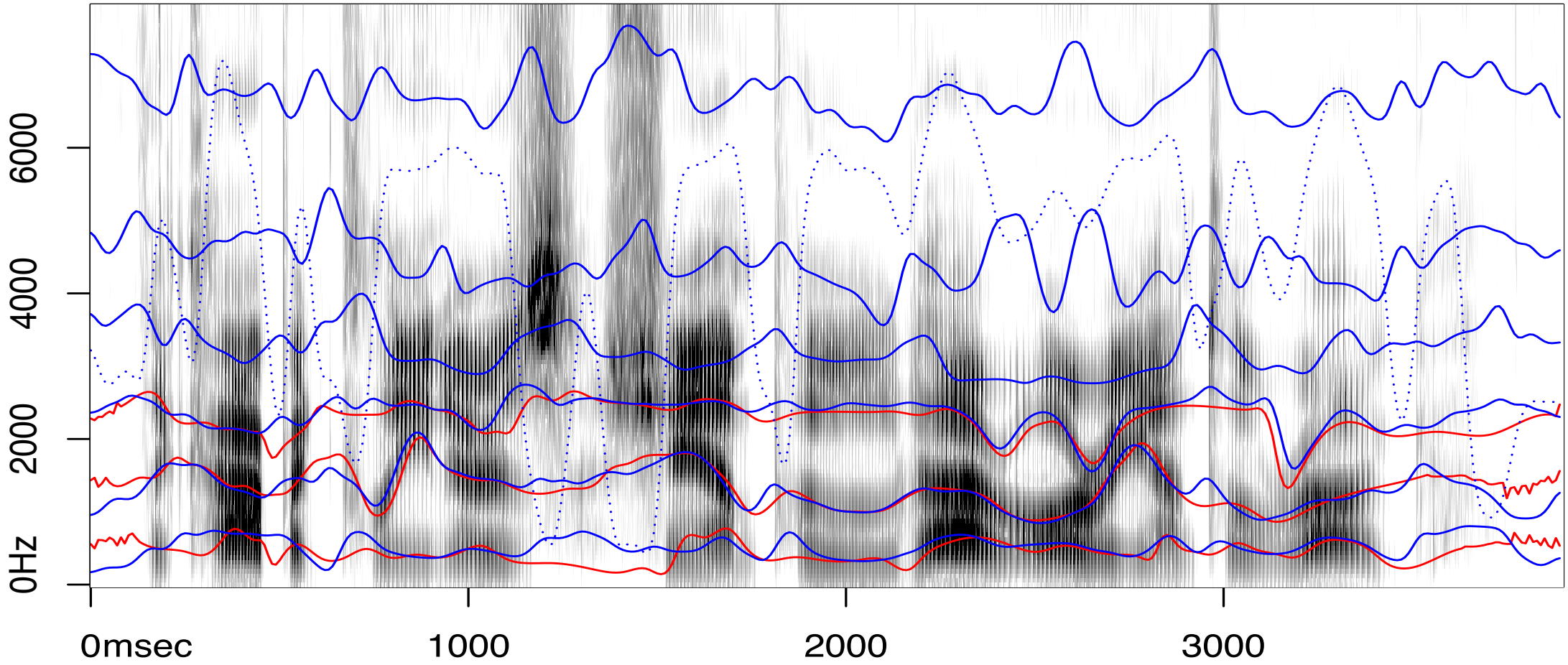
- Performed just as well as larger LSTM, BLSTM, and CNN models

# Results: Comparison with other published error rates (taken from Dissen et al. 2019)

| RMSE, all segments | | | | |
|---|---|---|---|---|
| Model | Mean | F1 | F2 | F3 |
| **FormantNet** | 173 | 143 | 177 | **195** |
| *Dissen et al. (2019):* | | | | |
| DF-RNN | **163** | 118 | **169** | 204 |
| DF-RCNN | 173 | 127 | 180 | 213 |
| *Others:* | | | | |
| KARMA | 220 | **114** | 226 | 320 |

| MAE, vowels | | | | |
|---|---|---|---|---|
| Model | Mean | F1 | F2 | F3 |
| **FormantNet** | **76** | 64 | 75 | **90** |
| *Dissen et al. (2019):* | | | | |
| DF-RNN | 82 | 54 | 81 | 112 |
| DF-RCNN | 78 | **53** | **72** | 108 |
| *Others:* | | | | |
| MSR 2006 | 98 | 64 | 105 | 125 |
| WaveSurfer | 106 | 70 | 94 | 154 |
| Praat | 209 | 130 | 230 | 267 |

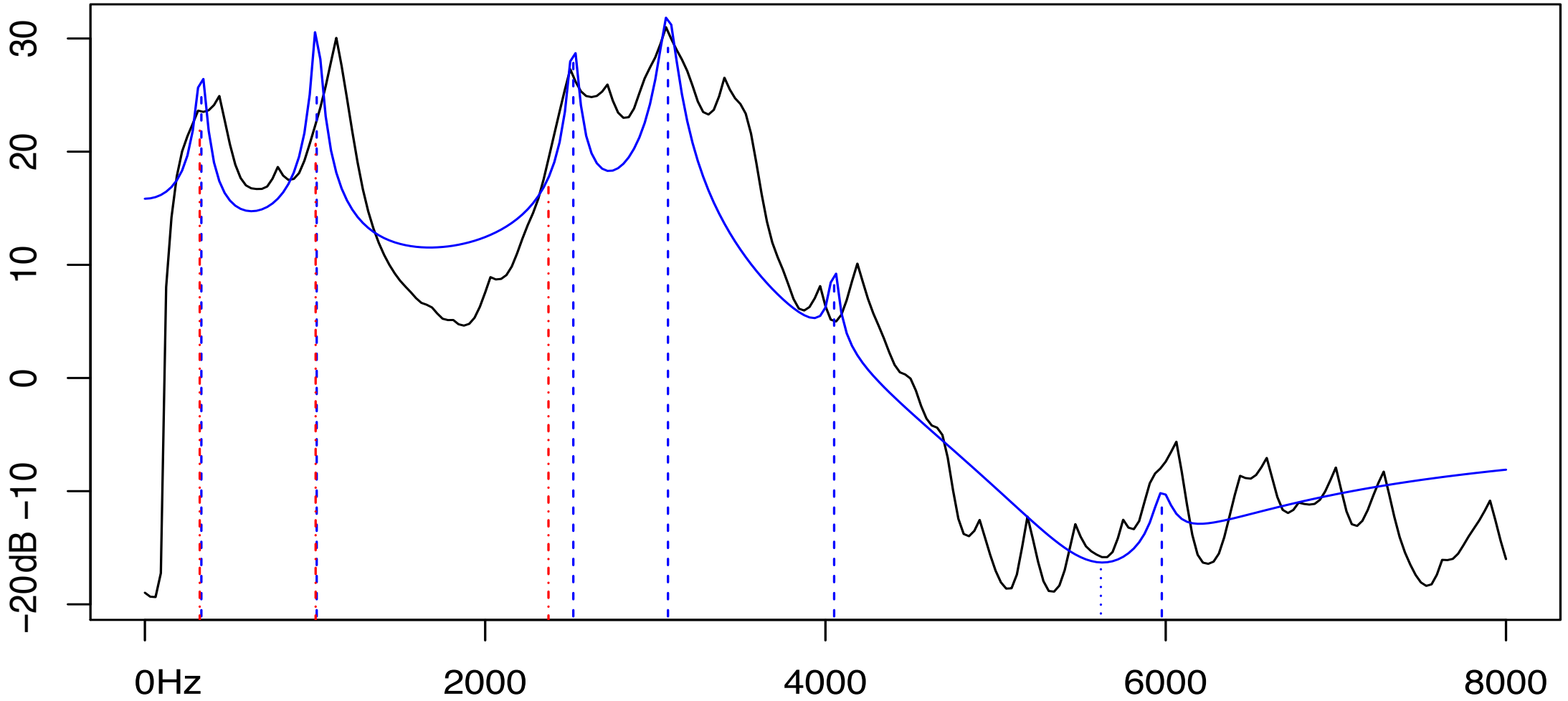
# Example spectrogram




Red: Hand-measured frequencies (F1-F3)

Blue: Model frequencies (F1-F6, A1)

# Example input vs. model spectra: /u/



# Improvements since IS submission (presented by Lilley & Bunnell at PaPE 2021, June 23)

1) Delta-Frequency ($\Delta f$) Loss
2) Changes in input source function modeling

# Delta-Frequency ($\Delta f$) Loss

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

$$\Delta f Loss = \frac{1}{T \times N} \sum_{t=1}^{T} \sum_{n=1}^{N} f_{n,t} - f_{n,t-1} \quad (5)$$

- Total loss is regular loss plus **weighted** sum of $\Delta f$ loss:

$$Loss = \frac{1}{T \times F} \sum_{t=0}^{T} \sum_{f=0}^{F} \left( h_{Dec}(f) - h_{Inp}(f) \right)^2 + \boldsymbol{W} \times \Delta f Loss \quad (6)$$

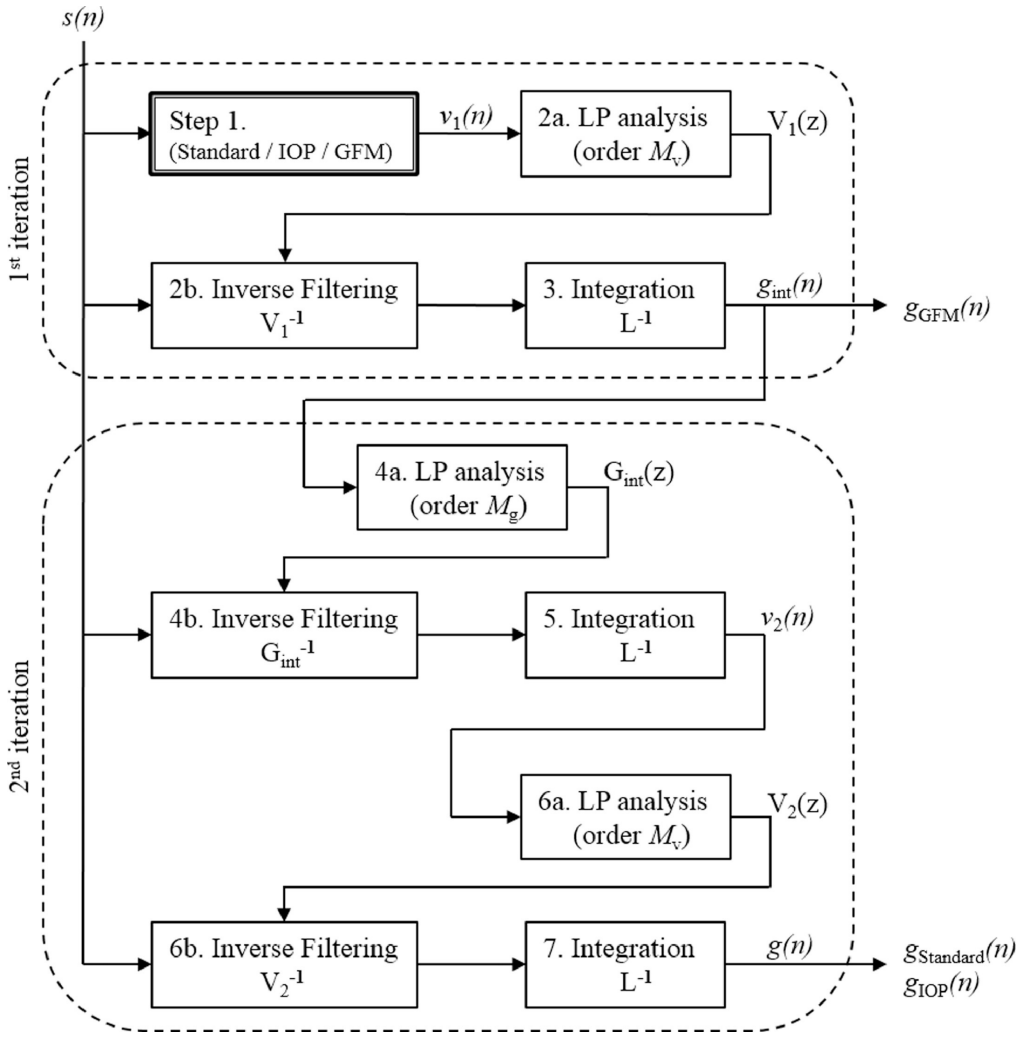
# Source function modeling

**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: $\Delta f$ 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 |
| NS | 0 | 115 | 94 | 121 | 130 |
| IAIF | .05 | 105 | 97 | 99 | 118 |
| PE | .15 | **98** | **88** | **98** | **109** |
| NS | .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 |
| NS | 0 | 76 | 61 | 80 | 92 |
| IAIF | .05 | 75 | 62 | 72 | 92 |
| PE | .15 | 71 | 58 | **70** | **86** |
| NS | .10 | **70** | **51** | 72 | 87 |

PE = Pre-Emphasis model
NS = No Source model

# Results: MAE per phonetic class
## (Adapted from Dissen et al. 2019, Table VI)

Table 2: *Mean absolute error over all speech in test set, divided by phonetic class.*

| | Praat | | | WaveSurfer | | | MSR | | | DF-RNN | | | DF-RCNN | | | FormantNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** |
| vowel | 130 | 230 | 267 | 70 | 94 | 154 | 64 | 105 | 125 | 54 | 81 | 112 | **53** | **72** | 108 | 64 | 75 | **90** |
| semivowel | 136 | 295 | 334 | 89 | 126 | 222 | 83 | 122 | 154 | **67** | 114 | 168 | 68 | 111 | 160 | 79 | **93** | **124** |
| nasal | 219 | 409 | 381 | 96 | 229 | 239 | 67 | **120** | **112** | **66** | 175 | 151 | 69 | 191 | 158 | 98 | 213 | 143 |
| fricative | 564 | 593 | 700 | 209 | 263 | 439 | **129** | **108** | **131** | 131 | 135 | 159 | 139 | 142 | 167 | 160 | 135 | 161 |
| affricate | 730 | 515 | 583 | 292 | 407 | 390 | **141** | **129** | **149** | 164 | 162 | 189 | 174 | 173 | 195 | 186 | 186 | 186 |
| stop | 258 | 270 | 351 | 168 | 210 | 286 | 130 | **113** | **119** | 131 | 135 | 168 | **123** | 135 | 170 | 135 | 158 | 166 |

"MSR": Microsoft Research (Deng et al. 2004) --
Used as starting point for hand labels

# Results: MAE per phonetic class (cont.)
## (Adapted from Dissen et al. 2019, Table VI)

Table 2b: *Mean absolute error over all speech in test set, divided by phonetic class.*

| | MSR | | | DF-RNN | | | DF-RCNN | | | FNet IAIF0 | | | FNet PE15 | | | FNet NS10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** |
| vowel | 64 | 105 | 125 | 54 | 81 | 112 | 53 | 72 | 108 | 64 | 75 | 90 | 58 | **70** | **86** | **51** | 72 | 87 |
| semivowel | 83 | 122 | 154 | **67** | 114 | 168 | 68 | 111 | 160 | 79 | 93 | **124** | 73 | **90** | 128 | **67** | **90** | 129 |
| nasal | 67 | **120** | 112 | 66 | 175 | 151 | 69 | 191 | 158 | 98 | 213 | 143 | 96 | 147 | 111 | **65** | 142 | **110** |
| fricative | **129** | **108** | 131 | 131 | 135 | 159 | 139 | 142 | 167 | 160 | 135 | 161 | 138 | 117 | **123** | 137 | 117 | 126 |
| affricate | **141** | **129** | **149** | 164 | 162 | 189 | 174 | 173 | 195 | 186 | 186 | 186 | 172 | 169 | 177 | 177 | 160 | 178 |
| stop | 130 | **113** | **119** | 131 | 135 | 168 | 123 | 135 | 170 | 135 | 158 | 166 | **113** | 123 | 135 | 146 | 124 | 137 |

"MSR": Microsoft Research (Deng et al. 2004) --
Used as starting point for hand labels

# Results: RMSE error over all segments
## (Adapted from Dissen et al. 2019, Table VIII)

Table 3b: *Root mean square error over all segments, test set.*

| Method | All | F1 | F2 | F3 |
|---|---|---|---|---|
| KARMA | 220 | **114** | 226 | 320 |
| DF-RNN | 163 | 118 | 169 | 204 |
| DF-RCNN | 173 | 127 | 180 | 213 |
| FNet IAIF0 | 173 | 143 | 177 | 195 |
| FNet PE15 | **150** | 125 | 146 | **174** |
| FNet NS10 | 152 | 133 | **144** | 176 |

KARMA: Mehta et al. (2012)

# Results: RMSE per gender, voiced frames

- Adapted from Schiel & Zitzelberger (2018)
- They compared 3 LPC-based trackers as well as DF-RNN
- They only tested DF-RNN on the test set (since it was trained on the training set)

Table 4b: *Root mean square error over all segments, split by gender, voiced frames only.*

| | F1 | | F2 | | F3 | |
|---|---|---|---|---|---|---|
| **Tracker** | **f** | **m** | **f** | **m** | **f** | **m** |
| *All utterances:* | | | | | | |
| SNACK | 126 | 100 | 291 | 227 | 313 | 375 |
| ASSP | 113 | 96 | 479 | 211 | 512 | 225 |
| PRAAT | 116 | 234 | 217 | 338 | 249 | 404 |
| FNet IAIF0 | 108 | 99 | 180 | 129 | 178 | **149** |
| FNet PE15 | 96 | 93 | 140 | **108** | 162 | 159 |
| FNet NS10 | **80** | **75** | **136** | 110 | **154** | 166 |
| *Test utterances:* | | | | | | |
| DF-RNN | 120 | 97 | 195 | 167 | 252 | 169 |
| FNet IAIF0 | 104 | 93 | 182 | 132 | 192 | **157** |
| FNet PE15 | 94 | 86 | 153 | **116** | 179 | 166 |
| FNet NS10 | **77** | **76** | **138** | 117 | **161** | 174 |

# Results: FDR of FormantNet vs. QCP-FBCOV

- Adapted from Gowda, Airaksinen, & Alku (JASA 2017)
- "Quasi-Closed Phase Forward-Backward Covariance-based" linear prediction analysis
- Formant Detection Rate (FDR): % of formant measurements within both an absolute and relative threshold of ground truth

Table 5b: *Formant detection rates (FDR) at different thresholds, for vowels and semivowels in the test set.*

| Ratio | F1 | F2 | F3 |
|---|---|---|---|
| *FDR within 20% and 200 Hz dev* | | | |
| QCP-FBCOV | 84.9 | 85.0 | 83.9 |
| FNet IAIF0 | 79.6 | 93.7 | 90.4 |
| FNet PE15 | 81.7 | 94.5 | **91.4** |
| FNet NS10 | **87.7** | **94.5** | 91.3 |
| *FDR within 25% and 250 Hz dev* | | | |
| QCP-FBCOV | 90.4 | 90.5 | 89.1 |
| FNet IAIF0 | 85.9 | 96.5 | 93.8 |
| FNet PE15 | 87.5 | 96.8 | **94.1** |
| FNet NS10 | **92.4** | **97.1** | **94.1** |
| *FDR within 30% and 300 Hz dev* | | | |
| QCP-FBCOV | 93.4 | 93.9 | 92.1 |
| FNet IAIF0 | 90.1 | 97.9 | 95.5 |
| FNet PE15 | 91.4 | 98.0 | **95.6** |
| FNet NS10 | **95.0** | **98.3** | **95.6** |

# Conclusions

- “FormantNet” method can be used to train a tracker on a corpus without prior hand-obtained formant measures for training
- Lower error than popular LPC-based trackers and supervised DNNs
- Produces a model of the **entire** vocal-tract signal: Frequencies, bandwidths, and amplitudes of **all** formants
  - Useful for e.g. formant speech synthesis
- Available now: github.com/NemoursResearch/FormantNet

# 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.

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.

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

J. Lilley and H. T. Bunnell, “A Neural-Network-Based Formant Tracker Trained on Unlabeled Data,” in *PaPE 2021 (4th Phonetics and Phonology in Europe): Book of Abstracts*, Brno, Czechia [virtual meeting], June 23, 2021, pp. 337-8. pape2021.upf.edu/program-4/

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.

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.

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 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.

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. 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.

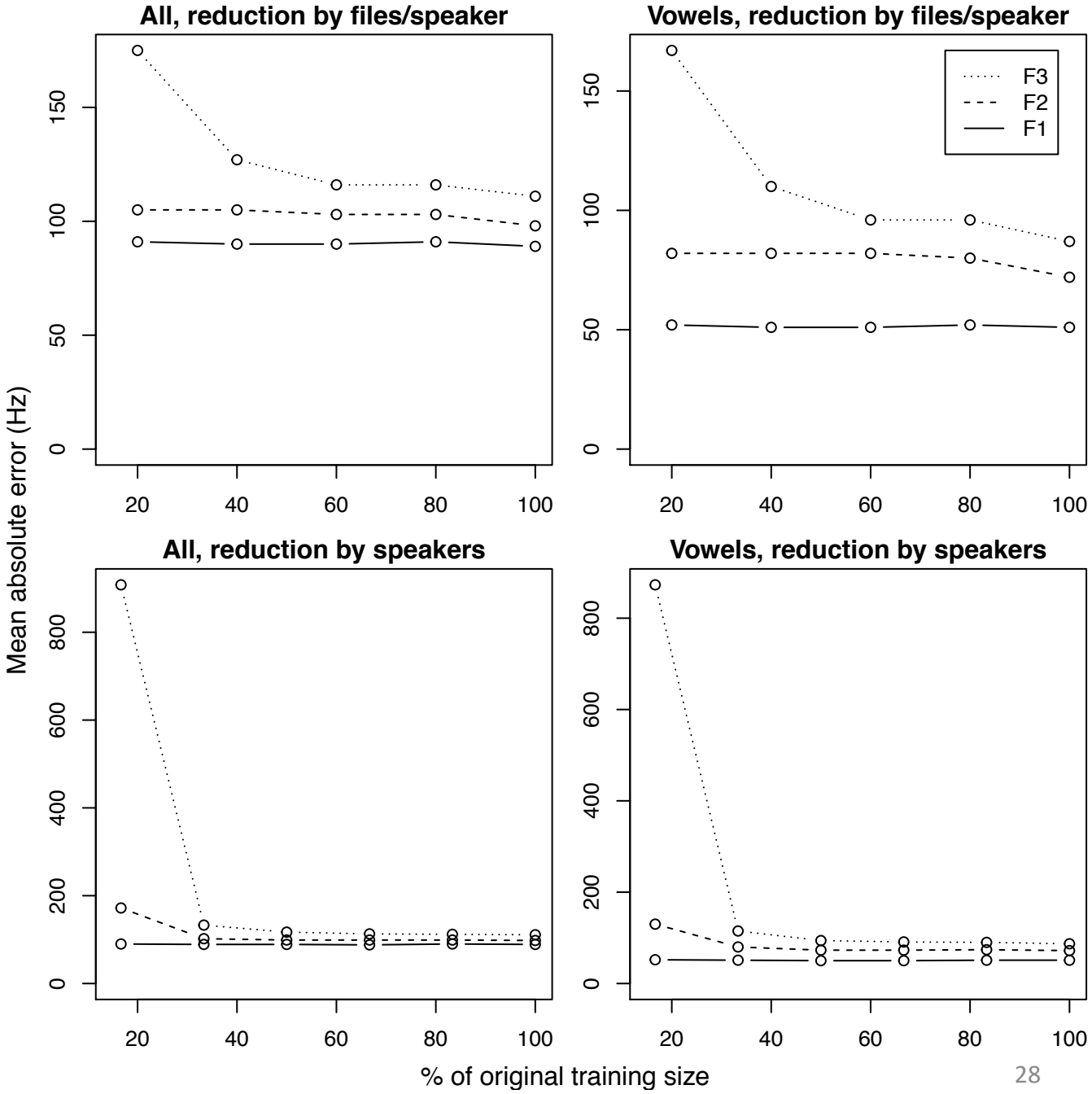
This is the end of the video slides. More results are reproduced below.

# 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

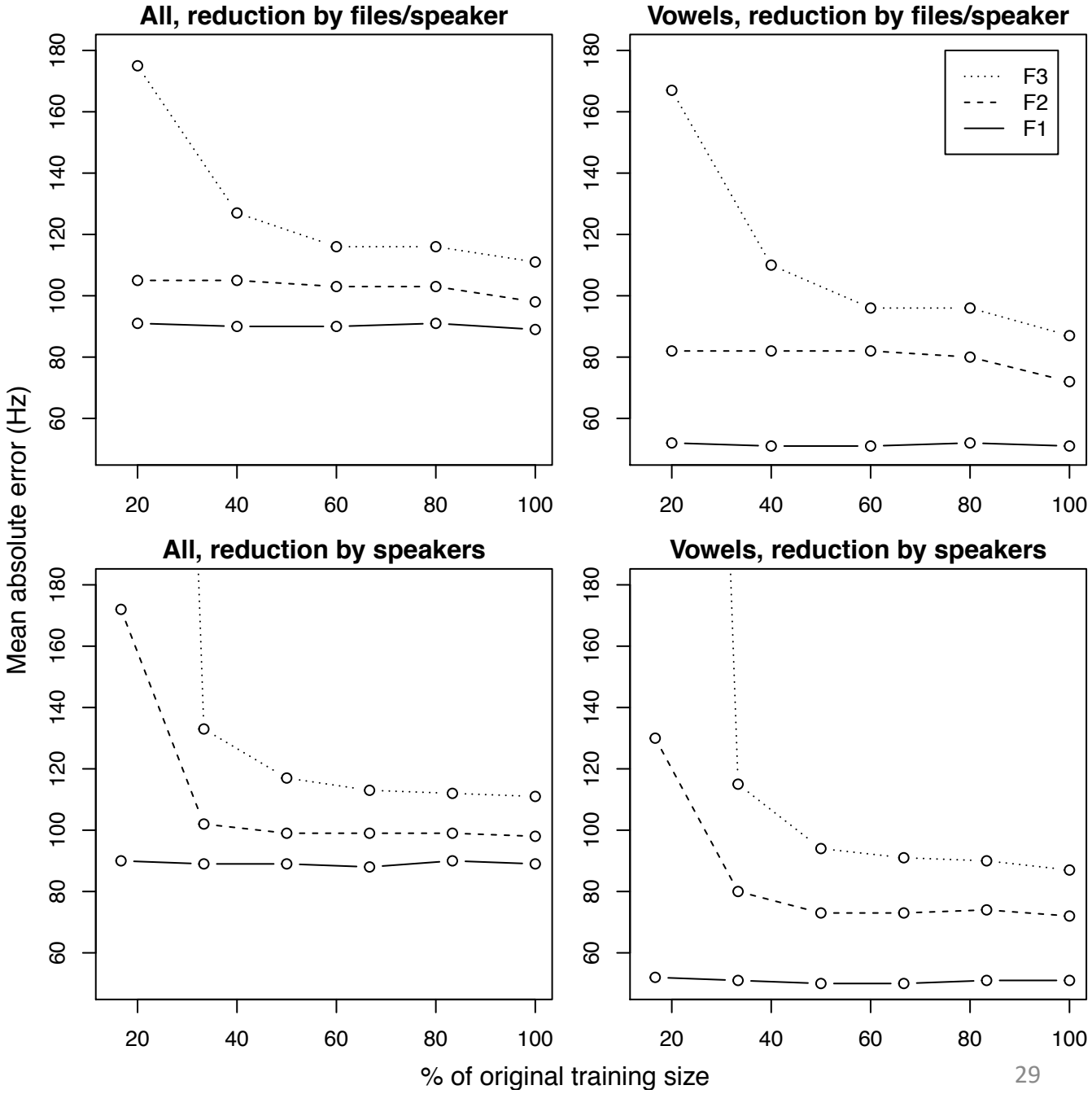


# Effect of Training Size (figure zoom-in)

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



# 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 |

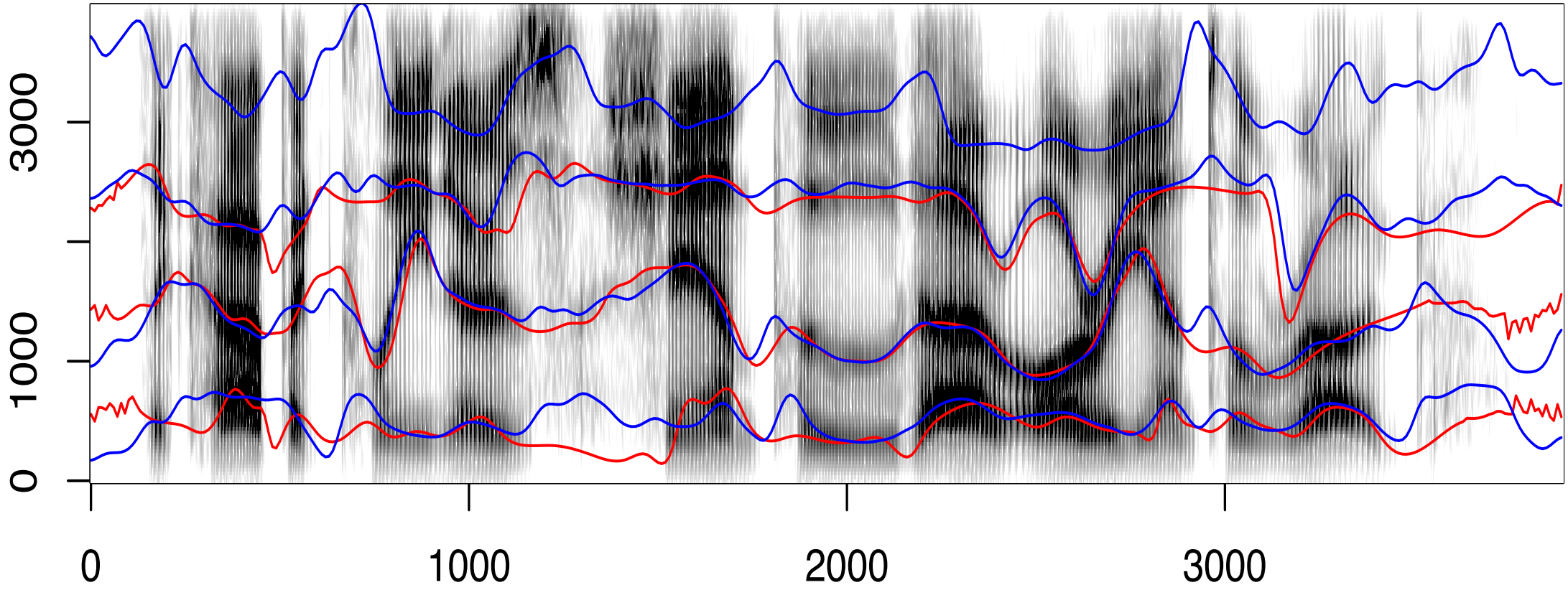
| MAE, vowels | | | | | |
|---|---|---|---|---|---|
| Training set | n | Mean | F1 | F2 | F3 |
| All train | 4140 | 70 | 51 | 72 | 87 |
| Partial train + test | 4140 | 71 | 51 | 73 | 88 |
| All train + test | 5820 | 72 | 51 | 75 | 90 |
| Train + test + val | 6300 | 71 | 51 | 74 | 87 |

# 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
- 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 |
| 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 |

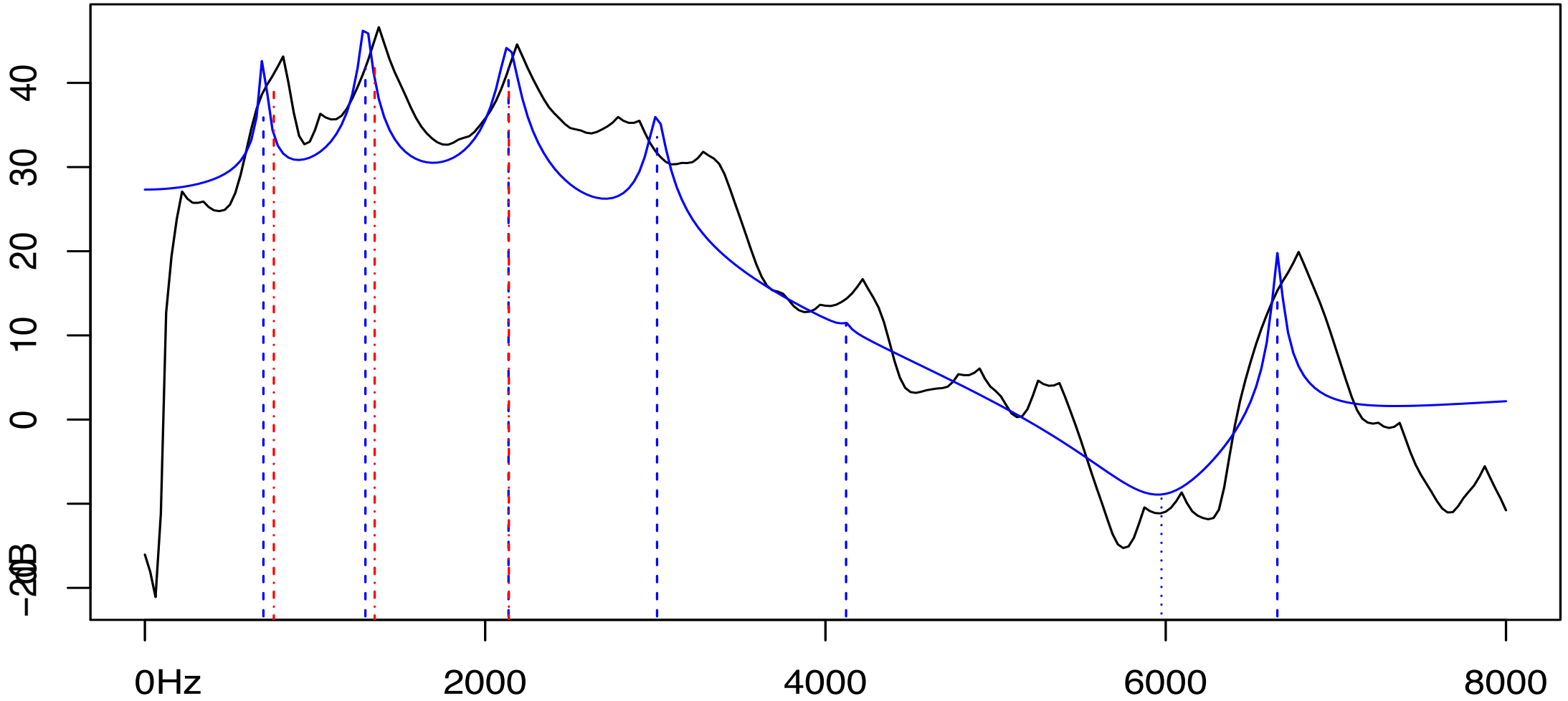
# Example spectrogram



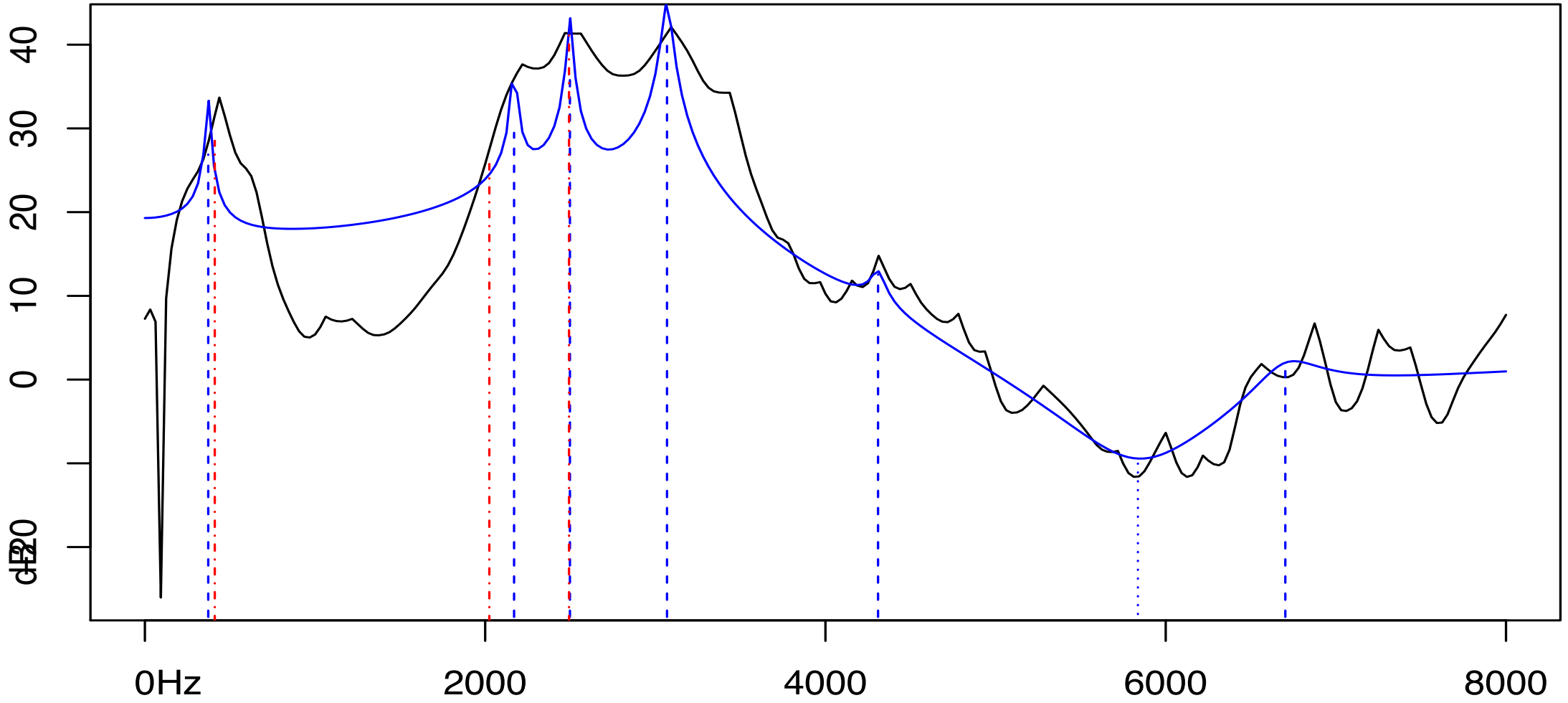

Red: Hand-measured frequencies (F1-F3)

Blue: Model frequencies (F1-F4)

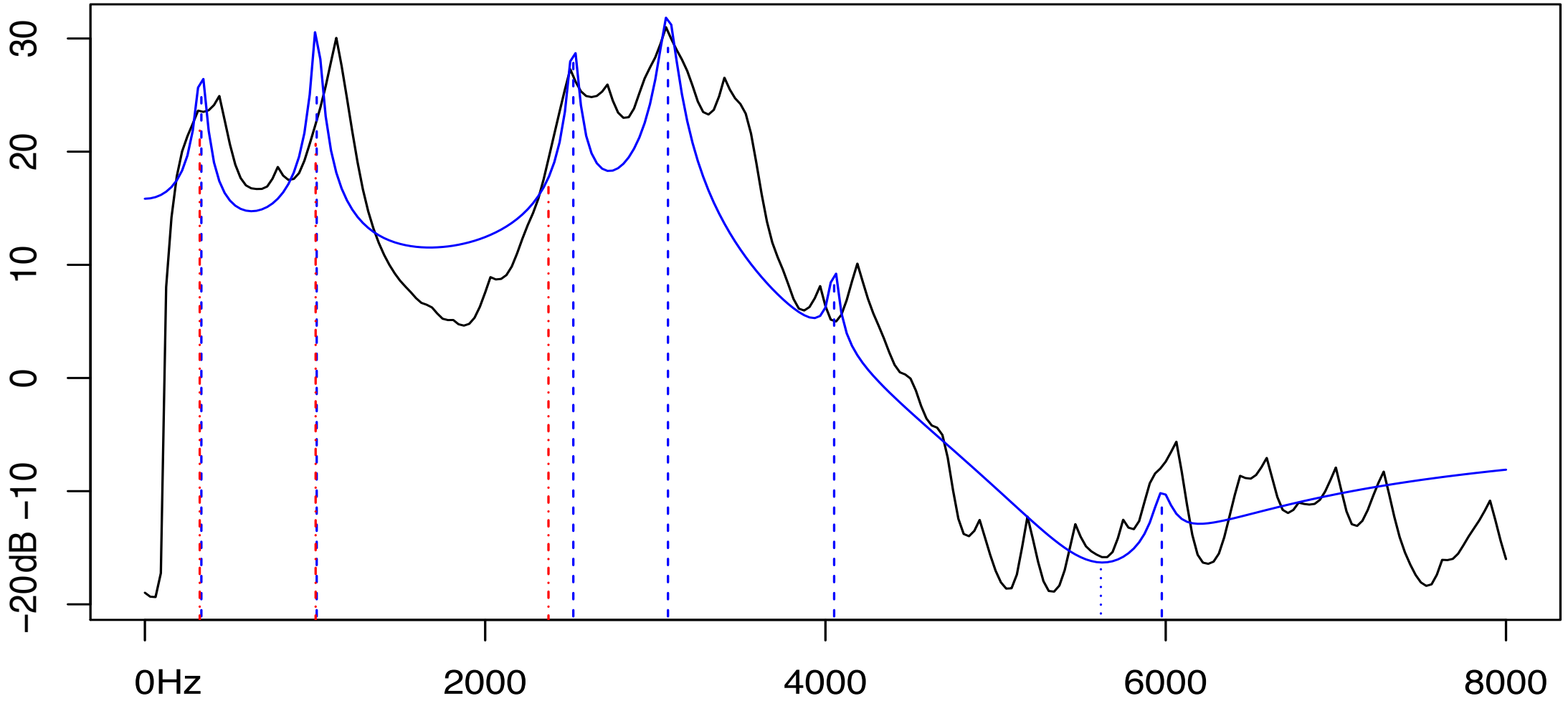
# Example spectra: /a/



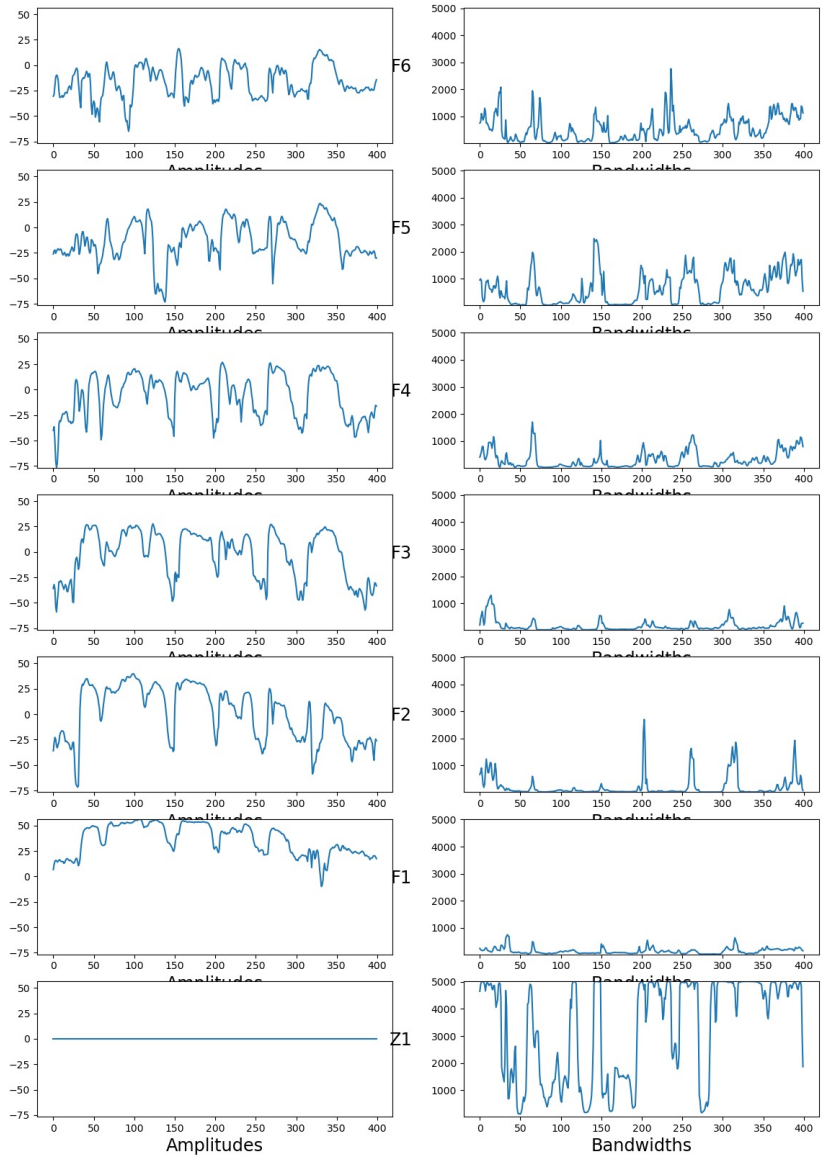
# Example spectra: /i/



# Example spectra: /u/



# Amplitudes and Bandwidths



# $\Delta f$ 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 (w=0)

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

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

# Source Modeling (w=.05)

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

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

# Source Modeling (w=.10)

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

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

# Source Modeling (w=.15)

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

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

# Source Modeling w/ $\Delta f$ Loss ("best" weights)

| MAE, all segments | | | | | |
|---|---|---|---|---|---|
| Source | *W* | Mean | F1 | F2 | F3 |
| IAIF | .05 | 105 | 97 | 99 | 118 |
| Pre-emphasis | .15 | **98** | **88** | **98** | **109** |
| None | .10 | 99 | 89 | 98 | 111 |

| MAE, vowels | | | | | |
|---|---|---|---|---|---|
| Source | *W* | Mean | F1 | F2 | F3 |
| IAIF | .05 | 75 | 62 | 72 | 92 |
| Pre-emphasis | .15 | 71 | 58 | **70** | **86** |
| None | .10 | **70** | **51** | 72 | 87 |

# Improved results: Comparison with other trackers

| RMSE, all segments | | | | |
|---|---|---|---|---|
| Model | Mean | F1 | F2 | F3 |
| FNet PE15 | **150** | 125 | 146 | **174** |
| FNet NS10 | 152 | 133 | **144** | 176 |
| FNet LSTM1 | 173 | 143 | 177 | 195 |
| DF-RNN | 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 |
| FNet LSTM1 | 76 | 64 | 75 | 90 |
| DF-RNN | 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 |