Combining Heterogeneous Sensors with Standard Microphones for Noise Robust Recognition

Horacio Franco[1], *Martin Graciarena*[1,2]
Kemal Sonmez[1], *Harry Bratt*[1]

[1] SRI International
[2] University of Buenos Aires

General Problem & Approach

- **Problem**
  - Current speech recognition systems are brittle with regard to changes in the acoustic environment. Need to increase robustness!
- **Approach**
  - Enrich standard microphone signal stream with multiple additional speech signals from *alternative sensors.*
- **Rationale**
  - Alternative sensors may be more isolated from environmental noise $\Rightarrow$ convey complementary robust information about signal components degraded with a standard microphone

Alternative Sensors

- **Throat, ear, skull microphones:** Alternative, more robust paths for some signal components.
- **Electroglottography (EGG):** A technique used to register laryngeal behavior indirectly by a measuring the change in electrical impedance across the throat during speaking.
- **Glottal Electro Magnetic Sensors (GEMS):** Low power radar-like sensor, can measure conditions of articulators, in particular voice excitation. (Lawrence Livermore Labs)
- **Nasal accelerometers:** Measure of nasal airflow.

Problems

- How to fuse both microphones' data to improve noisy recognition
- How to train acoustic models; (with very little "stereo" data available)

Proposed Approach

- Extend the Probabilistic Optimum Filtering (POF) technique to map noisy standard and throat microphones features, juxtaposed as an extended vector $\Rightarrow$ <u>estimate clean std microphone feature</u> (mel-cepstra features).
  - First problem: estimated std microphone features computed in MMSE sense to real clean std microphone features
  - Second problem: need for small to medium "stereo" database. Estimated std microphone features can be recognized with SRI's DECIPHER system

POF Introduction

• POF mapping is a <u>piece-wise linear transformation</u> from noisy feature space to clean feature space.

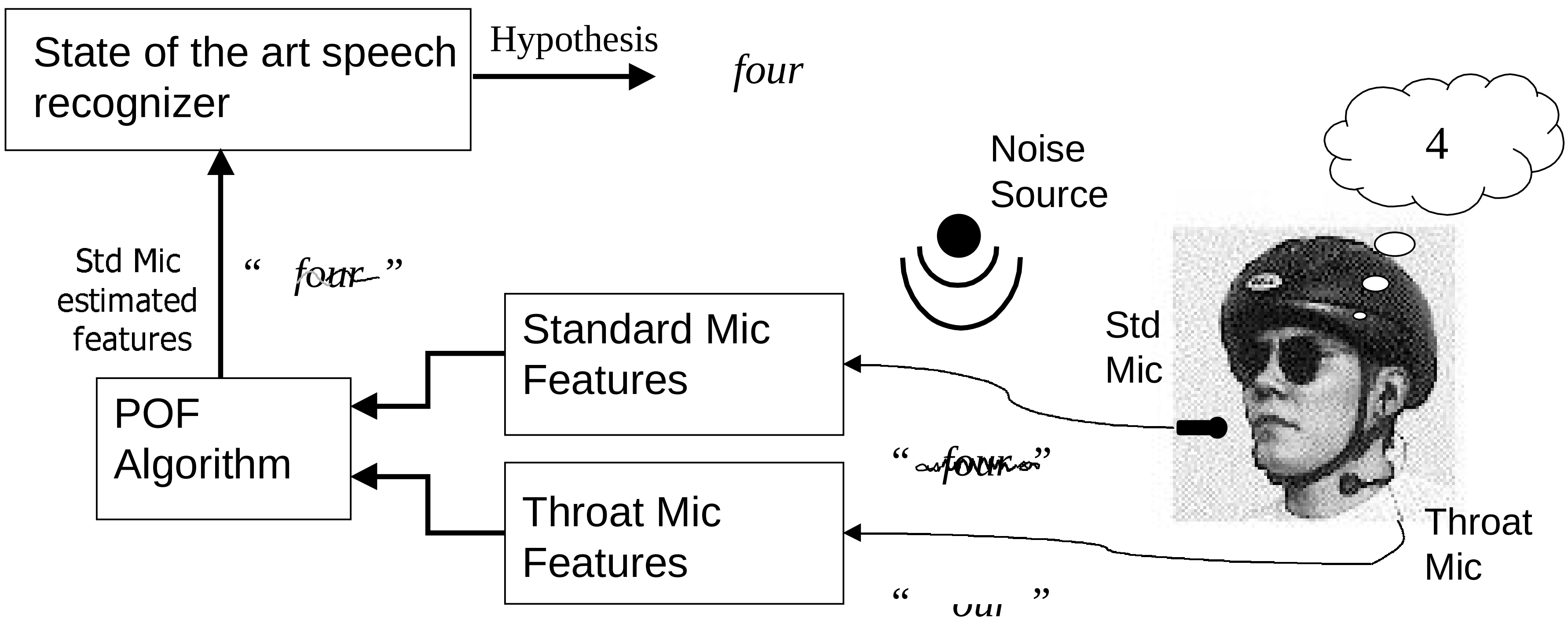
• Each linear transformation assigned to region in a VQ partition of noisy feature space.

• Estimated clean feature vector:

– Compute Posterior probabilities of VQ regions using a <u>conditioning vector</u> (derived from noisy feature vector)

– Compute set of linear transformations weighted by the posterior probabilities from noisy speech feature vector (one or more time adjacent frames (window parameter)).

Standard and Throat Microphone Feature Combination



** Combined microphones provide clearer picture!*

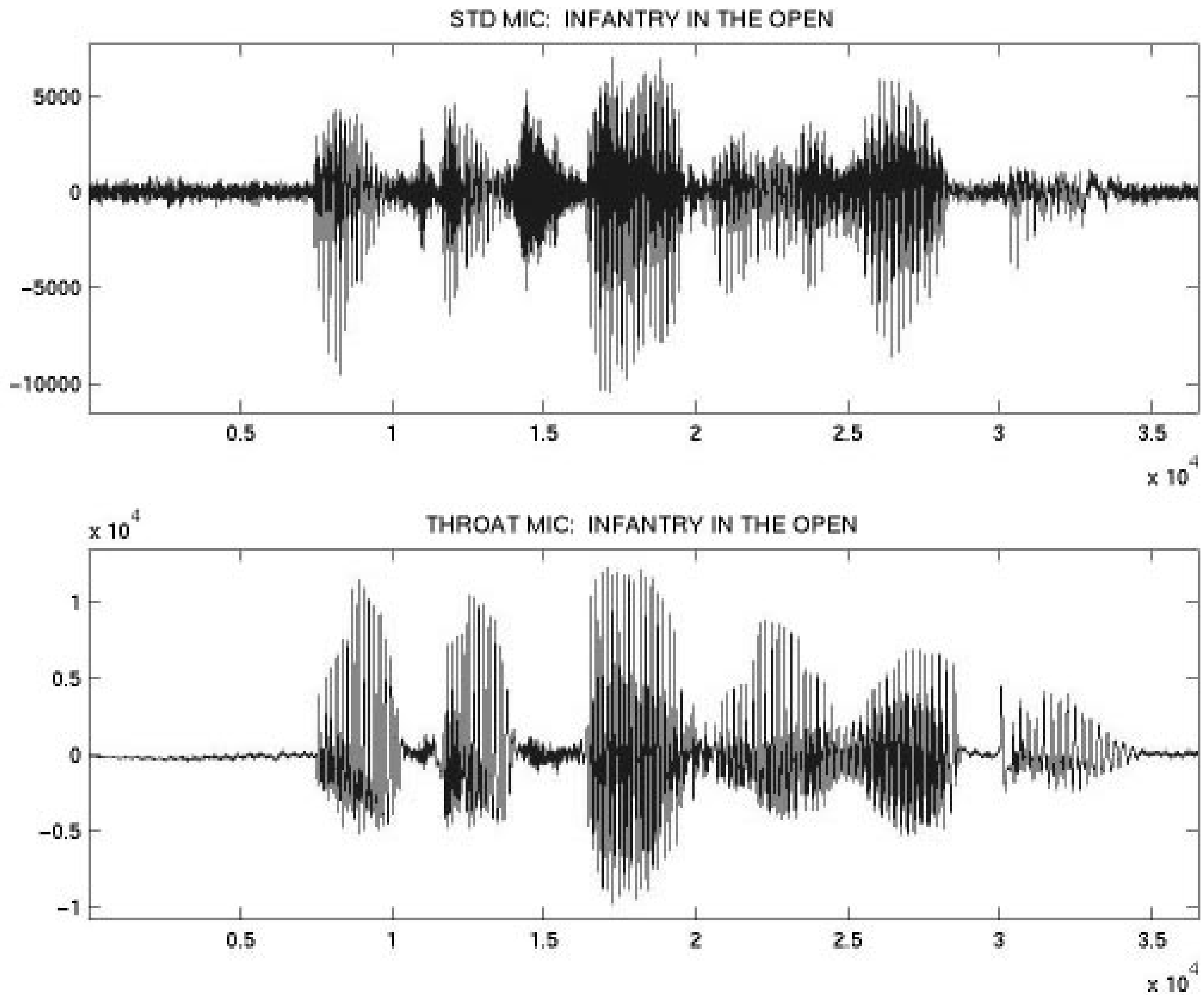
** Noise affects mostly Std microphone signal*

** Throat microphone signal almost immune to noise but has partial information*

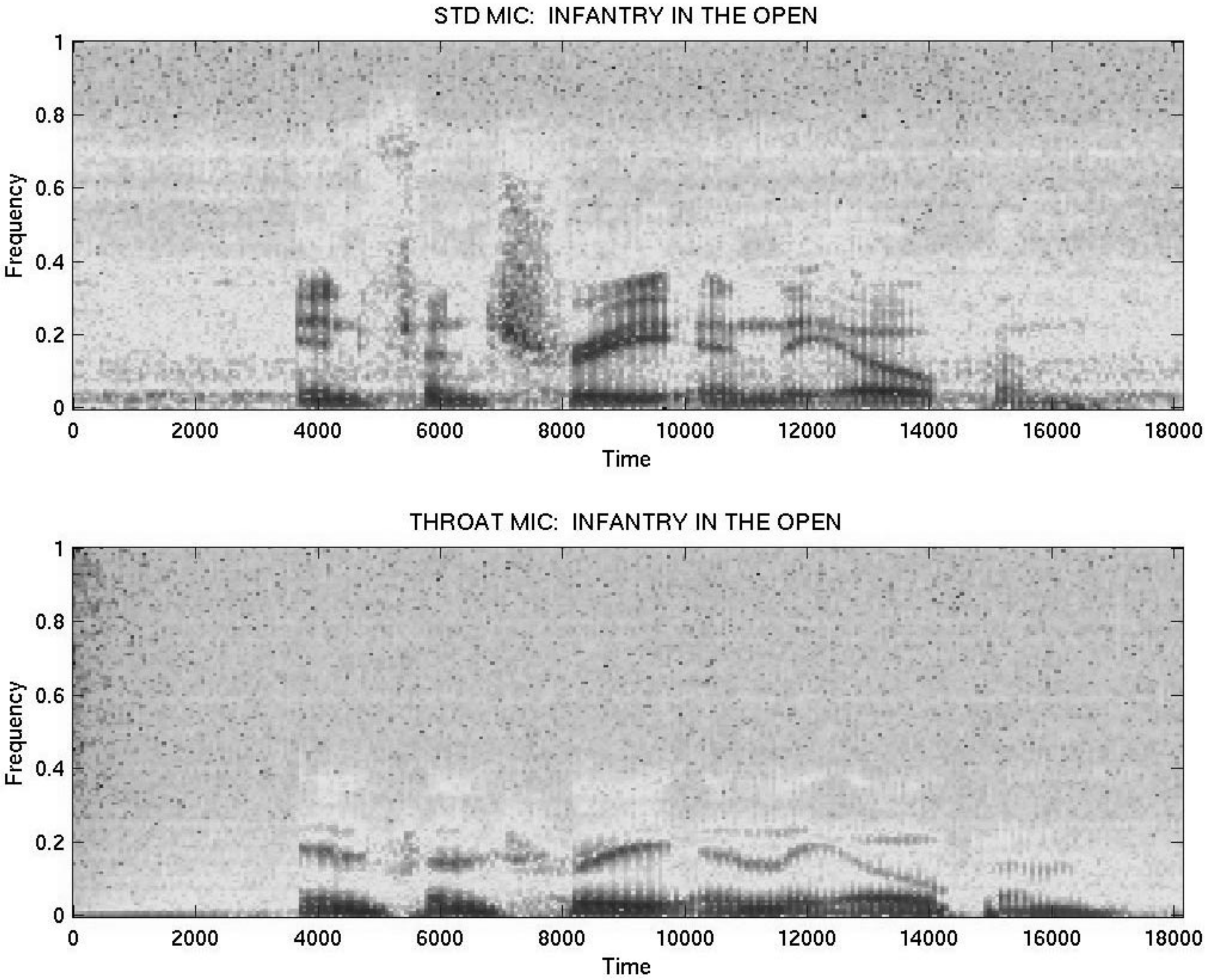
Throat Microphone

- Its a skin vibration transducer $\Rightarrow$ Highly immune to environmental noise due to <u>close contact with throat skin</u>.
- What type of Info it gives?
  - Robust voicing information
  - Some spectral information
- Production model for throat microphone signal $\Rightarrow$ Multipath signal?
- Robustness analysis: environmental noise energy captured by throat microphone is ~10 times lower than std microphone noise energy!

Std and Throat Microphone Signals



Std and Throat Microphone Signals



EXPERIMENT 1: *Artificially Added Noise*:

- M1 Tank noise artificially added only to std microphone in POF training database.
- Trained POF mappings from noisy features (std and std+throat) to std clean in "stereo" database
- Recognized noisy testing database. SNR's: Clean, 10dB, 6dB and 0dB. Mapped noisy to clean features with POF.
- *G3 company* corpus. Databases: POF training, 975 sent. & 30 speakers, Testing, 70 sent. & 7 indep. Speakers.
- Acoustic models: H4'98, adapted on the POF training database. LM: weighted combination of bigram LM's trained on H4 and Brown corpus. 5k vocabulary (no OOV)

RESULTS EXPT 1: *Artificially Added Noise*:

Results: WER % (distortion)

Compensation Method	# Window, # VQ Regions	Clean	10dB SNR	6dB SNR	0dB SNR
No Compensation Standard Mic.		18.2%	51.3% (.831)	73.9% (.892)	95.6% (.975)
POF Compensation Standard Mic.	5,100		46.0% (.616)	57.7% (.681)	88.5% (.777)
POF Combined Mic. Mapping (Throat C0 only)	5,100		37.9% (.616)	49.1% (.677)	76.1% (.765)
POF Combined Mic. Mapping (Throat Full vector)	3,100		35.7% (.590)	46.7% (.643)	66.4% (.715)
POF Combined Mic. Mapping+VQ (Throat Full vector)	3,100		29.3% (.577)	37.9% (.625)	53.8% (.687)
MLLR Adaptation Unsupervised on POF train database with FB align and clean rec. transcripts			47.1 %	58.7 %	80.5 %

EXPERIMENT 2: *Recorded Noisy Speech*:

- Recognition of recorded M1 Tank noisy speech
- Approach: SNR varies across sentences! Have to use SNR dependent mapping
- Estimate SNR ⇒ Apply mapping for that SNR
- Selected SNR's: >25dB (Clean), 8-12dB, 4-8dB.
- Used trained POF mappings from Expt. 1
- Database, SNR conditions: >25dB 91 sent, 8-12dB 116 sent, 4-8dB 75 sent.
- Same acoustic models as Expt. 1, same LM but had to interpolate uniform unigrams from test database. 5k Voc. (no OOVs)
- Database problems : click artifacts and misalignments!!

RESULTS EXPT 2: *Recorded Noisy Speech*

Results: WER %

Compensation Method	# Window, # VQ Regions	>25dB SNR	8-12dB SNR	4-8dB SNR
No Compensation Standard Mic.		19.6%	55.0%	62.4%
POF <u>Combined Mic.</u> Mapping+VQ (Throat Full vector)	3,100		41.2%	44.8%

Conclusions

• Proposed technique to combine noisy std and throat microphone features to estimate std microphone clean features.

• Experiments show robust complementary information is provided by the throat microphone.

Applications

• Robust recognition with throat microphone in cars, military vehicles, etc.

• Robust endpointing for highly noisy environments

Future Work

- Data collection
  - small pilot
  - single-speaker
    - easier/cheaper to collect
    - provide enough training for speaker-dependent models
    - expect results will generalize to speaker-independent systems
  - expect to collect 2 to 3 hours of WSJ utterances
    - first use WSJ "lsd_trn" speaker (includes at least 3 hours of speech) to train systems to determine how much data is sufficient
  - collect training data in high SNR conditions, a few test sets in different levels of noise

Future Work

- Signal Processing/Frontends
  - Combination of inputs in spectral domain
    - Reconstruction of a more robust spectral representation from components
  - Signal-adaptive front-ends for heterogeneous inputs
    - Each signal has unique time/frequency characteristics
- Testing/Analysis
  - determine WERs for each kind of microphone alone
  - determine WER reduction with different combinations of microphones
  - determine usefulness of combining features extracted from other devices with each microphone's feature vector