Bridging the Gap: Using Deep Acoustic Representations to Learn Grounded Language from Percepts and Raw Speech

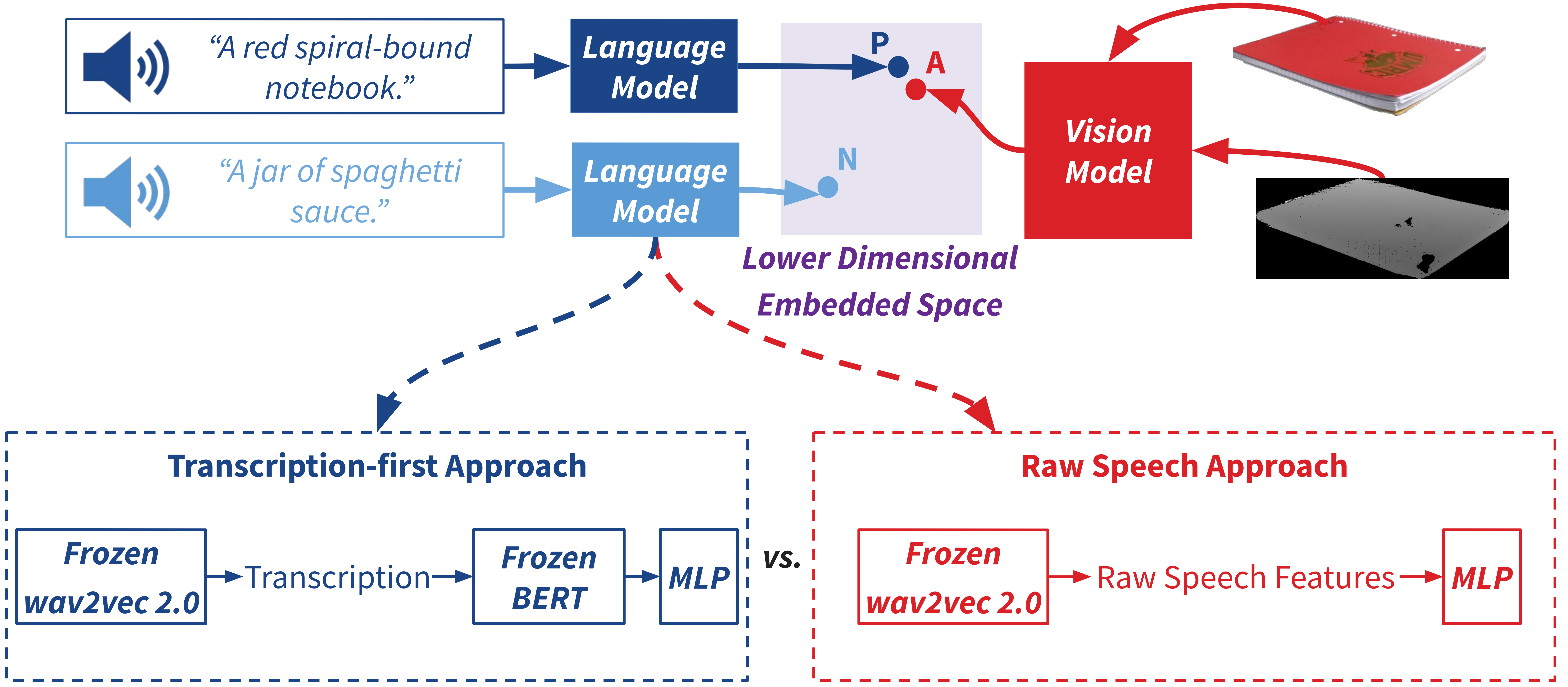
Gaoussou Youssouf Kebe,
Luke E. Richards, Edward Raff,
Francis Ferraro, Cynthia Matuszek

- **Grounded Language Learning** → Learning **natural language** as it relates to **sensory perception**
- We skip **transcriptions** and learn groundings from **raw speech** and visual percepts
- Speech **improves performance** and reduces performance differences across diverse groups

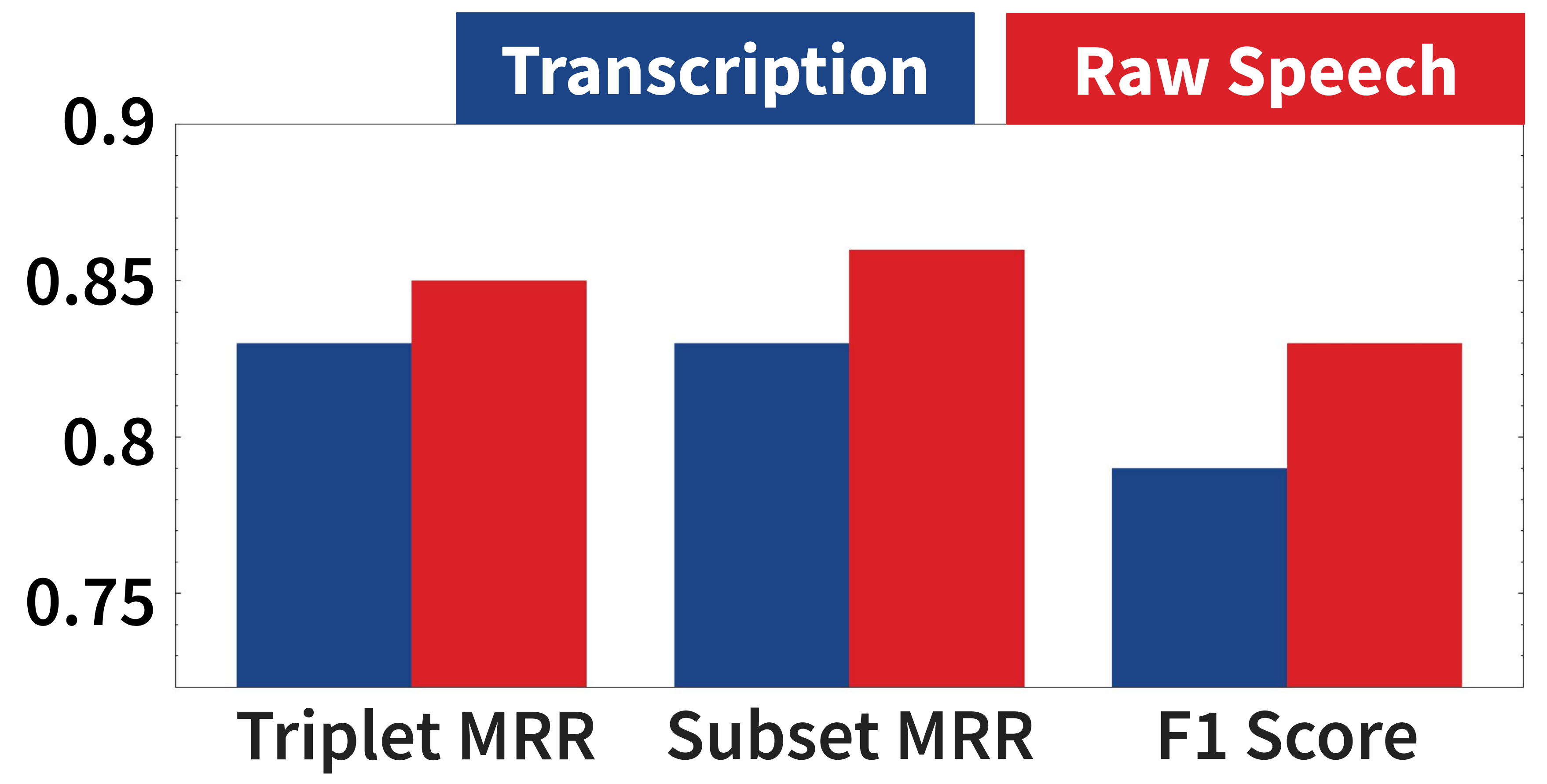
GoLD Dataset

RGB		*“It is a can with a pull top from Goya”*	**16500** Spoken Descriptions
Depth		*“This is a can of navy beans with a blue label”*	**552** Unique Speakers

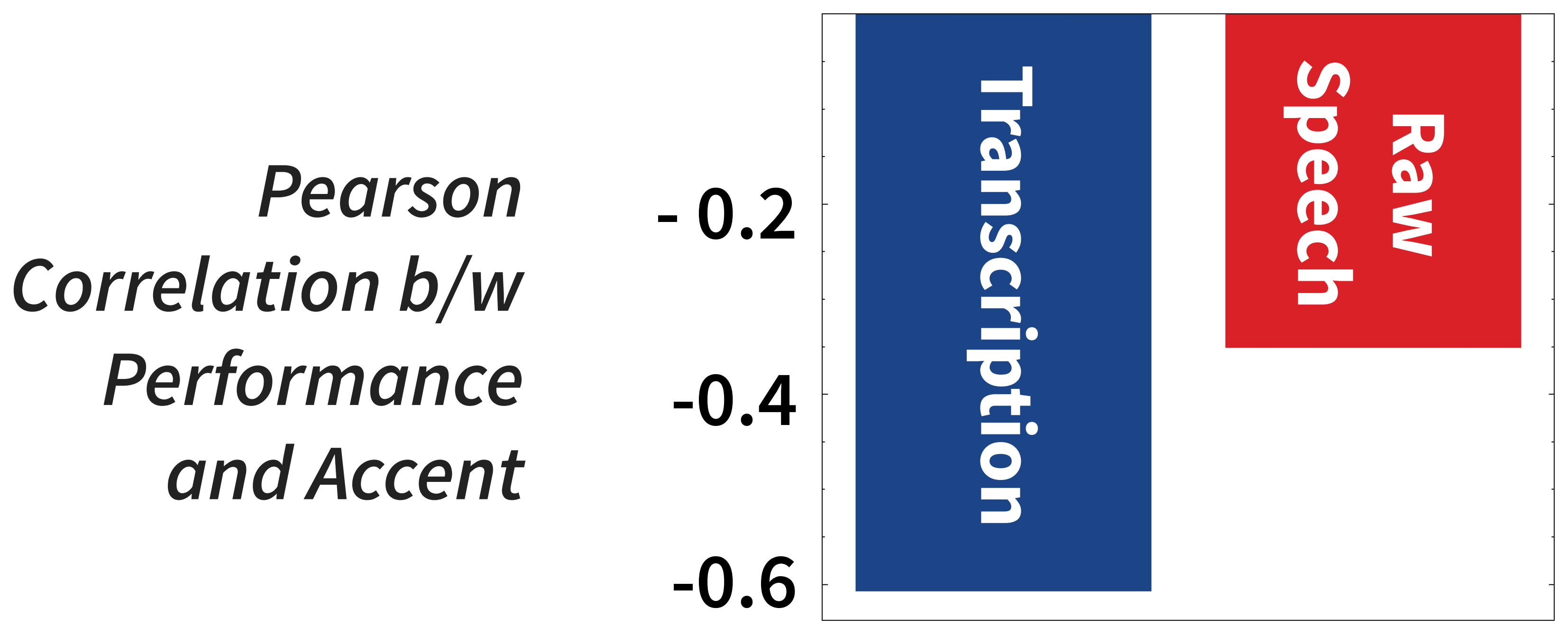
Triplet loss → *max(d(A, P) - d(A, N) + a, 0)*



Raw Speech approach improves performance



Raw Speech approach is more inclusive towards accented users


Pearson Correlation b/w Performance and Accent
Transcription
Raw Speech