Speaker Recognition from Coded Speech in Matched and Mismatched Conditions[1]

Bob Dunn
Tom Quatieri
Doug Reynolds
Joe Campbell[2]

MIT Lincoln Laboratory
[2]Department of Defense

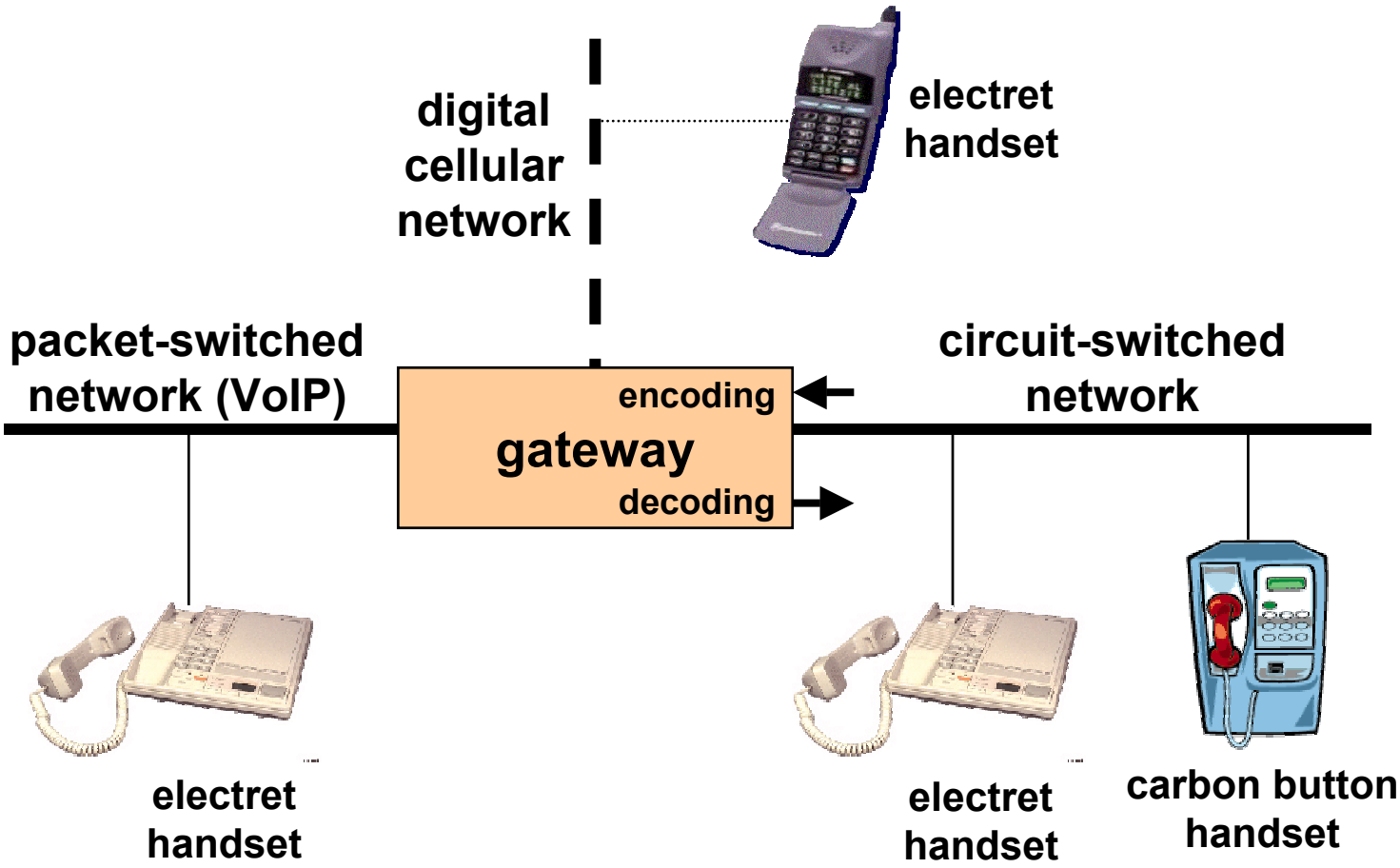
Odyssey
June 2001

[1]This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

Outline

- Introduction
- Speaker Detection System
- Experiments and Results
- Conclusions and Future Directions

Networks of Interest



- **Issues not addressed:** VoIP packet loss, cellular system transmission loss, speech coding prior to transmission in circuit-switched network

Recognition from Coded Speech

- **Why Coded Speech?**
  - **Past:**

    Data traffic was sent over analog voice networks
  - **Present and Future:**

    Voice is transmitted over digital data networks where speech coders are used for compression
  - **Applications:**

    Digital cell phone systems, satellite communication systems, VoIP networks

- **How does the speaker recognition system perform when the speech has been coded?**

- **How does a channel mismatch between model training and utterance recognition affect performance?**

Recognition from Coded Speech

- How does the speaker recognition system perform when the speech has been coded?
- How does a channel mismatch between model training and utterance recognition affect performance?

Classes of Speech Coders

- **Toll quality**
  - Quality as good as telephone system: 7-12 kb/s
  - Digital cellular, VoIP systems
  - Mixed parametric/waveform coders
    - Robust to non-speech signals
- **Medium quality**
  - Not as good as telephone system: 5-7 kb/s
  - VoIP systems
  - Mixed parametric/waveform coders
- **Communication quality**
  - Reduced speech quality but high intelligibility: 1-5 kb/s
  - Satellite telephone, secure voice (STU-III)
  - Completely parametric
    - Often poor performance for non-speech signals

Examples of Speech Coding

- **Original**  **64.0 kb/s**
  Switchboard telephone speech (μ-law)

- **GSM**  **12.2 kb/s**
  European digital cellular standard (ETSI Standard)

- **G.729**  **8.0 kb/s**
  Used for digital cellular and VoIP (ITU Standard)

- **G.723**  **6.3 kb/s and 5.3 kb/s**
  VoIP Standard (ITU Standard)

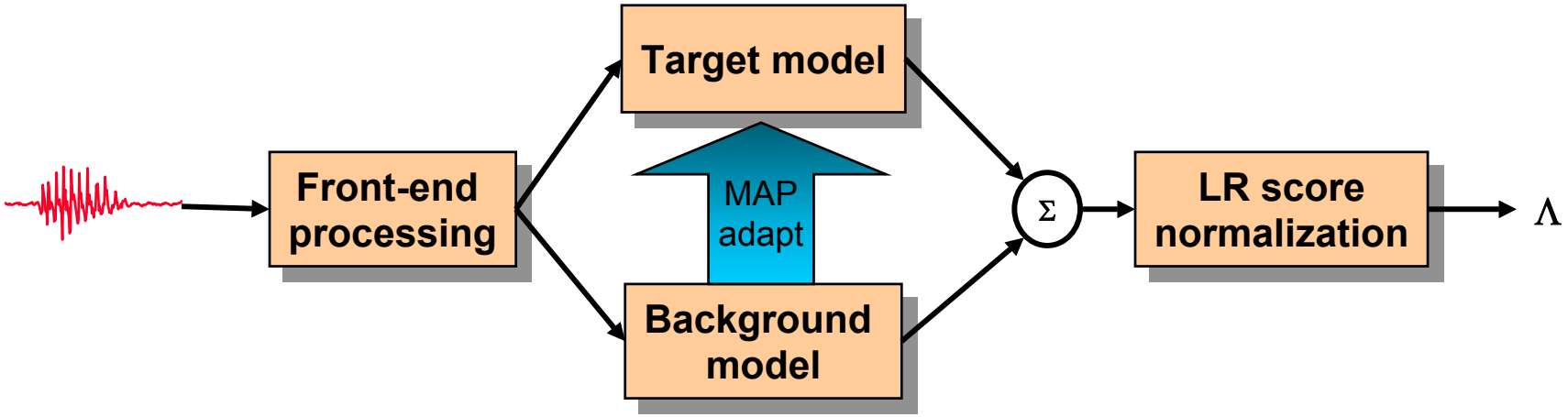
- **MELP**  **2.4 kb/s**
  Designed for satellite communications and secure voice
  (US Federal Standard)

Outline

- Introduction
- **Speaker Detection System**
- Experiments and Results
- Conclusions and Future Directions

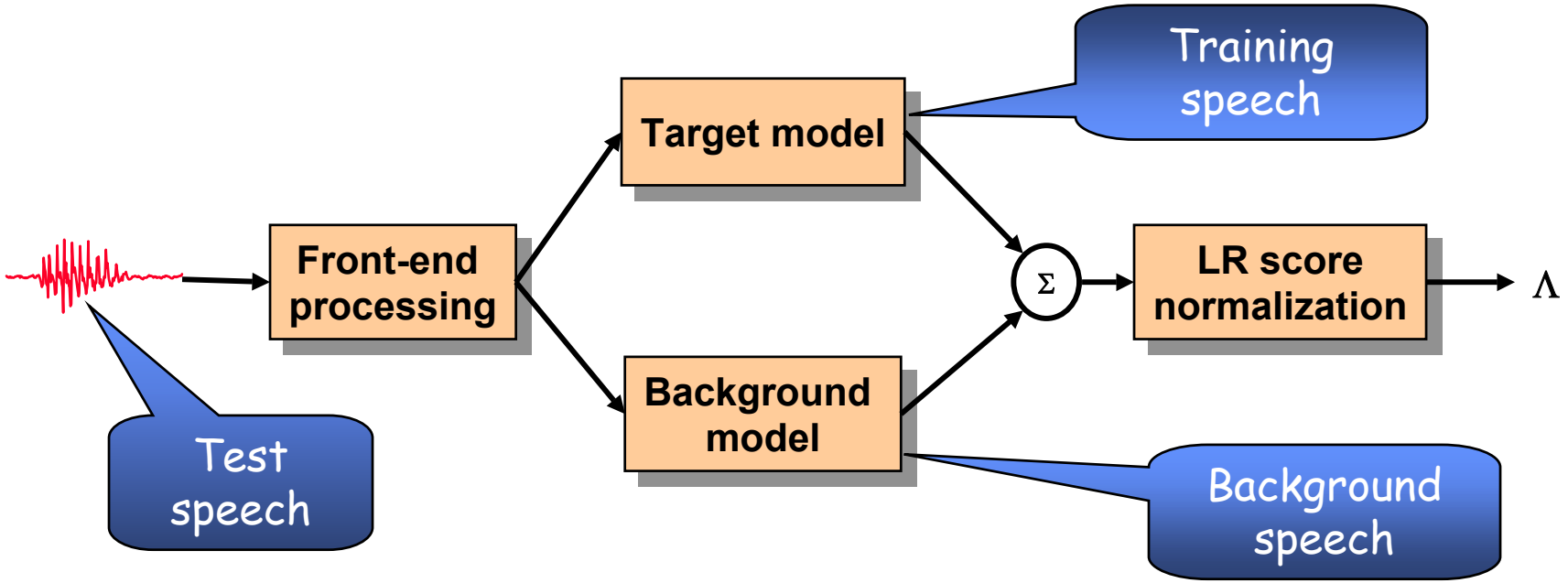
Speaker Detection System
GMM-UBM System

- Basic decision statistic is the likelihood-ratio



- **Front-end processing**: speech activity detection; bandlimited mel-filterbank cepstra with deltas; RASTA filtering
- **Background model**: 2048 Gaussian Mixture Model (GMM) trained on gender/handset balanced data
- **Target model**: Bayesian (MAP) adapted GMM derived from background model
- **Score normalization**: Handset and coder dependent score normalization

Coded Speech

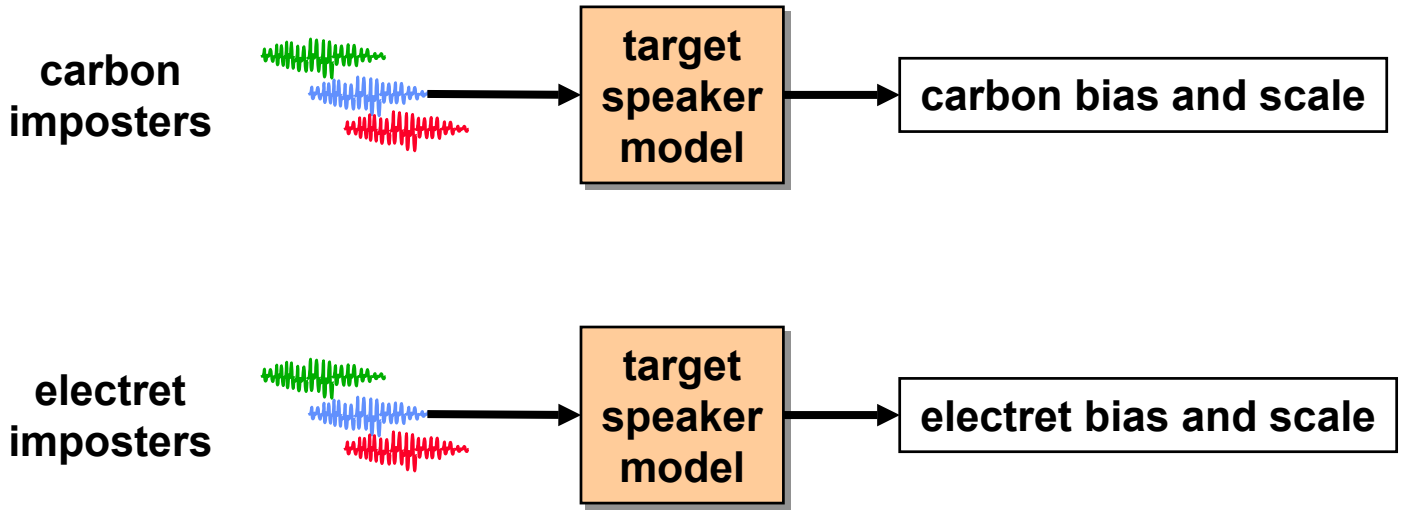


- Coded speech can occur in three locations in the system

	MATCHED BASELINE	MATCHED CODED	TRAIN CODED	TEST CODED	MISMATCHED BACKGROUND
Background Model	Telephone	Coded	Coded	Telephone	Telephone
Training	Telephone	Coded	Coded	Telephone	Coded
Testing	Telephone	Coded	Telephone	Coded	Coded

Score Normalization
Hnorm

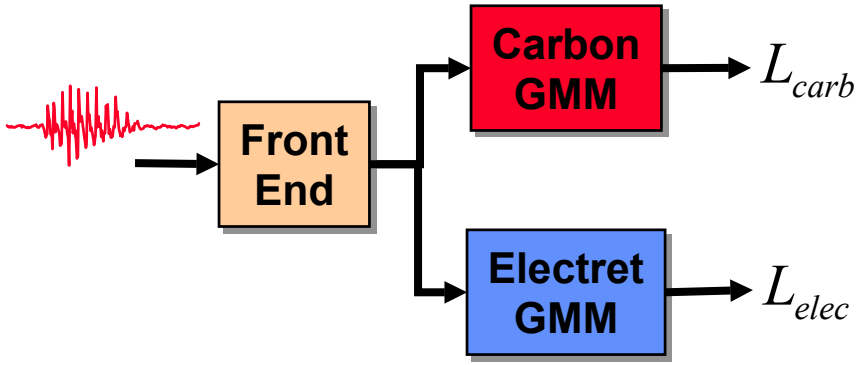
- Score normalization
  - Estimate and remove handset dependent bias and scale from target likelihood ratio score



- Handset-type detection is required for score normalization
  - Can handset-type be detected from coded speech?

Detection of Handset-Type
Coded Speech

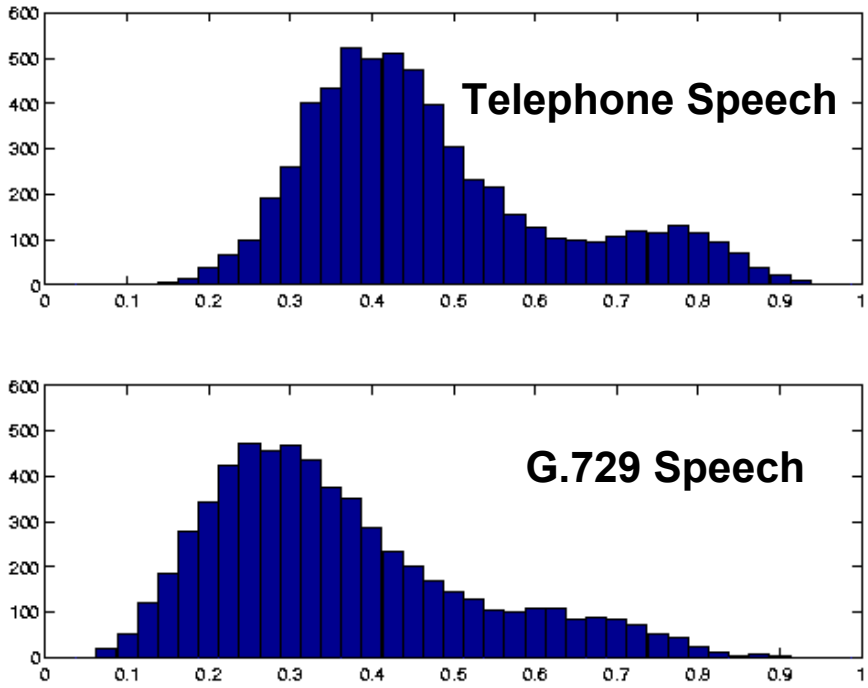
Compute Handset Likelihoods



$$P_{carb} = \frac{L_{carb}}{L_{carb} + L_{elec}}$$

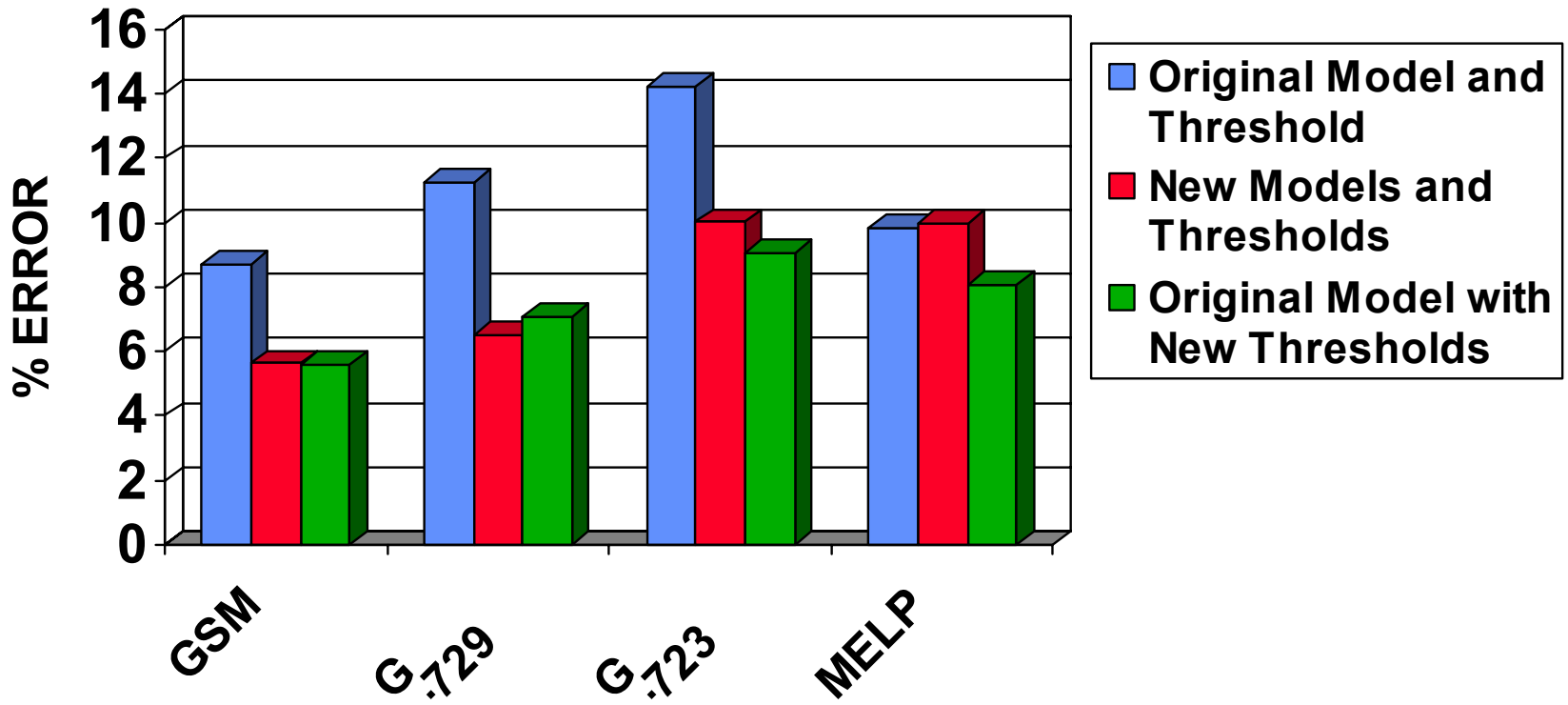
- **Histograms are clearly bimodal:**
  - **Distribution is shifted**
  - **Second mode is less distinct**

Histograms of Handset Probability



Posterior Probability of Carbon

Evaluation of Handset-Type Labels



- Errors relative to handset labels in NIST header
- Using a new handset models trained on coded speech does not significantly improve performance
- Using a coder dependent threshold with original model improves performance

Outline

- Introduction
- Speaker Detection System
- **Experiments and Results**
- Conclusions and Future Directions

Experimental Scenario

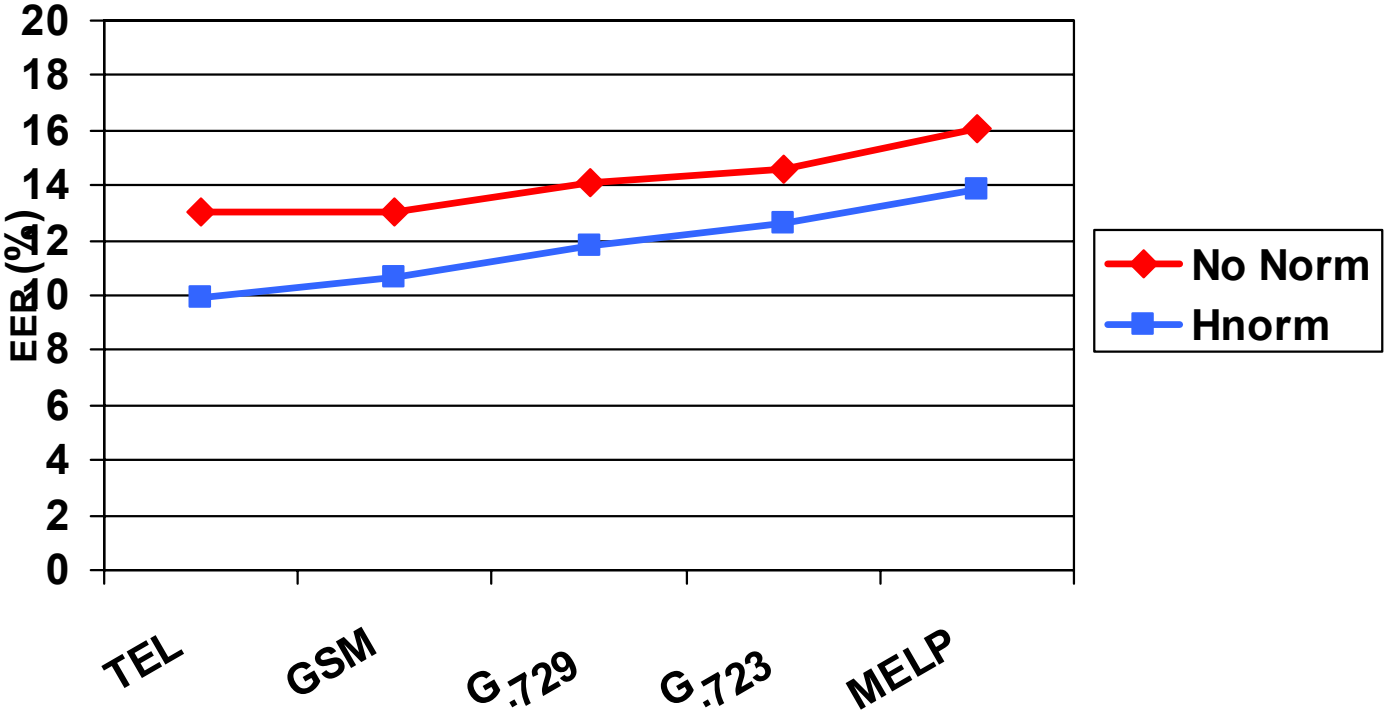
- **Database**
  - **Evaluation data:**
    Switchboard-II, phases 1 and 2
  - **Background model and normalization data:**
    Switchboard-II, phase 3

- **Switchboard speech is encoded and decoded with a speech coder to simulate transmission of telephone speech through a VoIP or digital cellular gateway**

- **Conditions**

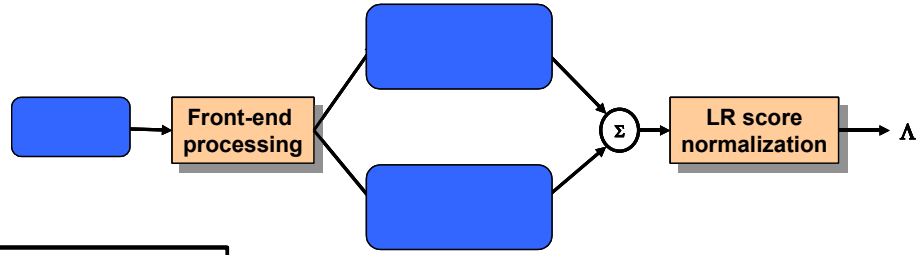
	MATCHED BASELINE	MATCHED CODED	TRAIN CODED	TEST CODED	MISMATCHED BACKGROUND
Background Model	Telephone	Coded	Coded	Telephone	Telephone
Training	Telephone	Coded	Coded	Telephone	Coded
Testing	Telephone	Coded	Telephone	Coded	Coded

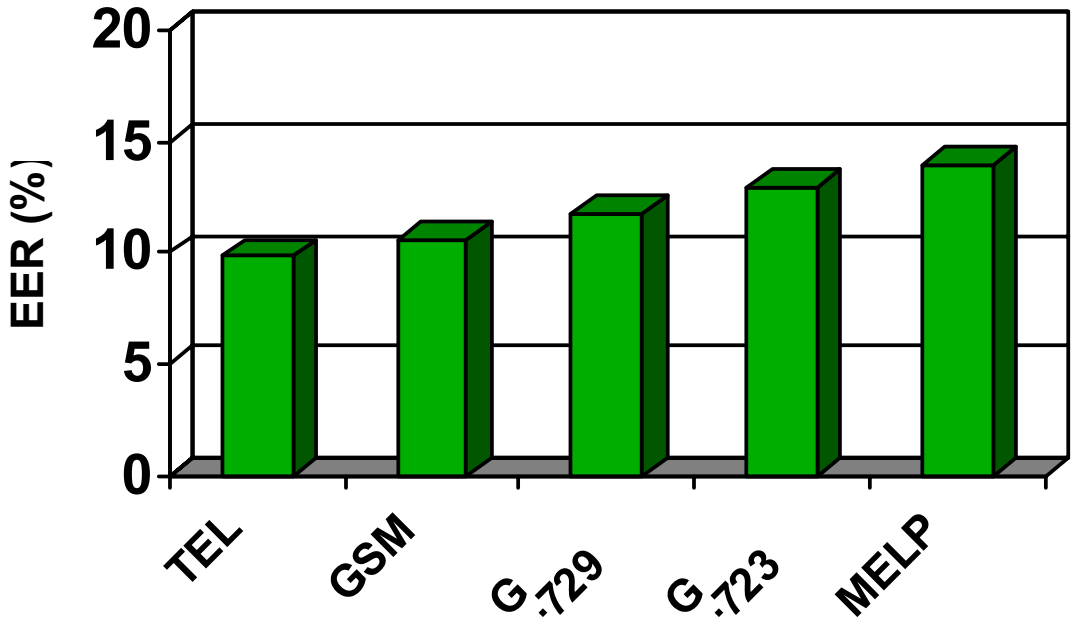
Matched Conditions
Score Normalization



- Coded speech is fully matched with respect to background model, training data and testing data
- EER increases as coder rate decreases
- Score normalization works in all conditions
  - handset and coder dependent
  - reduces EER by 2-3%

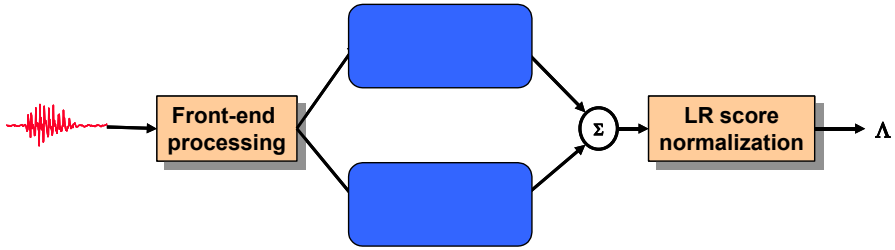
Mismatched Conditions
All Coded

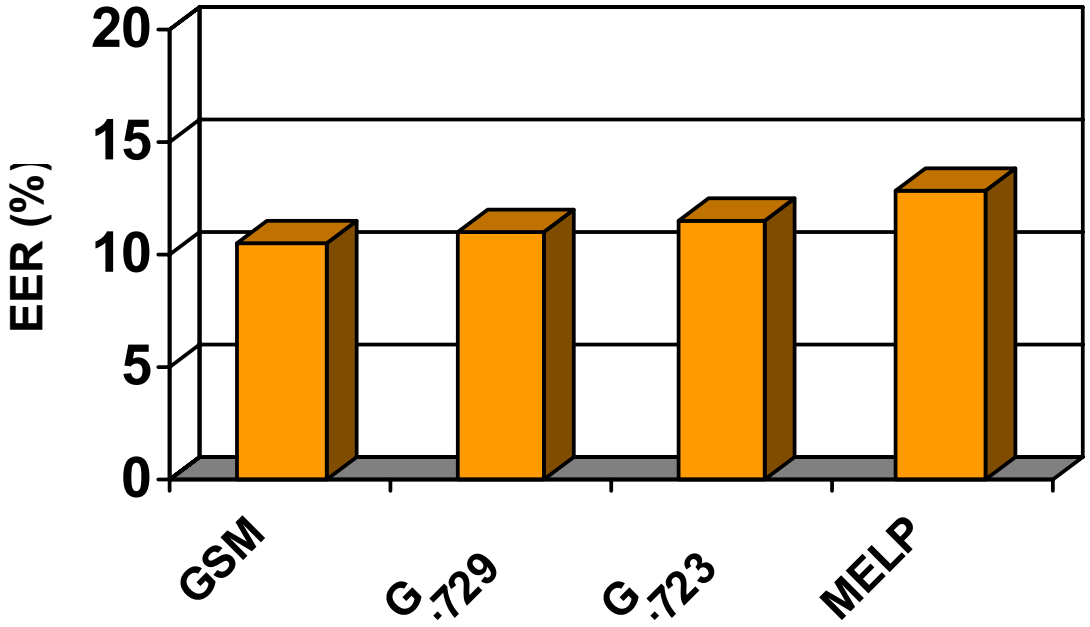




- Speaker detection results with handset and coder dependent score normalization

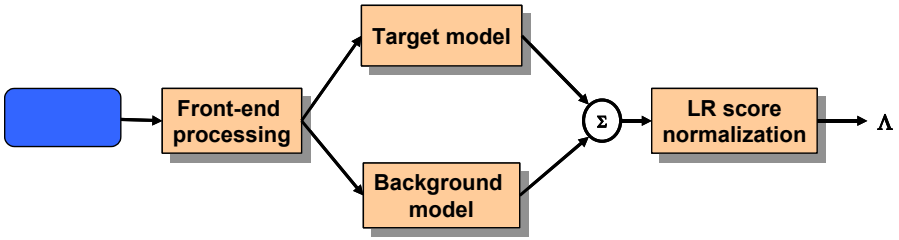
Mismatched Conditions
Bkg and Tgt Coded

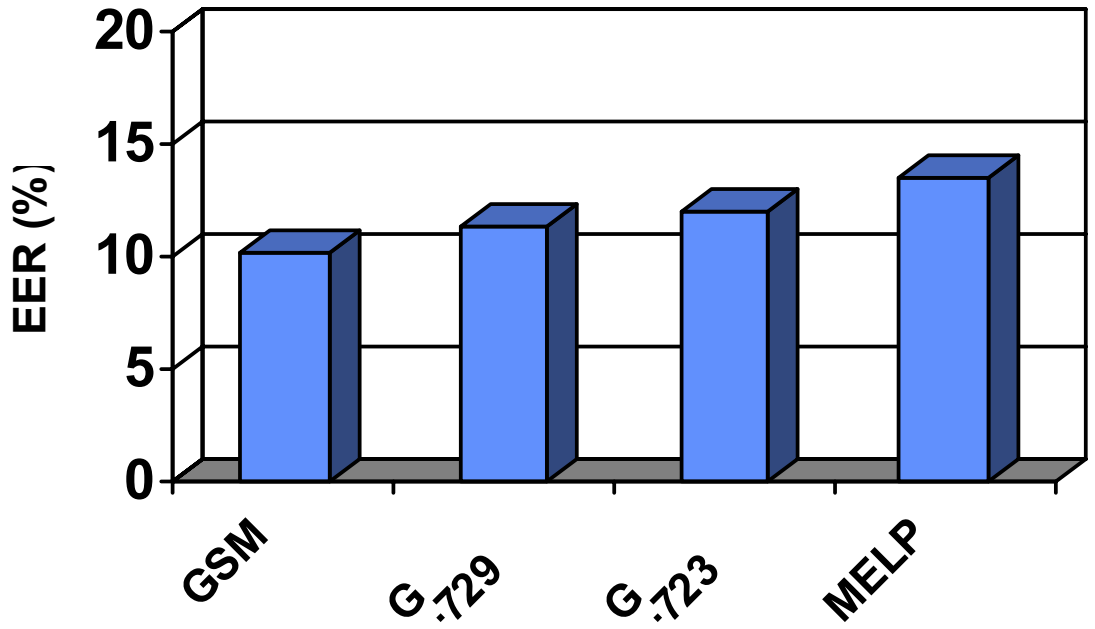




- Slightly better than all coded condition

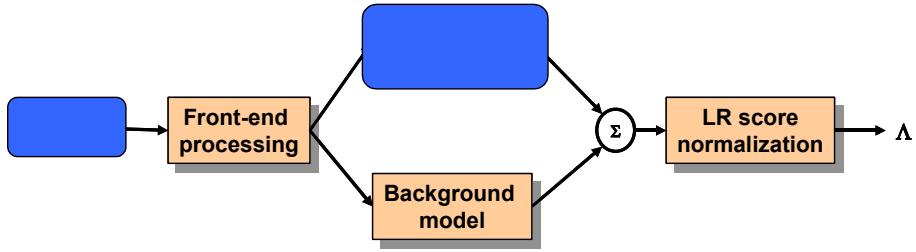
Mismatched Conditions
Test Coded

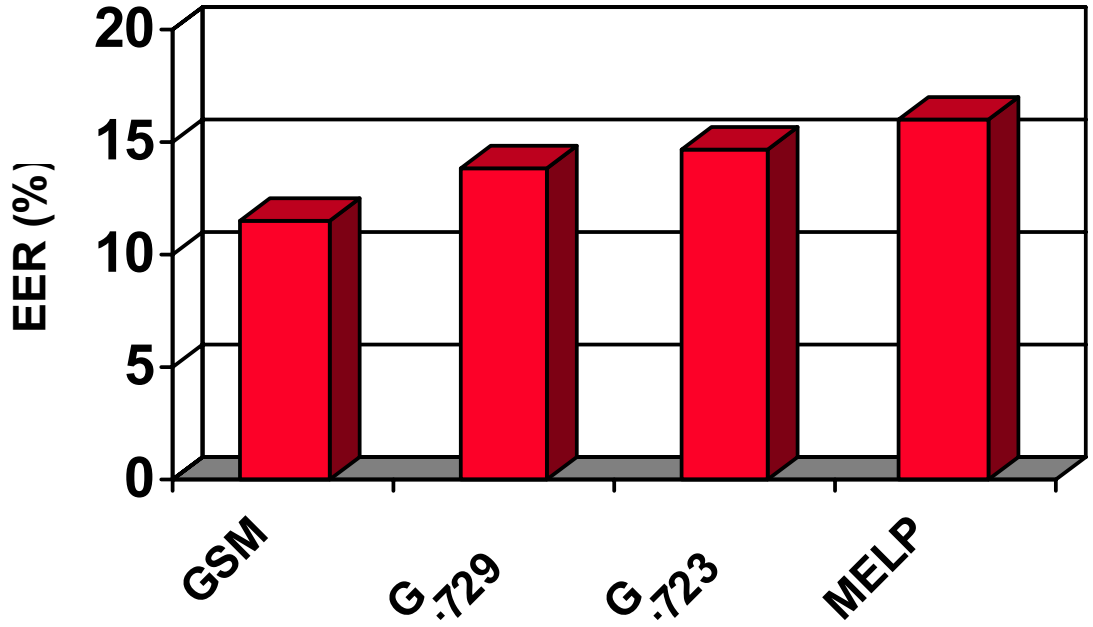




- Slightly better than all coded condition

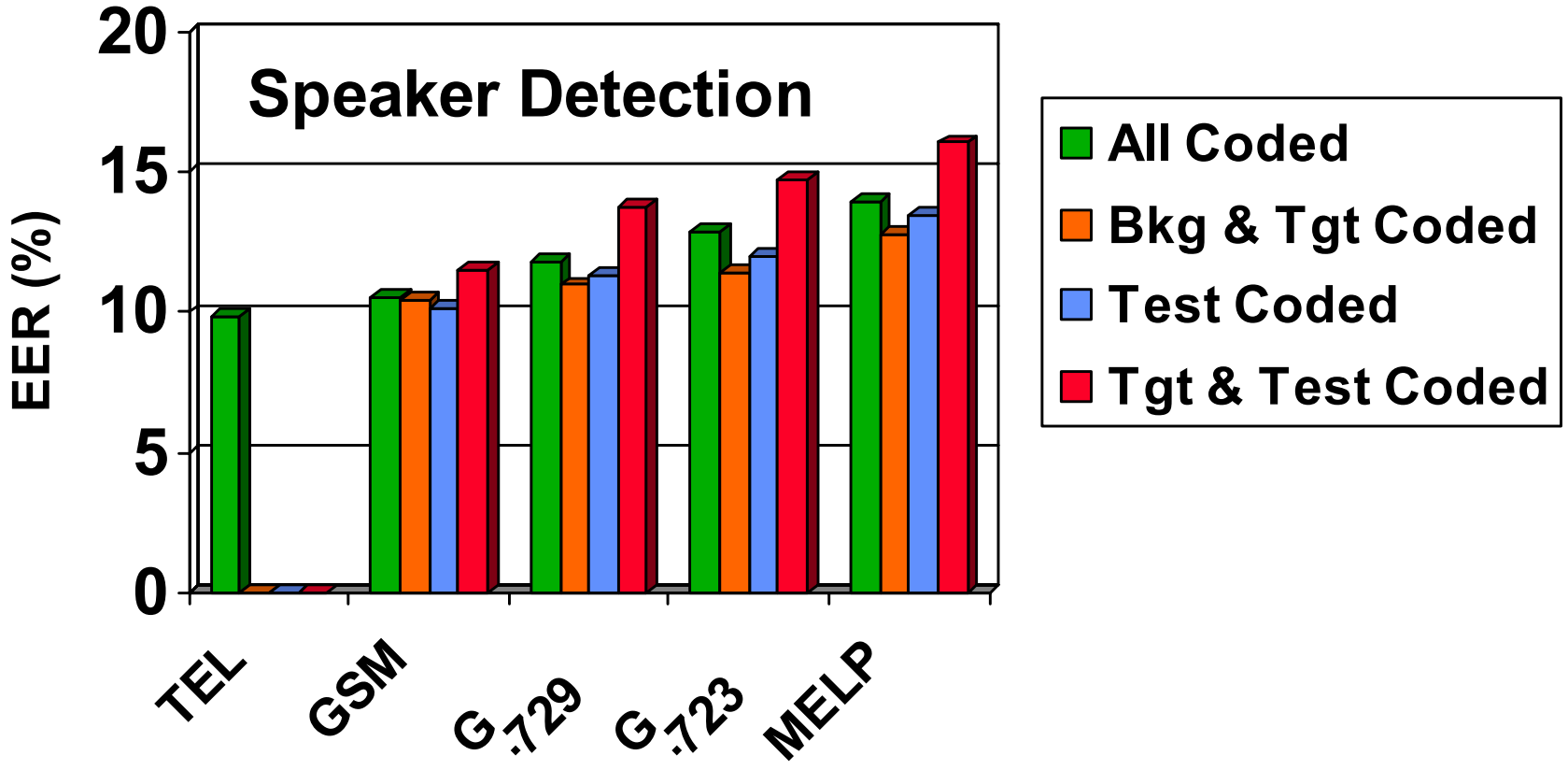
Mismatched Conditions
Tgt and Test Coded





- Mismatched background model has highest error rates

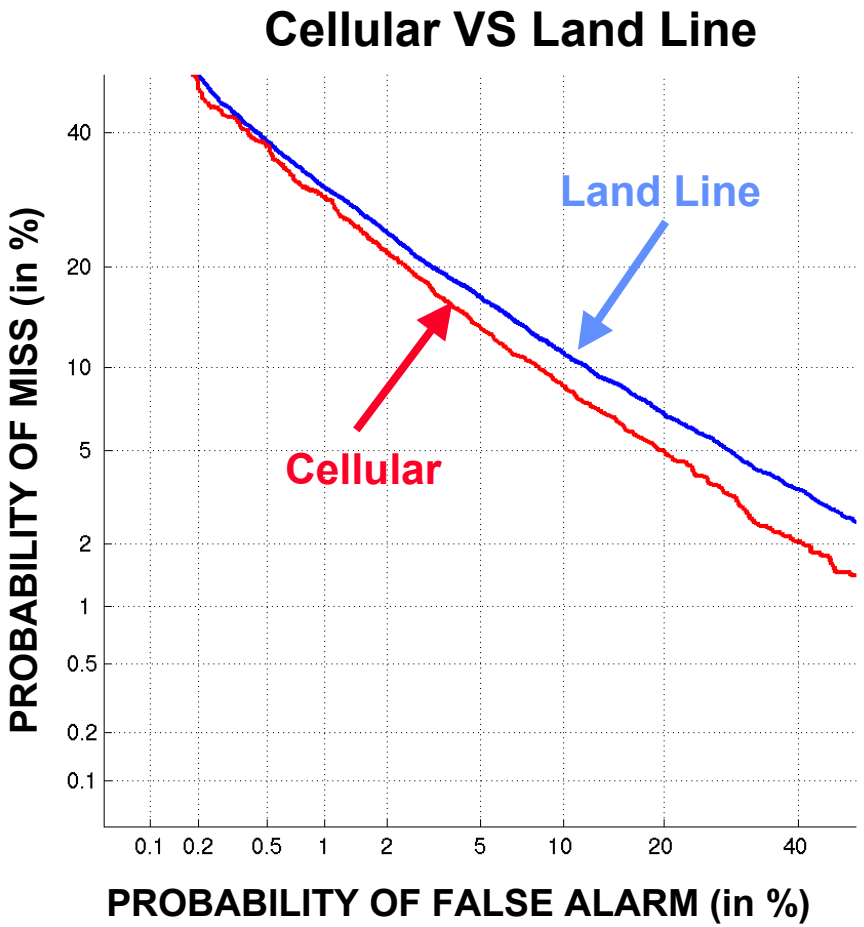
Mismatched Conditions



- Coded training or testing is as good or better than all coded
- Mismatched background model has reduced performance

Recent NIST Results

- NIST 2001 evaluation included a new Cellular database
- Cellular data is predominantly GSM (but database includes some land line)
- GMM-UBM system with Synthetic-Model-Synthesis (SMS) and Tnorm (no Hnorm)
- SMS with Tnorm is not the best performing system for land-line data



Recent NIST Results
Misc. Results

- Gender identification on cellular data
  - 1.3% error (27/2028)

- Cellular versus landline detection
  - ERR = 13%
  - Most errors are cell (gsm) calls being detected as landline
    - Probably from high quality cell calls

Outline

- Introduction
- Speaker Detection System
- Experiments and Results
- Conclusions and Future Directions

Conclusions

- Demonstrated speaker detection performance for speech coded by GSM, G.729, G.723, MELP
- Error rate increases as bit-rate (and speech quality) decreases
  - But not as dramatically as reported for speech recognition
- Handset-type detection can be used on coded speech
- Handset dependent score normalization (Hnorm) improves speaker detection for coded speech…error rate still increases as bit-rate decreases
- Background model training data should match target speaker training condition

Future Work

- Evaluate test score normalization (Tnorm) on coded speech
- Experiment with the use of a composite background model
- Test performance on directly coded speech (without local analog loop) (new database required)
- Examine effects of packet-loss and transmission errors
- Perform speaker detection directly from speech coder parameters
  - Initial experiments have found only small degradation