The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System

Thilo Köhler and Stephan Vogel

Nguyen Bach, Matthias Eck, Paisarn Charoenpornsawat, Sebastian Stüker,
ThuyLinh Nguyen, Roger Hsiao, Alex Waibel, Tanja Schultz, Alan W Black

Carnegie Mellon University, USA

IWSLT 2007 – Trento, Italy, Oct 2007

Outline

- Introduction & Challenges
- System Architecture & Design
- Automatic Speech Recognition
- Machine Translation
- Speech Synthesis
- Practical Issues
- Demo

Introduction & Challenges

TransTac program & Evaluation

Two-way speech-to-speech translation system

- Hands-free and Eyes-free
- Real time and Portable
- Indoor & Outdoor use
- Force protection, Civil affairs, Medical

Iraqi & Farsi

- Rich inflectional morphology languages
- No formal writing system in Iraqi
- 90 days for the development of Farsi system (surprised language task)

Outline

- Introduction & challenges
- **System Architecture & Design**
- Automatic Speech Recognition
- Machine Translation
- Speech Synthesis
- Practical Issues
- Demo

System Designs

Eyes-free/hands-free use

- No display or any other visual feedback, only speech is used for a feedback
- Using speech to control the system
  - *“transtac listen”* : turn translation on
  - *“transtac say translation”*: say the back-translation of the last utterance

Two user modes

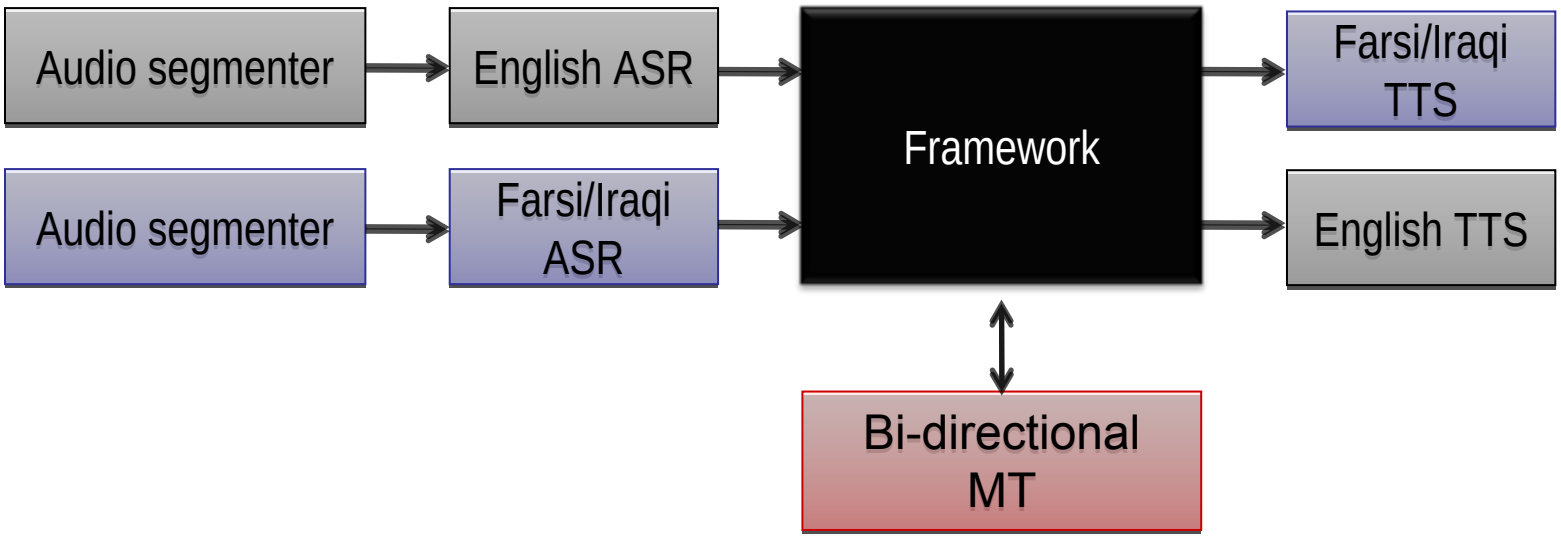
- **Automatic mode:**

  automatically detect speech, make a segment then recognize and translate it

- **Manual mode:**

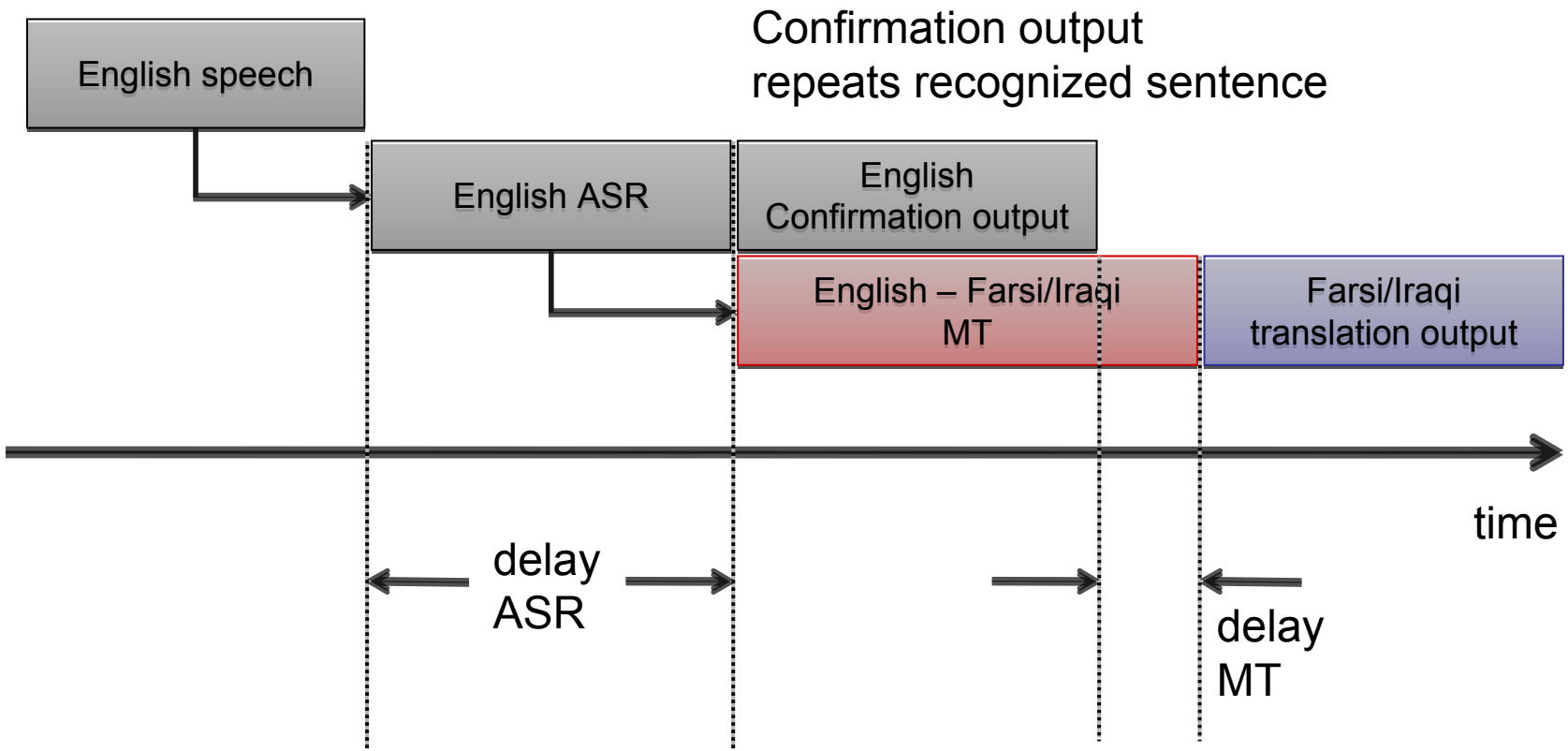
  providing a push-to-talk button for each speaker

System Architecture



Process Over Time

English to Farsi/Iraqi



CMU Speech-to-Speech System

interACT

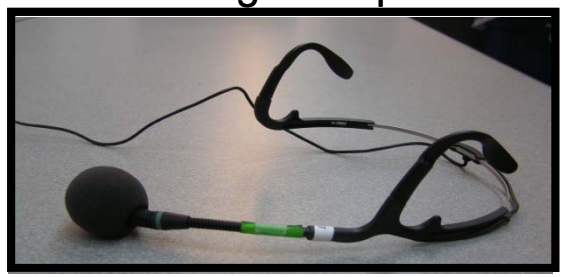
MOBILE TECHNOLOGIES LLC.

epstral

Optional speech control
Push-to-Talk Buttons

Small powerful Speakers

Close-talking Microphone



Laptop secured in Backpack

Outline

- Introduction & challenges
- System Architecture & Design
- **Automatic Speech Recognition**
- Machine Translation
- Speech Synthesis
- Practical Issues
- Demo

English ASR

3-state subphonetically tied, fully-continuous HMM

- 4000 models, max. 64 Gaussians per model, 234K Gaussians in total 13 MFCC, 15 frames stacking, LDA -> 42 dimensions

Trained on 138h of American BN data, 124h Meeting data

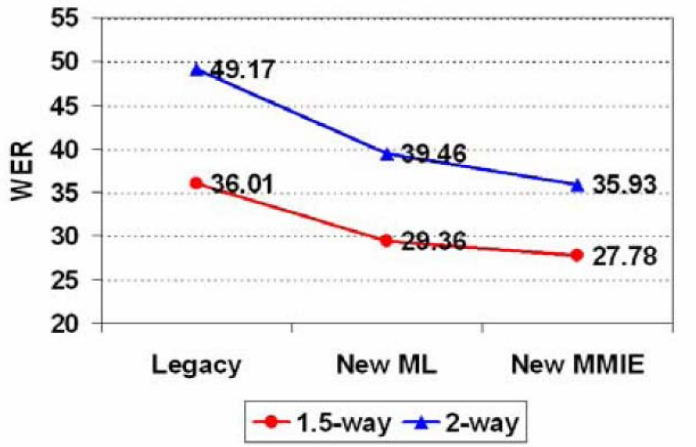
- Merge-and-split training, STC training, 2x Viterbi Training
- Map adapted on 24h of DLI data
- Utterance based CMS during training, incremental CMS and cMLLR during decoding

Iraqi ASR

ASR system uses the Janus recognition toolkit (JRTk) featuring the IBIS decoder.

Acoustic model trained with 320 hours of Iraqi Arabic speech data.

The language model is a tri-gram model trained with 2.2M words.



Iraqi ASR	2006	2007
Vocabulary	7k	62k
# AM models	2000	5000
#Gaussians/ model	≤ 32	≤ 64
Acoustic Training	ML	MMIE
Language Model	3-gram	3-gram
Data for AM	93 hours	320 hours
Data for LM	1.2 M words	2.2 M words

Farsi ASR

The Farsi acoustic model is trained with 110 hours of Farsi speech data.

The first acoustic model is bootstrapped from the Iraqi model.

- Two Farsi phones are not covered and they are initialized by phones in the same phone category.
- A context independent model was trained and used to align the data.
- Regular model training is applied based on this aligned data.

The language model is a tri-gram model trained with 900K words

Farsi ASR	2007
Vocabulary	33k
# AM models	2K quinphone
#Gaussians/ model	64max
Acoustic Training	MMIE/MAS/STC
Front-end	42 MFCC-LDA
Data for AM	110 hours
Data for LM	900K words

Farsi ASR	ML built	MMIE built
1.5-way	28.73%	25.95%
2-way	51.62%	46.43%

Outline

- Introduction & challenges
- System Architecture & Design
- Automatic Speech Recognition
- **Machine Translation**
- Speech Synthesis
- Practical Issues
- Demo

Typical Dialog Structure

English speaker gathers information from Iraqi/Farsi speaker

English speaker gives information to Iraqi Farsi speaker

English speaker:

- Questions
- Instructions
- Commands

Iraqi/Farsi:

- Yes/No - Short answers

English speaker	Farsi/Iraqi speaker
Do you have electricity?	
	No, it went out five days ago
How many people live in this house?	
	Five persons.
Are you a student at this university?	
	Yes, I study business.
Open the trunk of your car.	
You have to ask him for his license and ID.	

Training Data situation

	Source	Target
Iraqi→English		
Sentences	502,380	
Unique pairs	341,149	
Average length	5.1	7.4
Words	2,578,920	3,707,592
English→Iraqi		
Sentence pairs	168,812	
Unique pairs	145,319	
Average length	9.4	6.7
Words	1,581,281	1,133,230

	Source	Target
Farsi→English		
Sentences	56,522	
Unique pairs	50,159	
Average length	6.5	8.1
Words	367,775	455,306
English→Farsi		
Sentence pairs	75,339	
Unique pairs	47,287	
Average length	6.7	6.0
Words	504,109	454,599

Data Normalization

Minimize the mismatch in vocabulary between ASR, MT, and TTS components while maximizing the performance of the whole system.

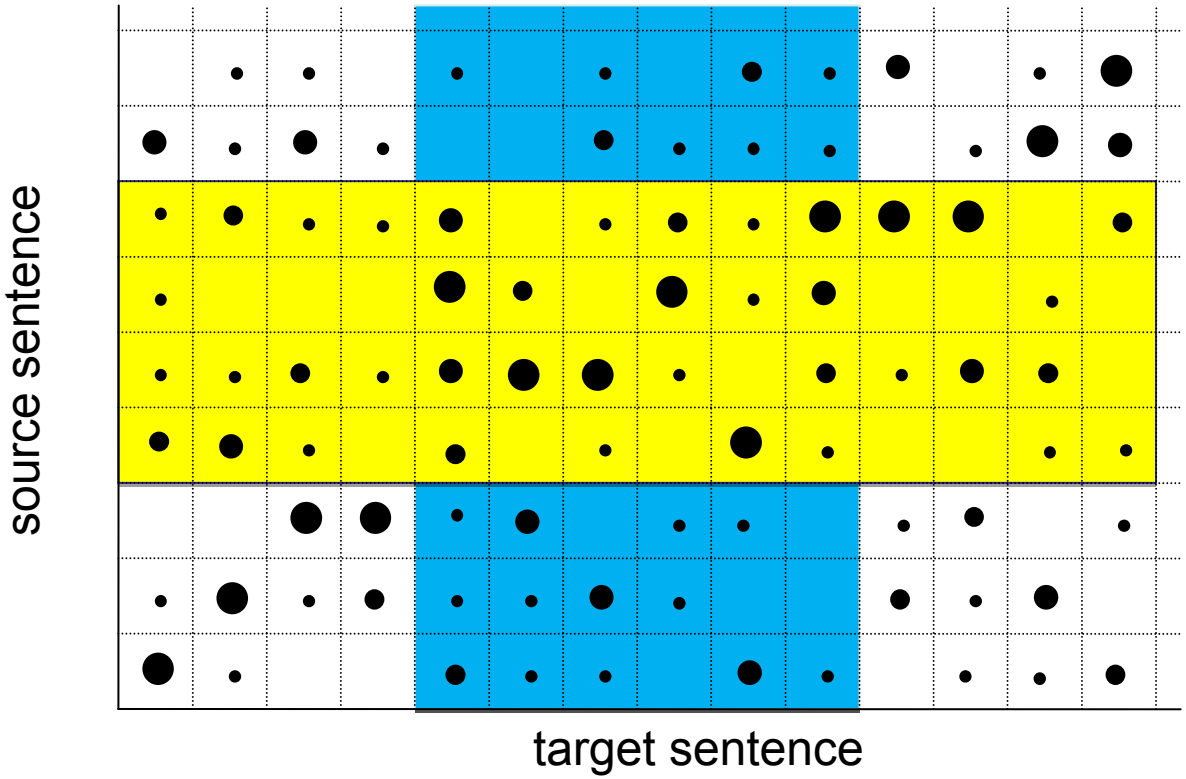
Sources of vocabulary mismatch

- Different text preprocessing in different components
- Different encoding of the same orthography form
- Lack of standard in writing system (Iraqi)
- Words can be used with their formal or informal/colloquial endings
  - *raftin* vs. *raftid* “you went”.
- Word forms (inside of the word) may be modified to represent their colloquial pronunciation
  - *khune* vs. *khAne* “house” ; *midam* vs. *midaham* “i give”

Phrase Extraction

For Iraqi – English: PESA Phrase Extraction

PESA phrase pairs based on IBM1 word alignment probabilities



PESA Phrase Extraction

Online Phrase Extraction

- Phrases are extracted as needed from the bilingual corpus

Advantage

- Long matching phrases are possible

  especially prevalent in the TransTac scenarios:

  “Open the trunk!”, “I need to see your ID!”, “What is your name?”

Disadvantage

- Slow speed: Up to 20 seconds/sentence

Speed Constraints

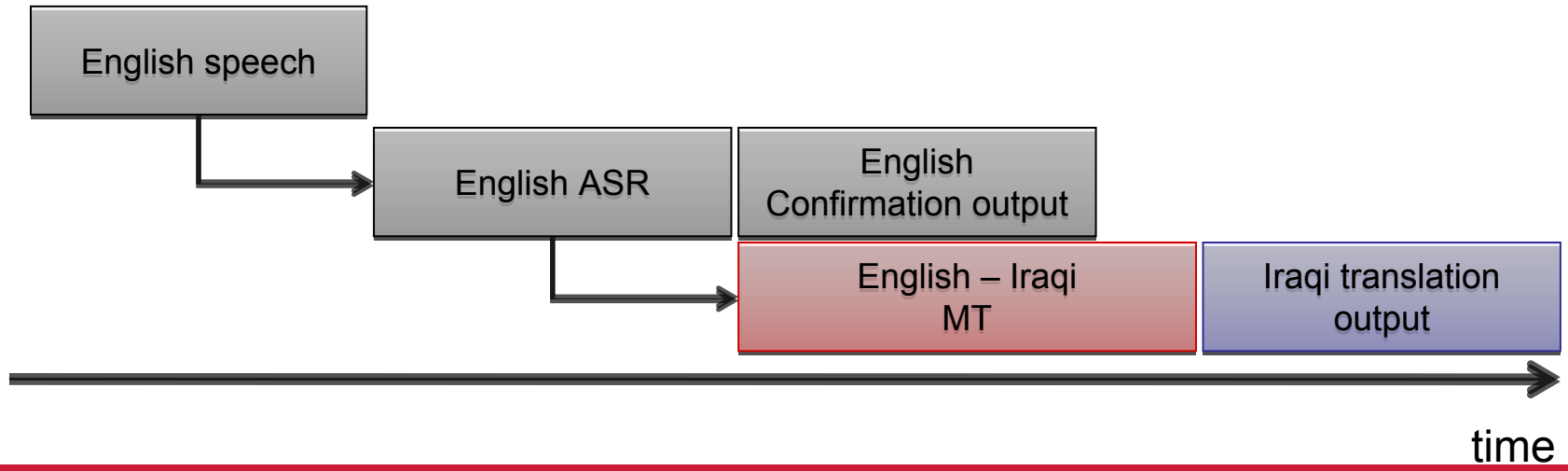
...20 seconds per sentence is too long

Solution: Combination of

- pre-extracted common phrases (→ speedup)
- Online extraction for rare phrases (→ performance increase)

Also Pruning of phrasetables in necessary

About 200 ms are available to do the translations



Pharaoh – Missing Vocabulary

Some words in the training corpus will not be translated because they occur only in longer phrases of Pharaoh phrase table.

- E2F and F2E: 50% of vocabulary not covered
- Similar phenomenon in Chinese, Japanese BTEC

PESA generates translations for all n-grams including all individual words.

Trained two phrase tables and combined them. Re-optimized parameters through a minimum-error-rate training framework.

English → Farsi	BLEU
Pharaoh + SA LM	15.42
PESA + SA LM	14.67
Pharaoh + PESA + SA LM	**16.44**

Translation Performance

Iraqi ↔ English

PESA Phrase pairs

(online + preextracted)

English → Iraqi	42.12
Iraqi → English	63.49

Farsi ↔ English

Pharaoh + PESA

(pre-extracted)

English → Farsi	16.44
Farsi → English	23.30

2 LM Options:

- 3-gram SRI language model (Kneser-Ney discounting)
- 6-gram Suffix Array language model (Good-Turing discounting)

English→Farsi	Dev Set	Test Set
Pharaoh + SRI LM	10.07	14.87
Pharaoh + SA LM	10.47	15.42

6-gram consistently gave better results

Outline

- Introduction & challenges
- System Architecture & Design
- Automatic Speech Recognition
- Machine Translation
- **Speech Synthesis**
- Practical Issues
- Demo

Text-to-speech

TTS from Cepstral, LLC's SWIFT

- Small footprint unit selection

Iraqi -- 18 month old

- ~2000 domain appropriate phonetically balanced sentences

Farsi -- constructed in 90 days

- 1817 domain appropriate phonetically balanced sentences
- record the data from a native speaker
- construct a pronunciation lexicon and build the synthetic voice itself.
- used CMUSPICE Rapid Language Adaptation toolkit to design prompts

Pronunciation

Iraqi/Farsi pronunciation from Arabic script

- Explicit lexicon: words (without vowels) to phonemes
- Shared between ASR and TTS
- OOV pronunciation by statistical model
  - CART prediction from letter context
- Iraqi: 68% word correct for OOVs
- Farsi: 77% word correct for OOVs

(Probably) Farsi script better defined than Iraqi script (not normally written)

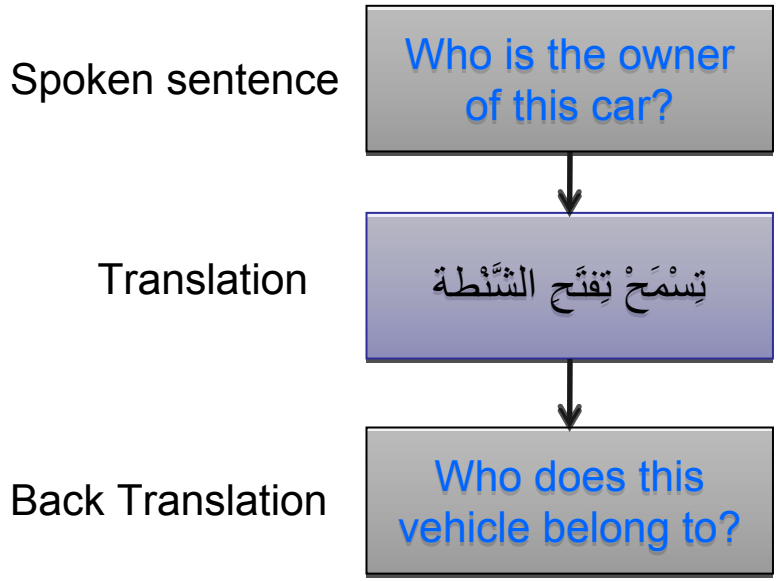
Outline

- Introduction & challenges
- System Architecture & Design
- Automatic Speech Recognition
- Machine Translation
- Speech Synthesis
- **Practical Issues**
- Demo

Back Translation

Play the “back translation” to the user

- Allows judgement of Iraqi output
- If “back translation” is still correct → translation was probably correct
- If “back translation” is incorrect → translation was potentially incorrect as well (repeat/rephrase)
- Very useful to develop the system



Back Translation

But the users…

- Confused by back translation

  “is that the same meaning?”

- Interpret it just as a repetition of their sentence

- mimic the non-grammatical output resulting from translating twice

Also:

- Underestimates system performance:

  Translation might be correct/understandable but back translation loses some information

  → User repeats but it would not have been necessary

Automatic Mode

Automatic mode translation mode was offered

- Completely hands-free translation

  System notices speech activity and translates *everything*

But the users…

- Do not like this loss of control
- Not everything should be translated, e.g. Discussions among the soldiers:

  “Do you think he is lying?”
- Definitely prefer “push-to-talk” manual mode

Other Issues

Some users: "*TTS is too fast to understand*"

- Speech synthesizers are designed to speak fluent speech, but output of an MT system may not be fully grammatical
- Phrase breaks in the speech could help listener to understand it

How to *use language expertise efficiently and effectively* when working on rapid development of speech translation components

- We had no Iraqi speaker and only 1 Farsi part timer
- How do you best use the limited time of the Farsi speaker?

  Check data, translate new data,

  fix errors, explain errors,

  use the system....?

Other Issues

User interface

- Needs to be as simple as possible
- Only short time to train English speaker
- No training of the Iraqi/Farsi speaker

Over-heating

- Outside temperatures during Evaluation reached 95 Fahrenheit (35° Centigrade)
- System cooling is necessary via added fans

DEMO: CMU Speech-to-Speech System

interACT

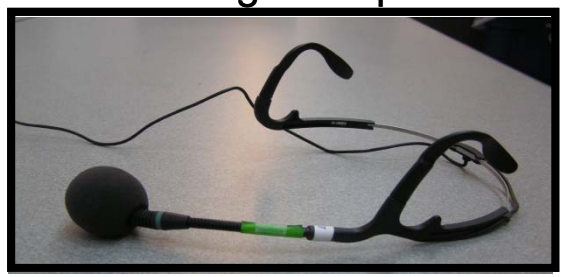
MOBILE TECHNOLOGIES LLC.

epstral

Optional speech control
Push-to-Talk Buttons

Small powerful Speakers

Close-talking Microphone



Laptop secured in Backpack

DEMO