# Multi-faceted Research in Speech Communication


## Prof Espy-Wilson

**Students: V Mitra, X Zhou, S Vishnubhotla, V Mahadevan, D Garcia-Romero**

**Collaborators: Haskins Labs, UIUC, UCLA, USC, BU, UC**


ISR @25

The Institute for Systems Research

A. JAMES CLARK SCHOOL OF ENGINEERING

## Articulatory information for Robust Automatic Speech Recognition

➢Current automatic speech recognition (ASR) systems assume speech is a string of non-overlapping phone units, an assumption that limits the ability of the acoustic models to properly account for variability such as coarticulation.
➢Articulatory information helps to model coarticulation and we have also shown that it also improves the robustness of ASR system

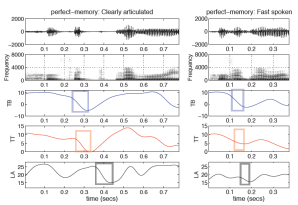
### Articulatory Information

Gestural Scores: Define constriction actions that specify the initiation and termination of a target driven articulatory constriction within the vocal tract.
TVs: describe geometrically the shape of the vocal tract in terms of constriction degree and location.

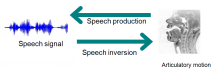
| Constriction organs | Vocal tract variables (TVs) | Articulators |
|---|---|---|
| Lip | Lip Aperture (LA) | Upper lip, lower lip, jaw |
| | Lip Protrusion (LP) | |
| Tongue Tip | Tongue tip constriction degree (TTCD) | Tongue body, tip, jaw |
| | Tongue tip constriction location (TTCL) | |
| Tongue Body | Tongue body constriction degree (TBCD) | Tongue body, jaw |
| | Tongue body constriction location (TBCL) | |
| Velum | Velum (VEL) | Velum |
| Glottis | Glottis (GLO) | Glottis |

TVs are the outcomes of the action units defined by the gestural scores.

Invariance property of gestures

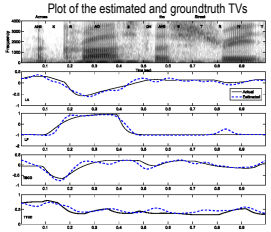



➢ Due to change of speech rate, intra-gestural dynamics (e.g. gestural reduction) and inter-gestural timing (e.g. increased overlap) can be altered, resulting in big acoustic changes
➢ However, the overall gestural pattern remains the same
➢ To obtain articulatory information from speech we need to perform 'speech inversion'

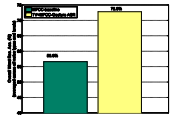



➢ Traditionally flesh-point based articulatory information has been used for speech-inversion.
➢ We have shown TVs are better for speech inversion
✓ TVs can be estimated more accurately than flesh-point data.
✓ TVs being relative measure demonstrates less non-uniqueness in the speech inversion problem than the absolute measure flesh-point trajectories.

Plot of the estimated and groundtruth TVs



The estimated TVs were used in a gesture-based speech recognition architecture where the input speech was noise contaminated with 8 different noise types at 7 signal to noise ratios from clean to -5dB



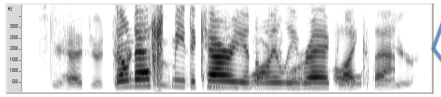
Use of articulatory information significantly improved the word recognition performance


### Acknowledgement

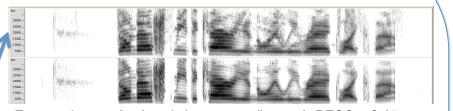
This research was supported by NSF Grant # IIS0703859, IIS-0703048 and IIS0703782.


## Speech Extraction Technology

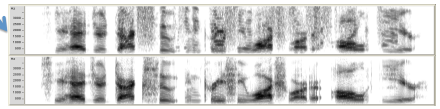
• Existing speech enhancement techniques aim at suppressing noise depending on statistical characteristics of the noise



Speech mixture at 6dB target to masker ratio, PESQ (target) = 2.28, PESQ (masker) = 1.34

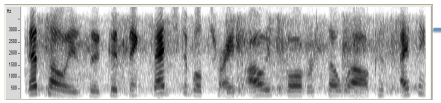


Extracted target (top) and clean target (bottom), PESQ = 3.15

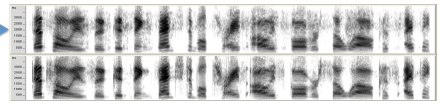


Extracted masker (top) and clean masker (bottom), PESQ = 2.06

• Our speech extraction technology works by modeling the speech and can be applied for both stationary (e.g., subway) and non-stationary (e.g., speech babble) background noise
• An objective measure for evaluating speech quality is the perceptual evaluation of speech quality (PESQ) which ranges from 0.5 (highly degraded) to 4.5 (perfect quality)



Speech in subway noise at 10dB signal to noise ratio, PESQ = 1.84

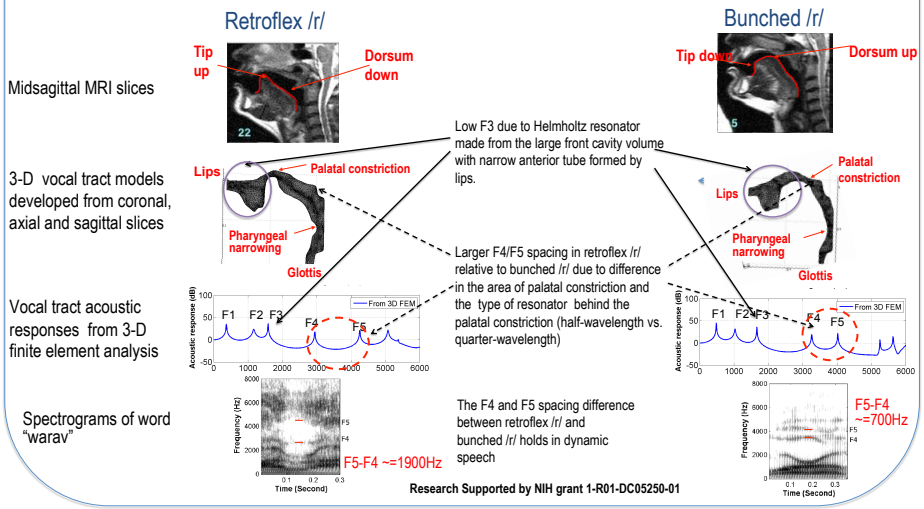


Extracted speech (top) and Clean speech (bottom), PESQ = 2.80

• Received University of Maryland, "Invention of the Year" award for multipitch estimation algorithm
• Technology can improve sound quality in communication devices like hearing aids and cell phones and improve the performance of other speech based technologies
• Research supported by NSF grants IIS-0812509 and BCS-0519256

## Speech production: Many-to-one articulatory to acoustic mapping

It has been long claimed that the acoustics of American English /r/ don't change as a result of the many ways in which this sound can be produced (a continuum between retroflex /r/ and bunched /r/. We have been able to show why the salient low F3 cue is stable regardless of vocal tract shape and, more importantly, that there are acoustic signatures in F4 and F5 for the various articulatory configurations.



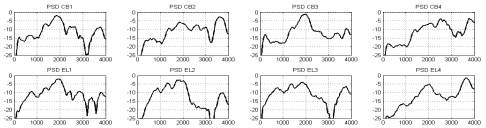


Research Supported by NIH grant 1-R01-DC05250-01


## Speech forensics: device identification and media authentication

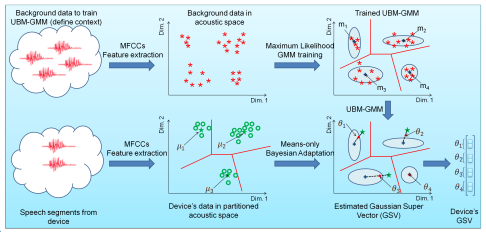
- MOTIVATION: imperceptible alterations of digital speech content pose a serious threat to a wide variety of fields such as intellectual property, criminal investigation, and law-enforcement
- GOAL: automatic extraction of forensic evidence about the mechanism involved in the generation of the speech recording by analysis of the acoustic signal
- NON-INTRUSIVE PARADIGM: Only have access to actual speech recordings

Magnitude squared of the frequency response (in dB) of 8 landline handsets.

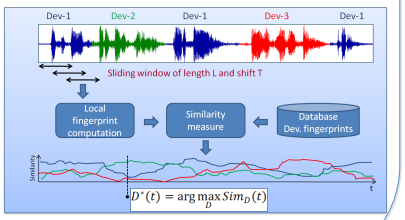



Confusion matrix for telephone handsets. Average accuracy 93.2 %

| | CB1 | CB2 | CB3 | CB4 | EL1 | EL2 | EL3 | EL4 |
|---|---|---|---|---|---|---|---|---|
| CB1 | **94.3** | 0.6 | 0.3 | 1.1 | 0.9 | 0.2 | 0.0 | 2.5 |
| CB2 | 1.6 | **97.0** | 0.3 | 0.2 | 0.5 | 0.0 | 0.5 | 0.0 |
| CB3 | 0.2 | 0.0 | **99.2** | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 |
| CB4 | 0.2 | 0.0 | 0.9 | **98.7** | 0.0 | 0.0 | 0.0 | 0.2 |
| EL1 | 1.1 | 2.5 | 0.0 | 0.2 | **86.9** | 2.5 | 3.3 | 3.5 |
| EL2 | 0.2 | 0.3 | 0.0 | 0.0 | 3.3 | **92.9** | 2.8 | 0.5 |
| EL3 | 0.3 | 1.9 | 0.0 | 0.0 | 7.4 | 6.2 | **83.4** | 0.8 |
| EL4 | 1.3 | 0.0 | 0.3 | 0.5 | 3.5 | 0.6 | 0.6 | **93.2** |




Mechanism to compute the a statistical intrinsic fingerprint of a device.




Use of device intrinsic fingerprint to detect multi-device composites.


Supported by NSF grant IIS-0917104