MERIT FAIR

BIEN 2009

Robust Speech Recognition

Articulatory Information to Account for Coarticulation

Rob Bailey

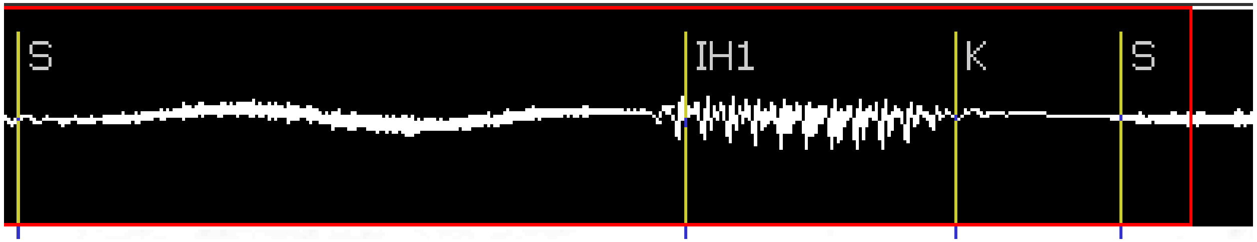
Kossivi Wody Edji

Vikramjit Mitra

Dr. Carol Espy-Wilson

Motivation

- Current Automatic Speech Recognition (ASR) systems are phone-based and assume phones to be distinctive regions.

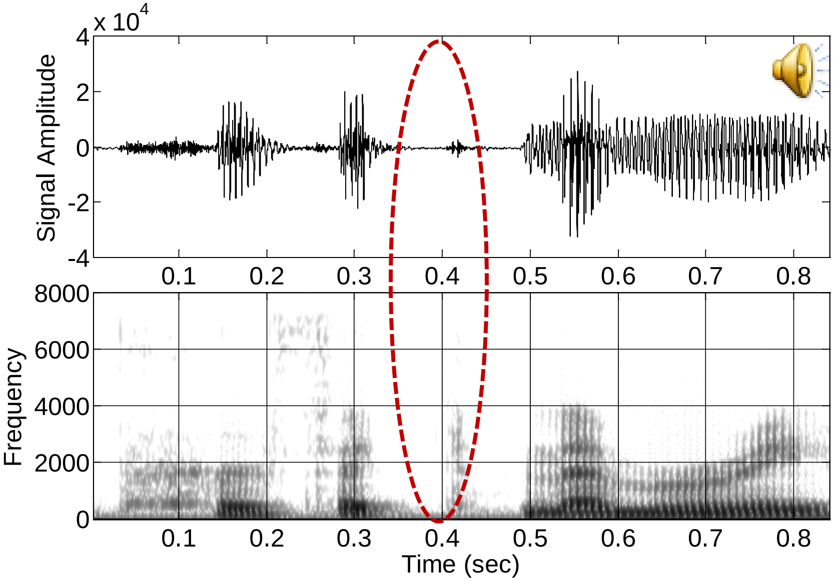


- Current state-of the art ASR systems need to impose limitations (e.g., clearly-articulated speech or limited vocabulary) in the recognition task in order to handle speech variability such as coarticulation

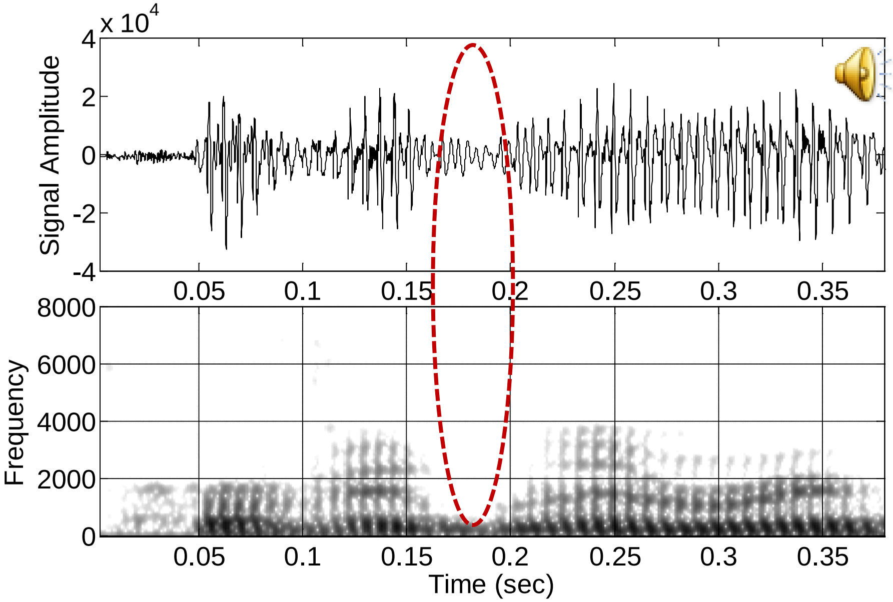
MERIT FAIR
BIEN 2009

Coarticulation

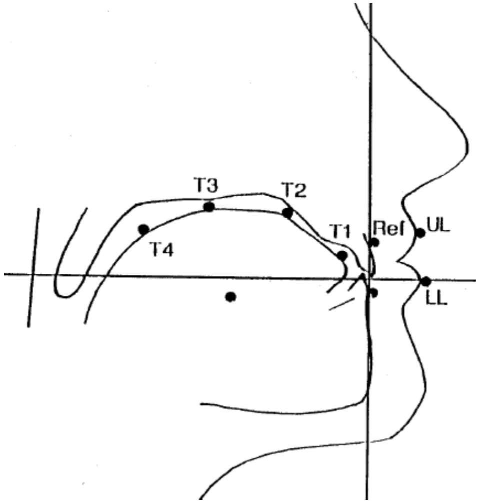
“perfect-memory” clearly articulated



“perfect-memory” quickly articulated



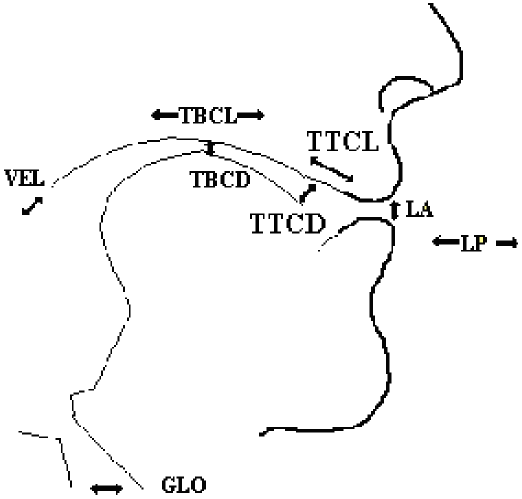
Our Approach



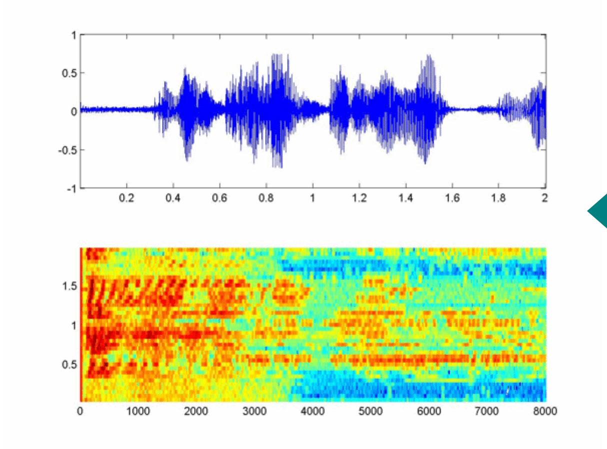
- We are using articulatory information instead of phones to account for coarticulation

- Previous studies have used articulatory information in the form of the Cartesian coordinates of the pellet locations.
  - Pellet data are often inconsistent and introduce more non-uniqueness.

- In our study, we are using tract variables instead of pellet information.
  - The tract variables are relative measures and reduce the non-uniqueness.



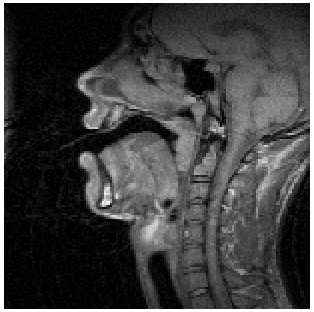
How to obtain articulatory information?



Speech signal

Speech production

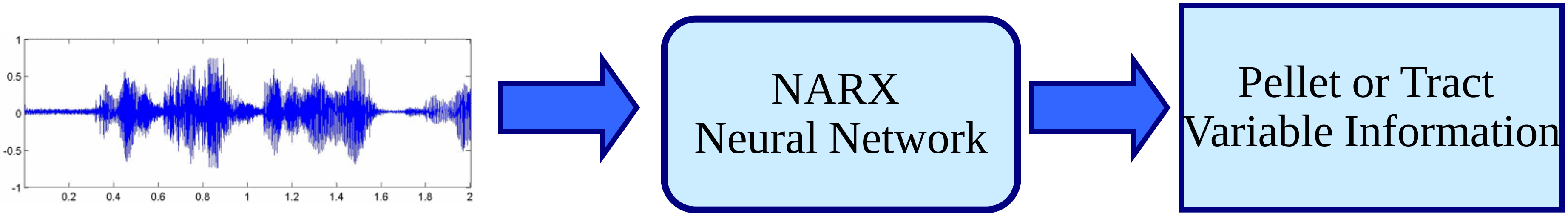
inversion



Articulatory motion

Task 1: Speech Inversion

- Objective:
  - – Train Neural Networks to estimate tract variables and pellet trajectories given a speech signal
    - NARX = Nonlinear Autoregressive Networks with Exogenous Inputs



NARX
Neural Network

Pellet or Tract
Variable Information

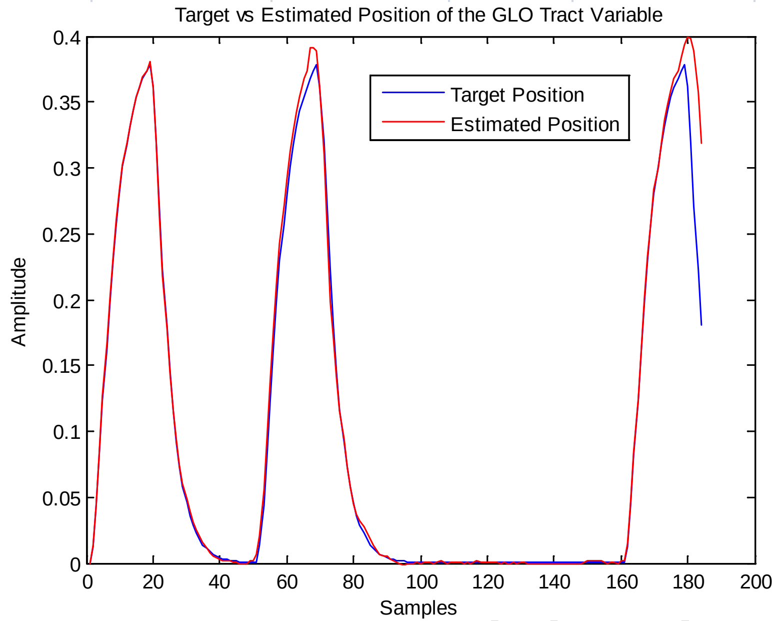
- Procedure:
  - – Implement the process of optimization through five trials of training for neural networks to achieve the most accurate network

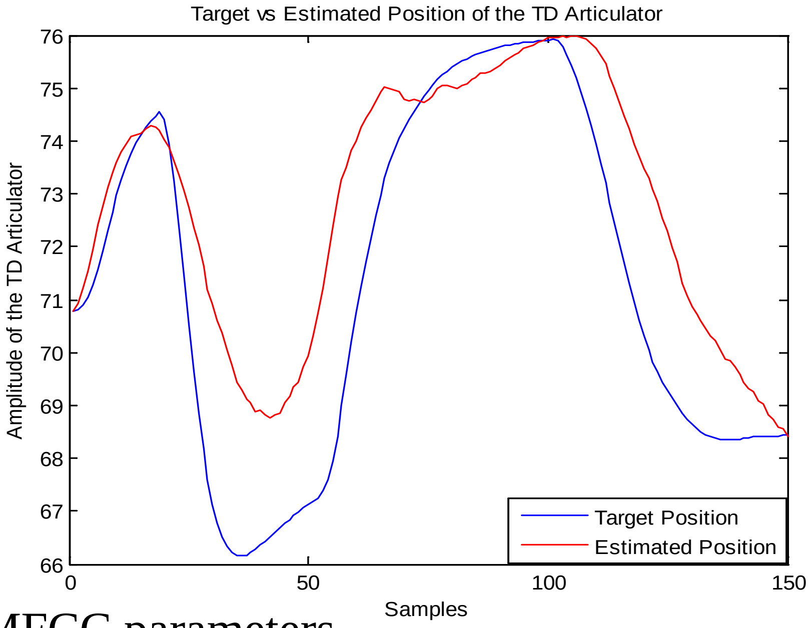
MERIT FAIR
BIEN 2009

Neural Network Training Results

Mel Frequency Cepstral Coefficients (MFCC)			
Tract Variables	Correlation	Pellets	Correlation
GLO	0.98	LL	0.64
VEL	0.90	UL	0.41
LA	0.85	JAW	0.85
LP	0.52	TD	0.93
TTCD	0.93	TF	0.89
TTCL	0.93	TR	0.93
TBCD	0.91	TT	0.84
TBCL	0.91		
Avg	0.87	Avg	0.78

Acoustic Parameters (AP)			
Tract Variables	Correlation	Pellets	Correlation
GLO	0.99	LL	0.60
VEL	0.73	UL	0.63
LA	0.76	JAW	0.83
LP	0.69	TD	0.88
TTCD	0.90	TF	0.82
TTCL	0.86	TR	0.88
TBCD	0.83	TT	0.75
TBCL	0.88		
Avg	0.83	Avg	0.77


Target vs Estimated Position of the GLO Tract Variable


Target vs Estimated Position of the TD Articulator

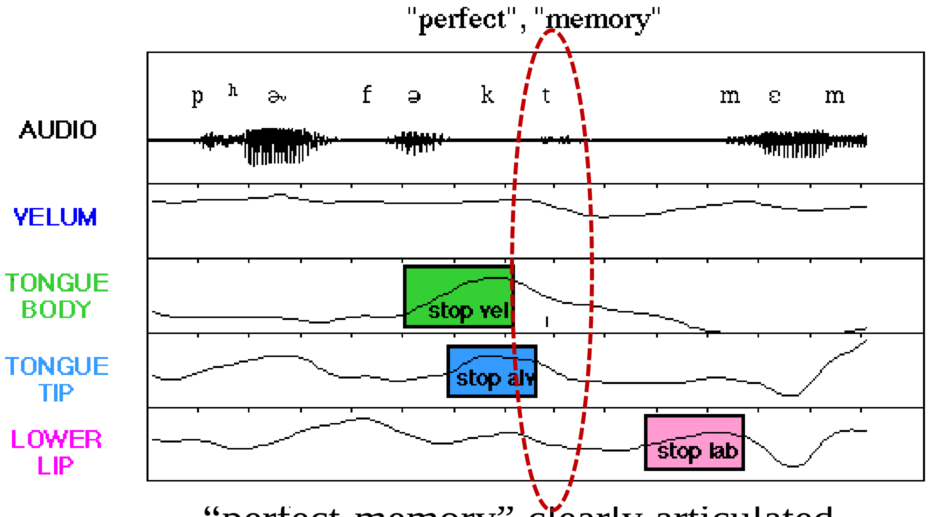
Graph obtained using MFCC parameters

MERIT FAIR
BIEN 2009

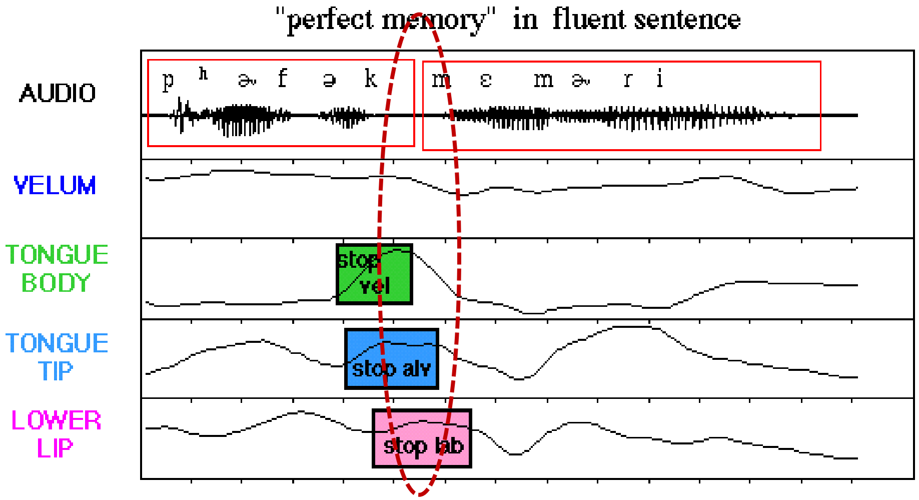
Task 2: Model Gestures for XRMB

- Gestures are constriction actions along the vocal tract and they are defined by dynamic parameters

- How will gestures account for coarticulation?



“perfect-memory” clearly articulated

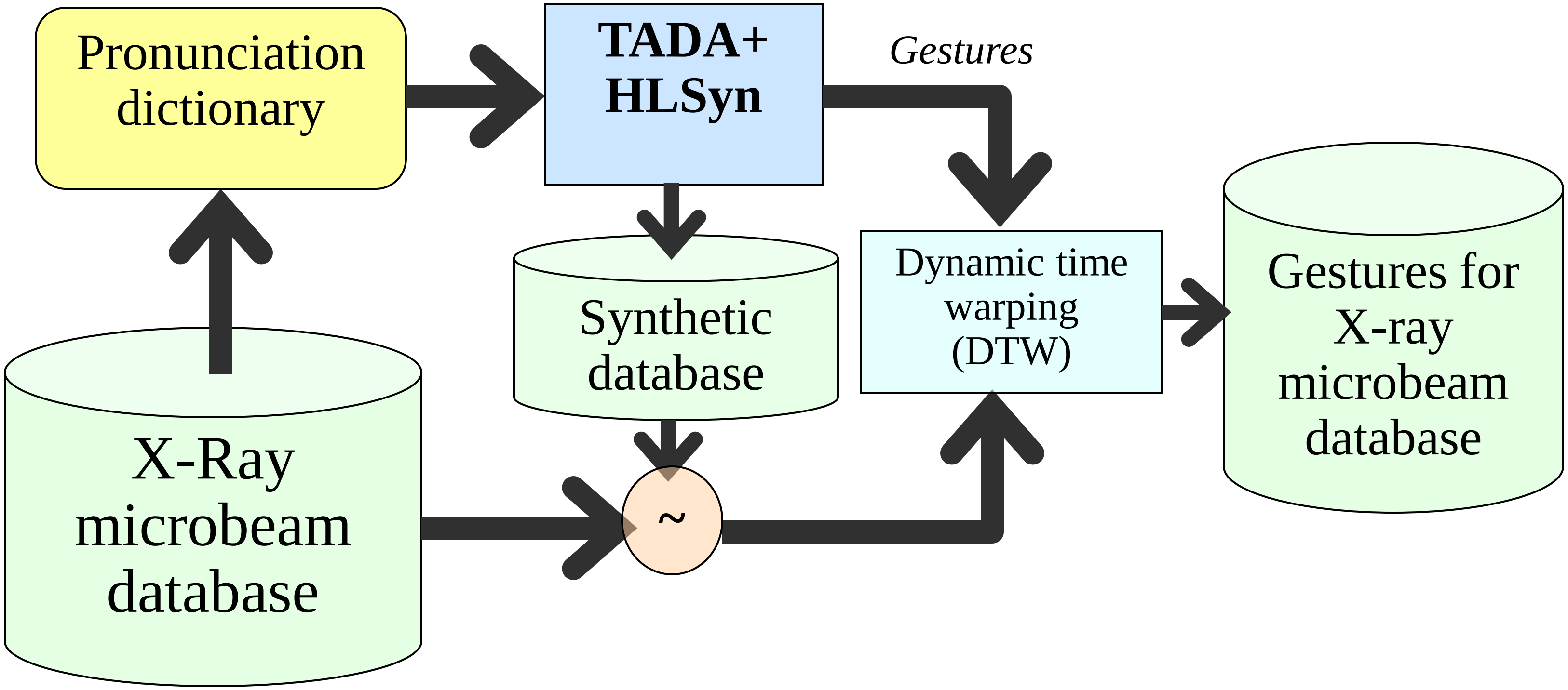


“perfect-memory” quickly articulated

MERIT FAIR
BIEN 2009

Task 2: Model Gestures for XRMB

- Procedure:

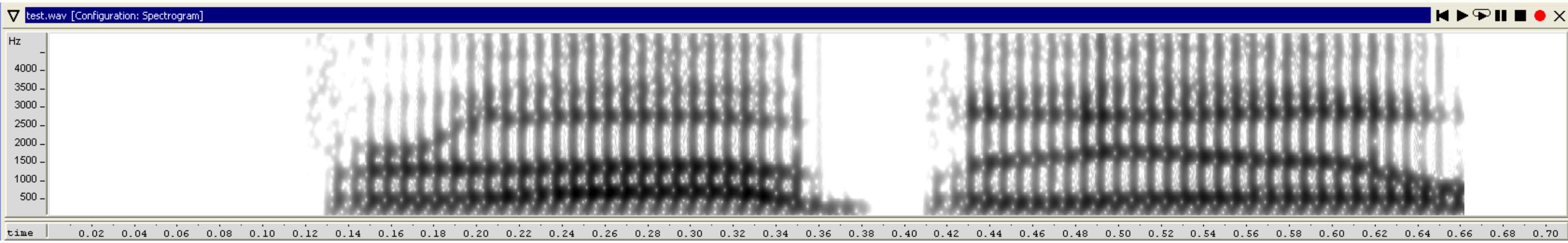


MERIT FAIR

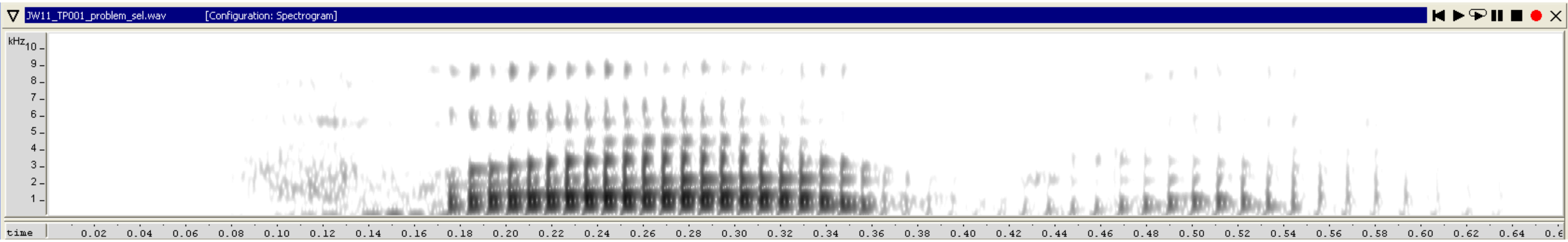
BIEN 2009

Dynamic Time Warping Results

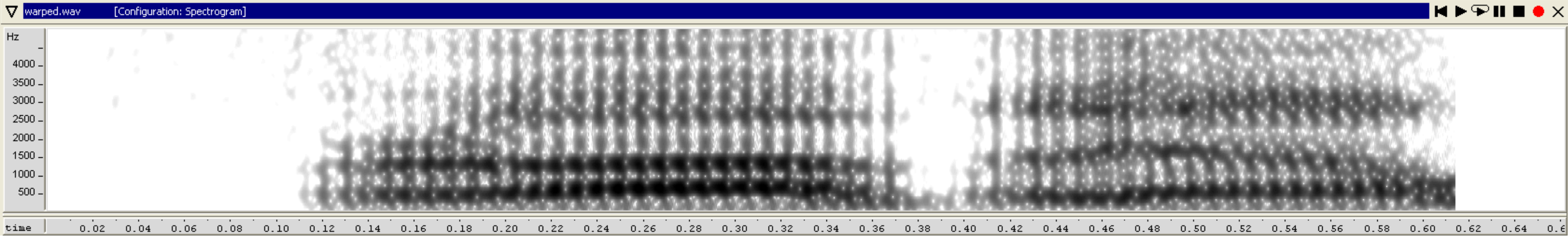
synthetic



natural



warped



Conclusions and Future Work

- Neural networks estimated the tract variables more accurately than the pellets

- We were able to warp the synthetic speech signal to the natural speech signal

- We have obtained the gestures for the natural speech from the gestures of the warped synthetic speech

- Our research is a preliminary step in designing an ASR systems which uses gestures obtained from tract variables to account for coarticulation

Acknowledgments

- National Science Foundation CISE award #0755224
- Dr. Carol Espy-Wilson
- Vikramjit Mitra
- MERIT BIEN Faculty

References

- H. Demuth, M. Beale and M. Hagan “Neural Network Toolbox 6 User’s Guide” The MathWorks, Natick, MA, 2008.
- J. Frankel and S. King, "ASR - Articulatory Speech Recognition", In proceedings of Eurospeech, pp. 599-602, Aalborg, Denmark, September 2001.
- J. Frankel and S. King, "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition", in Proceedings of Eurospeech, Interspeech-2005, pp.3045-3048, Lisbon, Portugal, 2005.
- H. Nam, L. Goldstein, E. Saltzman and D. Byrd, “Tada: An enhanced, portable task dynamics model in matlab”, Journal of the Acoustical Society of America, Vol. 115, no. 5, 2, pp. 2430, 2004.
- J. Westbury, “X-ray microbeam speech production database user’s handbook”, University of Wisconsin, 1994.