Robust Speech Recognition

Rob Bailey, Kossivi Wody Edji, Dr. Carol Espy-Wilson, Vikramjit Mitra

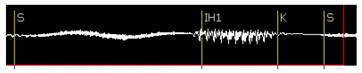
Objectives

- Compare how accurately machine learning techniques can estimate from acoustic data:
  - vocal tract constrictions and their locations (tract variables)
  - the location of various articulators (pellets)
- Create gestural specifications for a spontaneous speech database

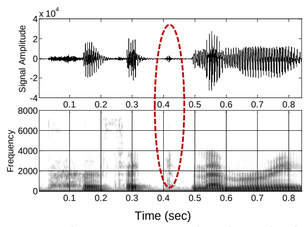
Motivation

Current ASR systems:

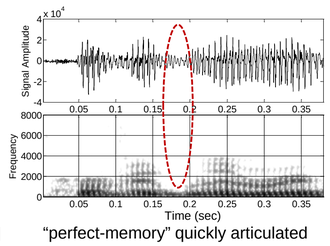
- Based on phones and assume phones to be distinctive regions



- Need to impose limitations in the recognition task in order to handle coarticulation



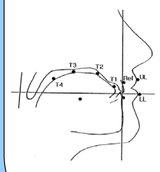
"perfect-memory" clearly articulated



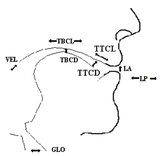
"perfect-memory" quickly articulated

Our Approach

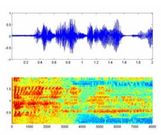
- Use articulatory information to better model coarticulation:
  - 7 pellets or 8 tract variables
  - Speech Inversion Technique



Pellets



Tract Variables



Speech signal

Speech production

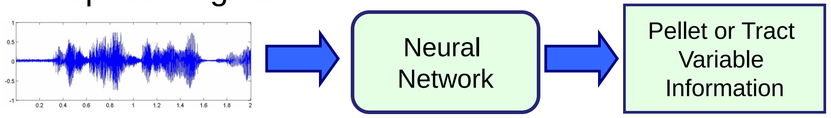
Speech inversion

Articulatory motion

Methodology

•Task 1: Speech Inversion

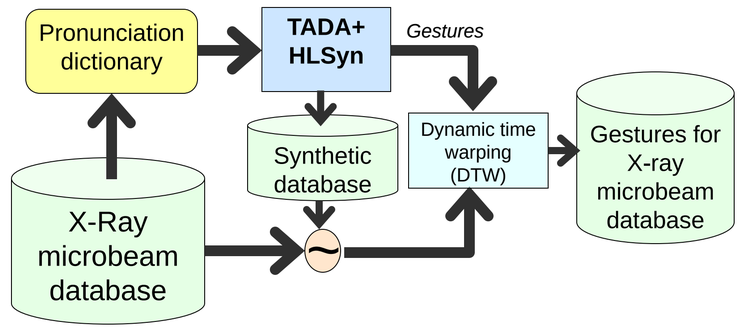
•Train artificial neural networks (ANNs) to estimate tract variables and pellet trajectories given a speech signal



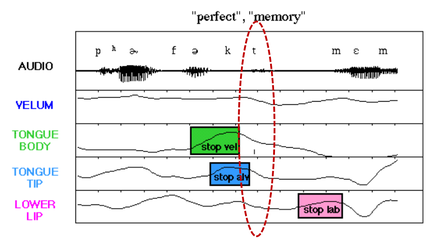
•Task 2: Gestural Modeling

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

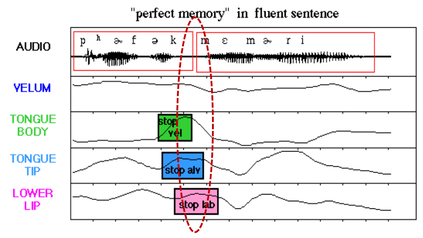
•Procedure



• Gestures are invariant which can account for coarticulation



"perfect-memory" clearly articulated



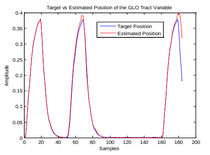
"perfect-memory" quickly articulated

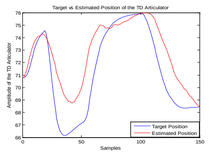
Results

•Neural Network Training Results

- Using Mel-Frequency Cepstral Coefficients (MFCCs)

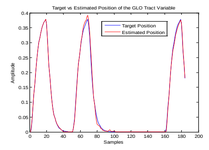
MFCC			
Tract Variables	Correlation	Pellets	Correlation
GLO	98.48	LL	63.70
VEL	89.55	UL	40.69
LA	84.74	JAW	84.75
LP	51.76	TD	93.42
TTCD	92.99	TF	88.92
TTCL	93.22	TR	93.40
TBCD	91.20	TT	84.26
TBCL	90.52		
AVG	86.56		78.45

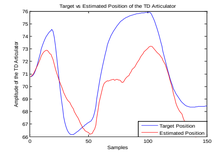




- Using Acoustic Parameters (APs)

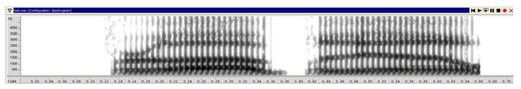
AP			
Tract Variables	Correlation	Pellets	Correlation
GLO	99.34	LL	59.56
VEL	73.04	UL	62.66
LA	76.04	JAW	82.76
LP	68.65	TD	87.54
TTCD	89.92	TF	82.10
TTCL	86.40	TR	88.30
TBCD	83.23	TT	75.40
TBCL	87.98		
AVG	83.07		76.90



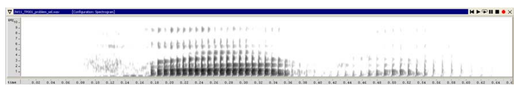


•Dynamic Time Warping Results

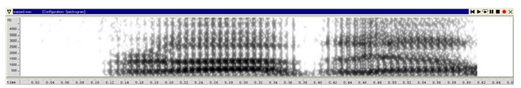
1.



2.



3.



1. Synthetic 2. Natural 3. Warped

Conclusions

- Estimated the tract variables more accurately than the pellets using neural networks
- Warped the synthetic speech signal to the natural speech signal
- Obtained the gestures for the natural speech from the warped synthetic speech

Acknowledgments

•This work was supported by NSF CISE award #0755224