Information Fusion Approaches for Distant Speech Recognition in a Multi-microphone Setting

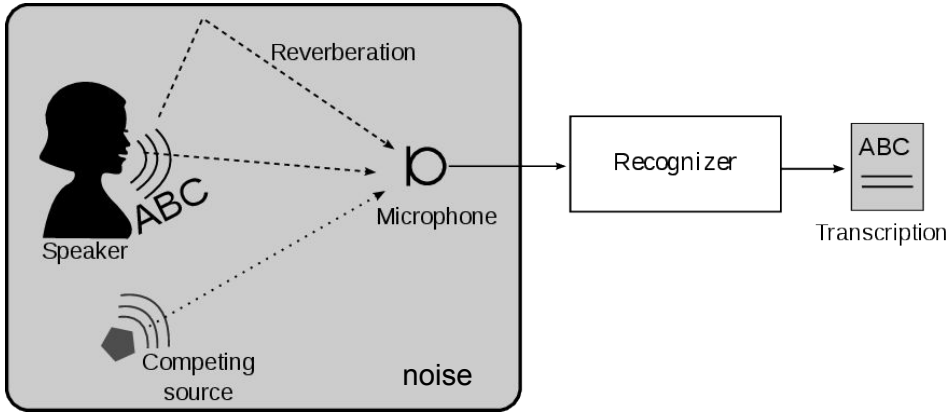
Cristina Guerrero Flores
Ph.D. Defense

Advisor: Maurizio Omologo

FONDAZIONE BRUNO KESSLER

What is **DSR**?

Why is **DSR hard**?



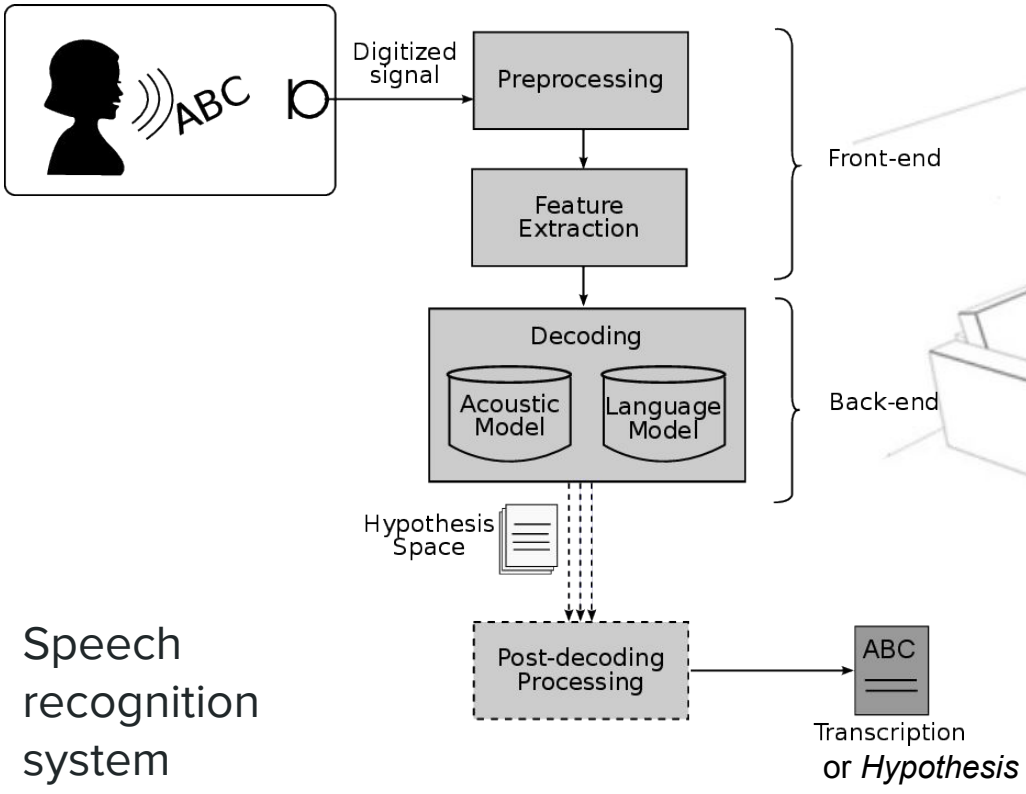
- On top of ASR issues ...

- Distance emphasizes **acoustic phenomena/ distortions**: noise, simultaneous sources, **reverberation**

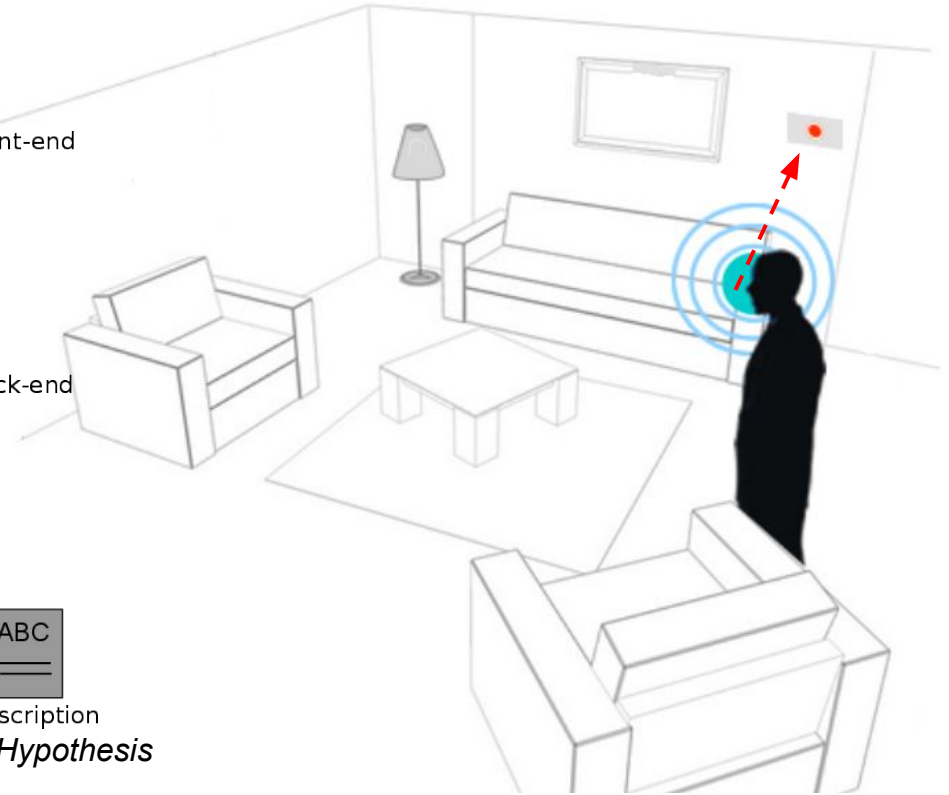
- **Modeling the acoustic variabilities** for speech recognition is **almost impossible** in practice

How is **DSR** addressed?

Single-mic DSR



Speech recognition system



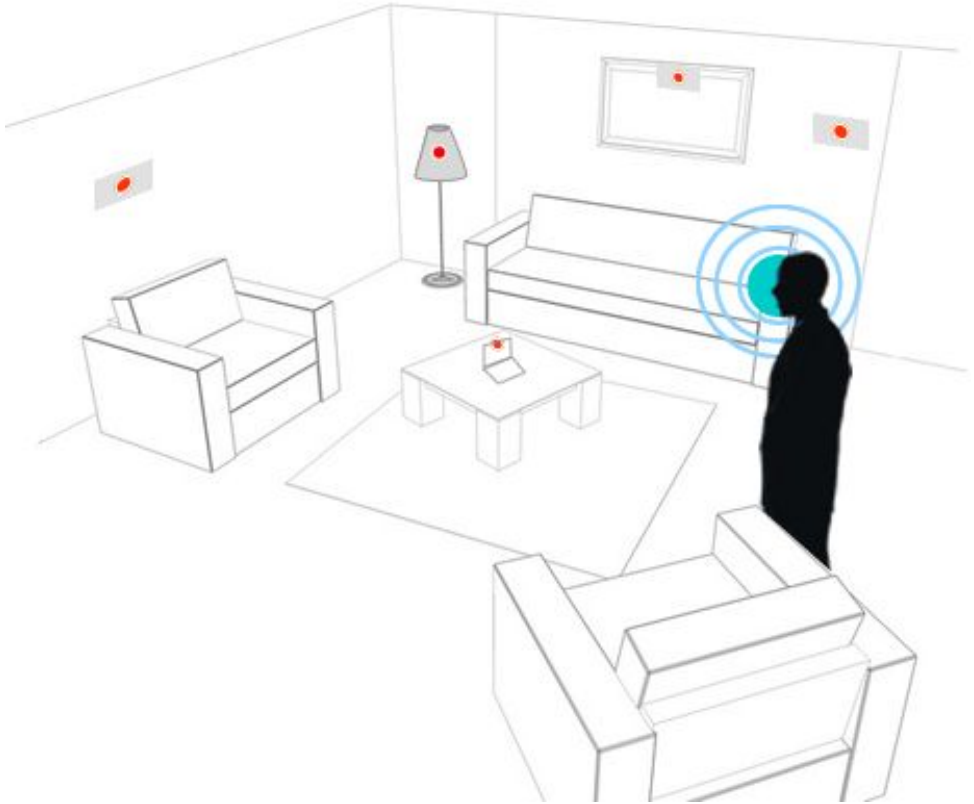
Multi-mic DSR

Challenges:

- **Decision**
- Resource representation
- Complexity

Benefits:

- Multiple **perspectives**
- Better **performance**



Multi-mic Processing

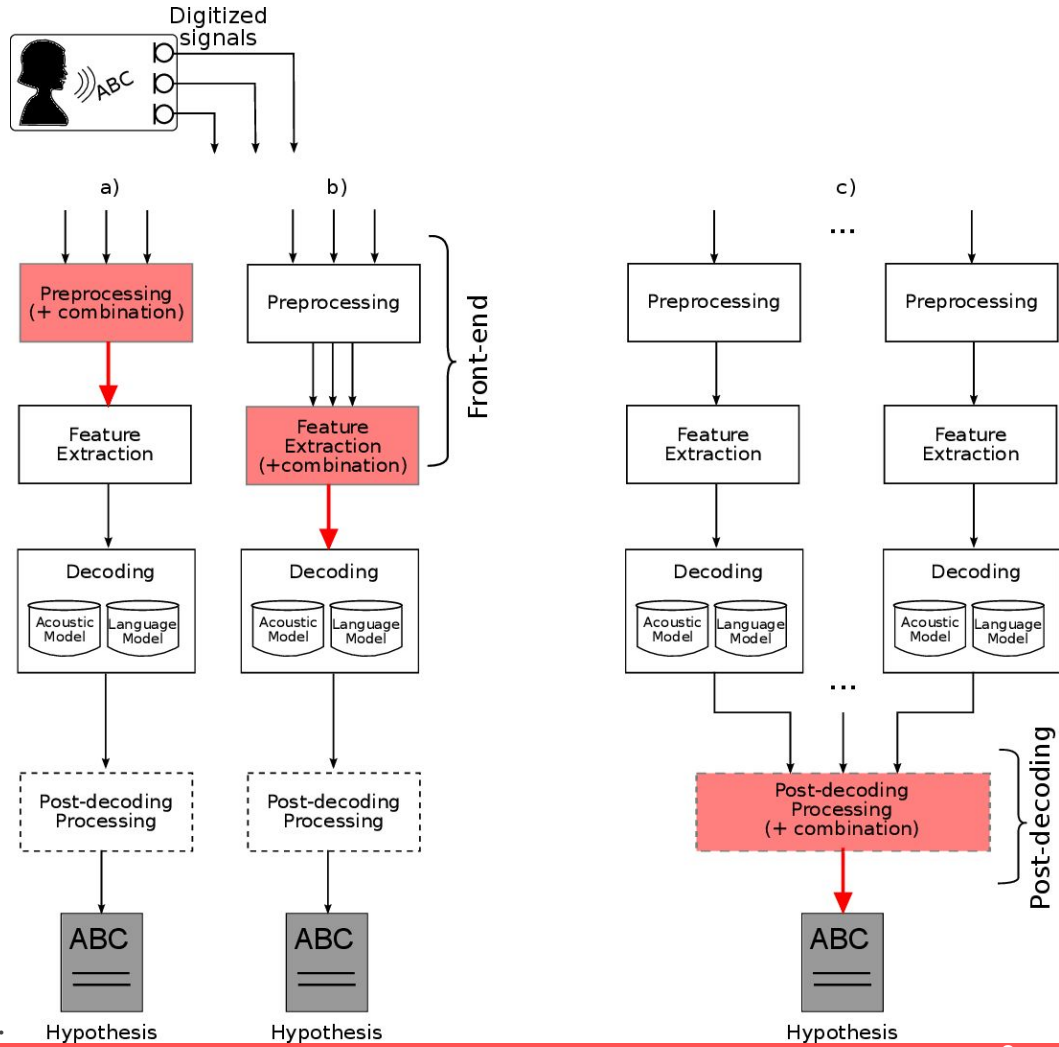
Front-end

Beamforming [Flanagan, J., et al., 1985],
Feature combination [Ma et al., 2010]
Channel selection [Wolf and Nadeu, 2014]
Enhancement [Benesty, J., et al, J. 2005],
Degradation [Droppo, J. and Acero, A. 2008],
Source separation [Makino et al., 2007].

Post-decoding

-Word level combination [Fiscus, 1997],
Hypothesis space combination [Stolcke, 2011]
Hypothesis selection [Stolcke et al., 1997, Obuchi, 2006].

Evaluation campaigns: CHiME, REVERB, AsPIRE.



Objective

To investigate and propose solutions for distant speech recognition in enclosures equipped with multiple largely-spaced mics.

- **Real scenario**: smart home + mic network European project DIRHA *

- **Information fusion** approaches at: **Front-end level** and **Post-decoding** level

*See: http://dirha.fbk.eu

Contributions

- Proposed an **objective-score based channel selection** framework.
- Introduced a novel **methodology for** channel selection **assessment**.
- Proposed a **method for combining hypothesis spaces** captured in a **multi-microphone scheme**.
- Implemented the proposed hypothesis combination method as an **extension of SRILM toolkit***.

Scientific production: CS work: Guerrero C., Tryfou G., Omologo M. INTERSPEECH 16, Guerrero C., Tryfou G., Omologo, M. -under submission at Computer, Speech and Language Journal. Hypothesis combination work: Guerrero C., Omologo M., HSCMA 14, Guerrero C., Omologo M., EUSIPCO 14, *Extension at: https://github.com/cristinagf/mmcn.

General outline:

Part I) Channel Selection

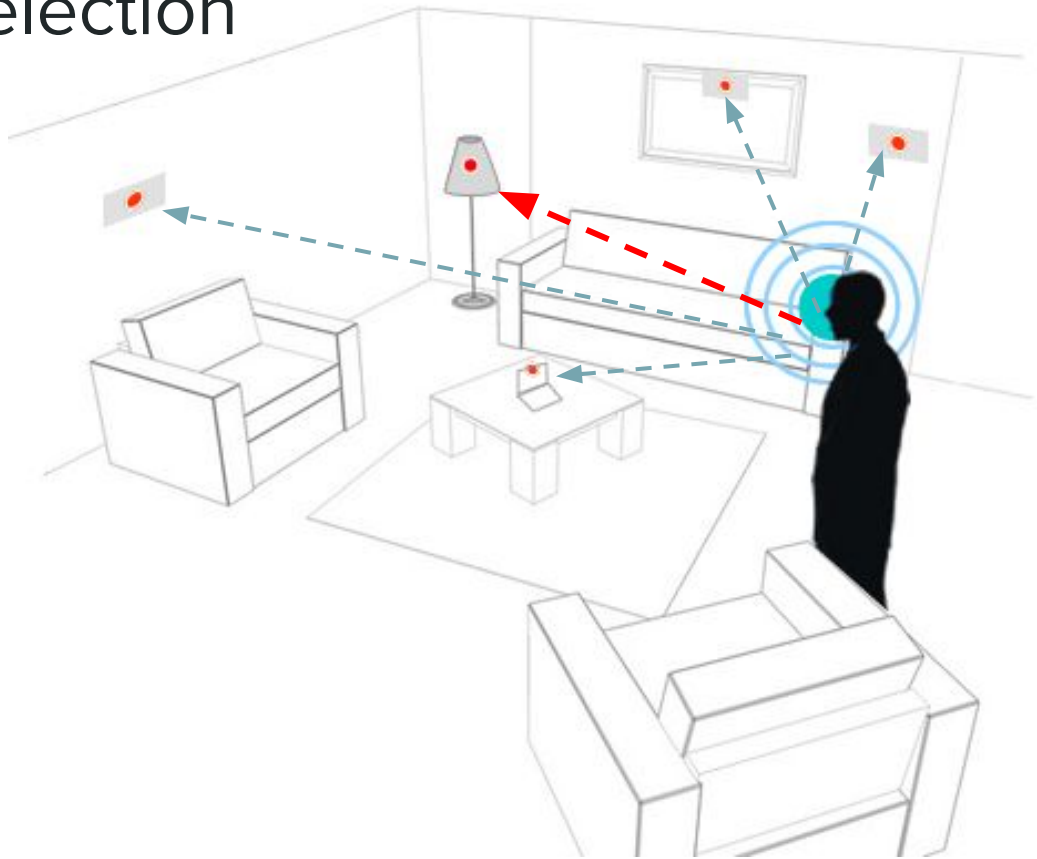
Part II) Hypothesis combination

Conclusions and future directions

Outline: Part I "*Channel Selection*"

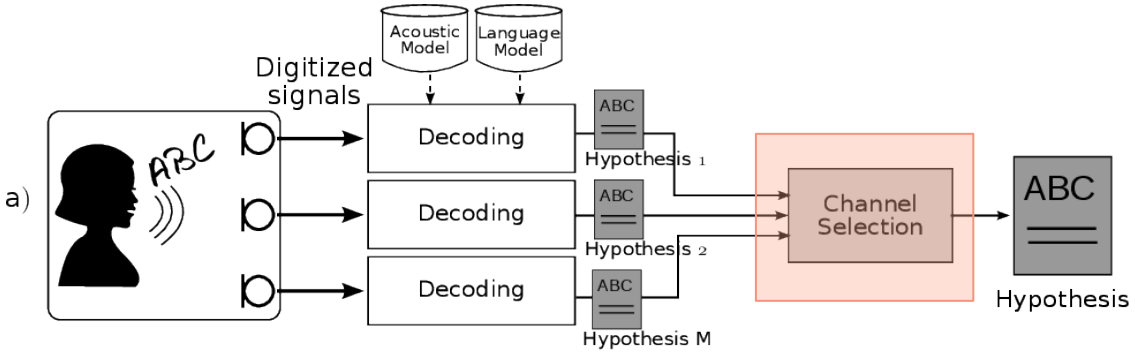
1. Channel Selection (CS)
2. Contribution: CS based on Cepstral Distance
3. Experiments
    - Effect of speaker location/mic-network on CS
    - CS in realistic scenarios
4. Results

Channel Selection

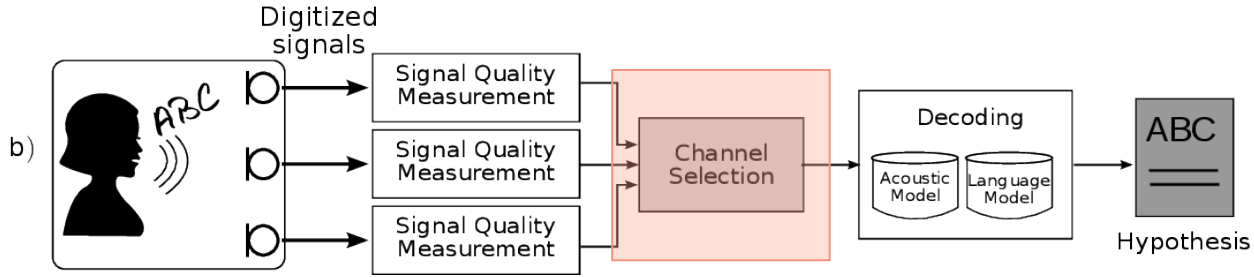


Channel Selection (CS)

- Maximization problem. **How to score the channels?**



Decoder based CS



Signal based CS

- [Wolf and Nadeu, 2014]

CS: Signal-based methods

[Guerrero C., Tryfou G., Omologo M. INTERSPEECH 16]

- **Informed:** reverberant signals + **target signal**
  Search the closest possible to ideal (clean speech).
  Not a real applicable solution, but as a tool for study.

- **Blind:** only use reverberant signals
  - e.g., Envelope variance [Wolf, 2013]

$$\hat{C} = \arg\max_{m} \sum_{k} \frac{V_m(k)}{\max_{m}(V_m(k))}$$

m: channel
k: frequency sub-band
$V_m(k)$: $sub-band\ variance$

CS Contribution:

Key: How good is a signal?

Objective measures to estimate signal quality

- Speech coding, speech enhancement, other speech applications (speech recognition, voice activity detection)

- Cepstral distance (**CD**)

  Inverse Fourier transform of the log of the spectrum

$$d(\vec{c}_x, \vec{c}_m)$$

c_x: cepstral coef. of the clean signal
c_m: cepstral coef. of a signal captured by the mic m

CS Contribution: CS based on CD

Informed:

$$\hat{M}_X = \underset{m}{\operatorname{argmin}} \quad d(\vec{c}_x, \vec{c}_m)$$

distance
(between clean / signal of mic m)

Blind:

Reference? Create a distortion reference[1].

Search the furthest from average distortion.

$$\hat{R}(t,\omega) = \frac{1}{M} \sum_m \log |X_m(t,\omega)|$$

$$\hat{M}_{\hat{R}} = \underset{m}{\arg\max}\, d(\vec{c}_{\hat{R}}, \vec{c}_m)$$

distance
(between reference/signal of mic m)

1: Estimated as the geometric mean spectrum of the acquired signals
[Guerrero C., Tryfou G., Omologo M., INTERSPEECH 16]

CS: Experiments

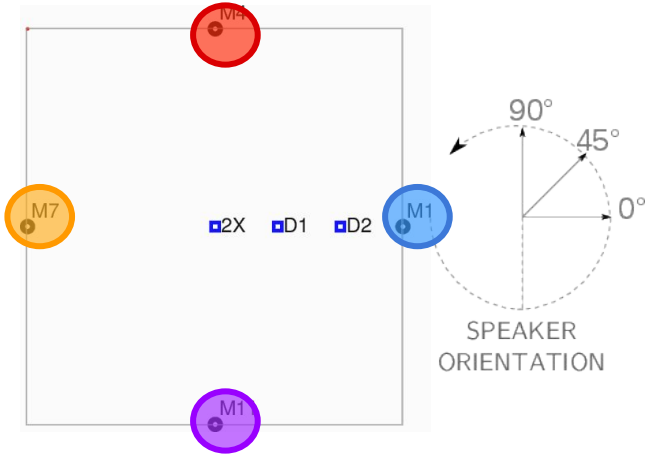
Understand:

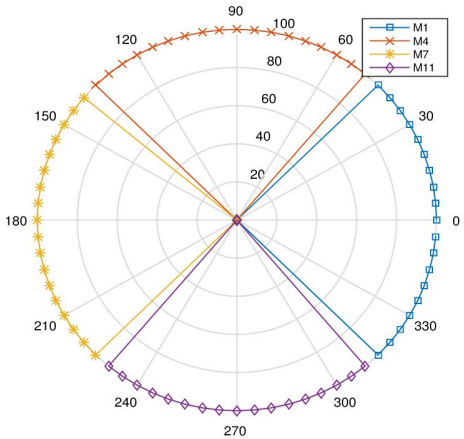
1) Effect of the speaker located at different positions/orientations, and the effect of the microphone network configuration on CS
2) CS in a realistic DSR

CS: Experiments

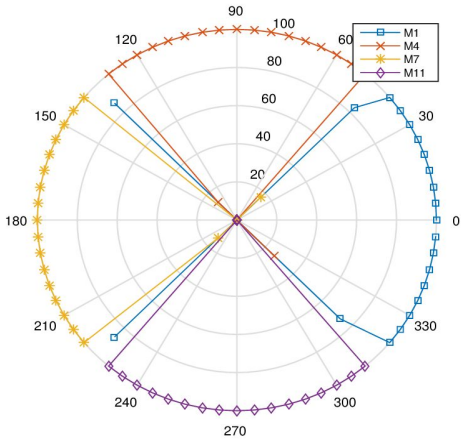
1.a) Effect of the speaker located at different positions

Speaker at 2X





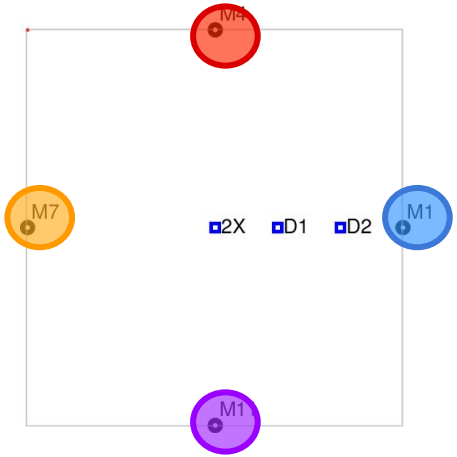
Channel selection using:
CD Informed

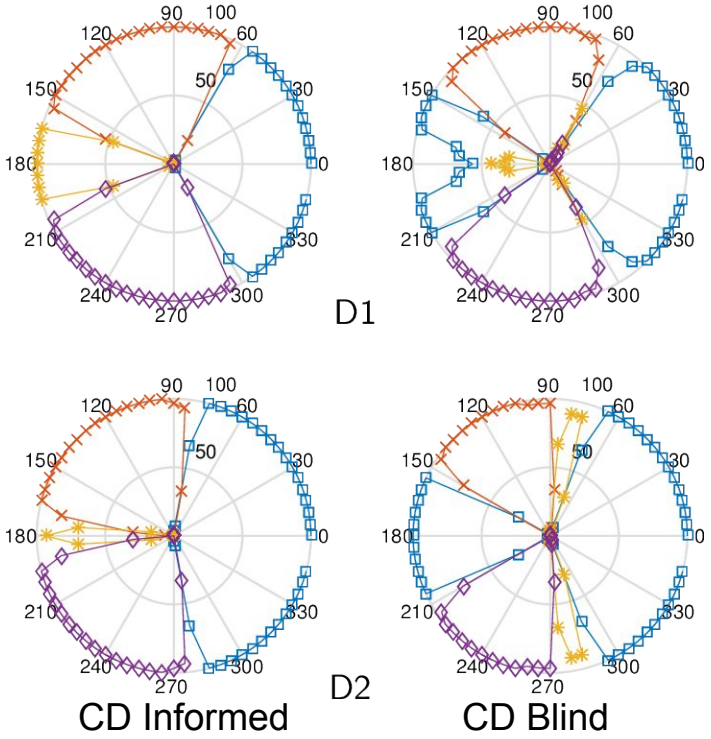


CD Blind

CS: Experiments

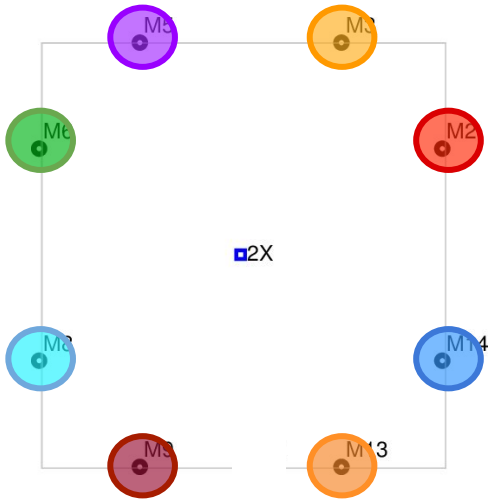
1.a) Effect of the speaker located at D1, D2

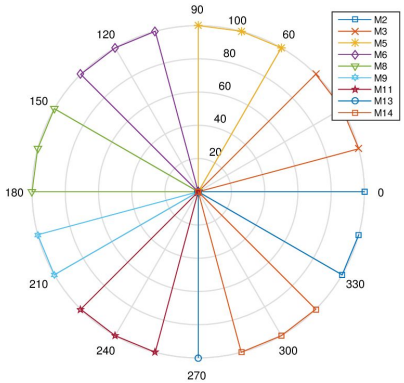




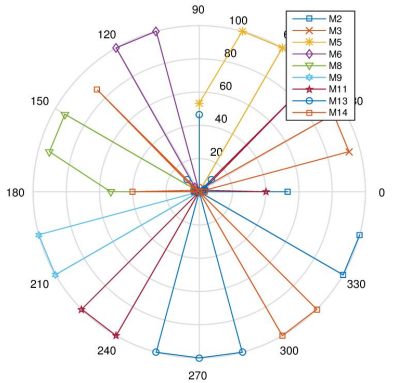
CS: Experiments

1.b) Effect of unbalanced microphone network





CD Informed



CD Blind

CS: Experiments

2) CS in a realistic scenario. Benefits for DSR.

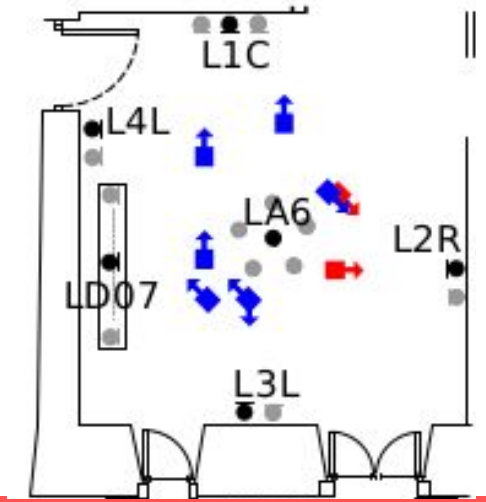
CS methods: CD informed, CD blind, EV

CS recognition performance: word error rate (WER)

Simulations (Sim), Real.

4 Datasets (by position/orientation):

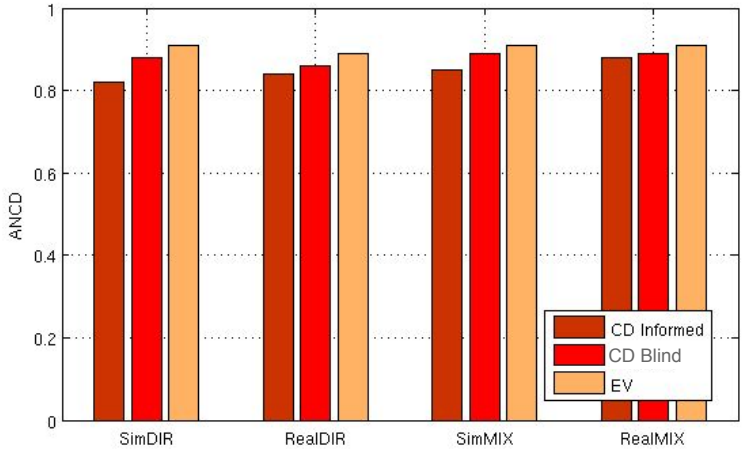
- Direct: SimDIR RealDIR (see figure)
- Mixed: SimMix RealMix



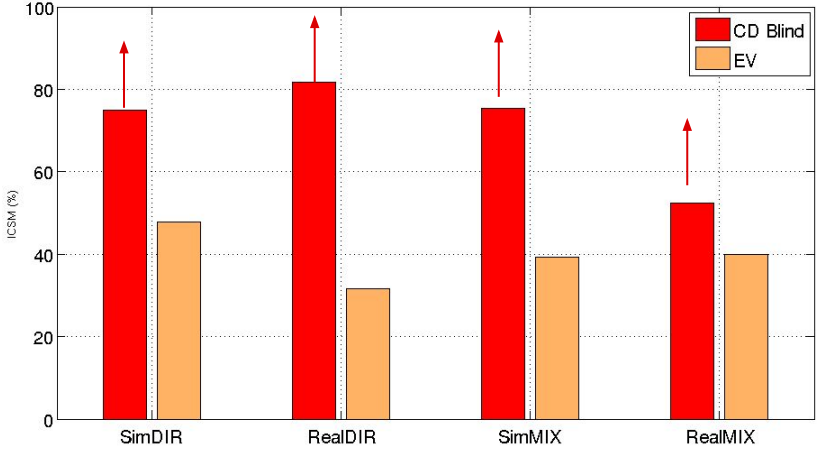
CS: Experiments

- Introduced metrics:

Average Normalized CD (**ANCD**)



Informed CS Matching rate (**ICSM**)



Agreement to an informed method!

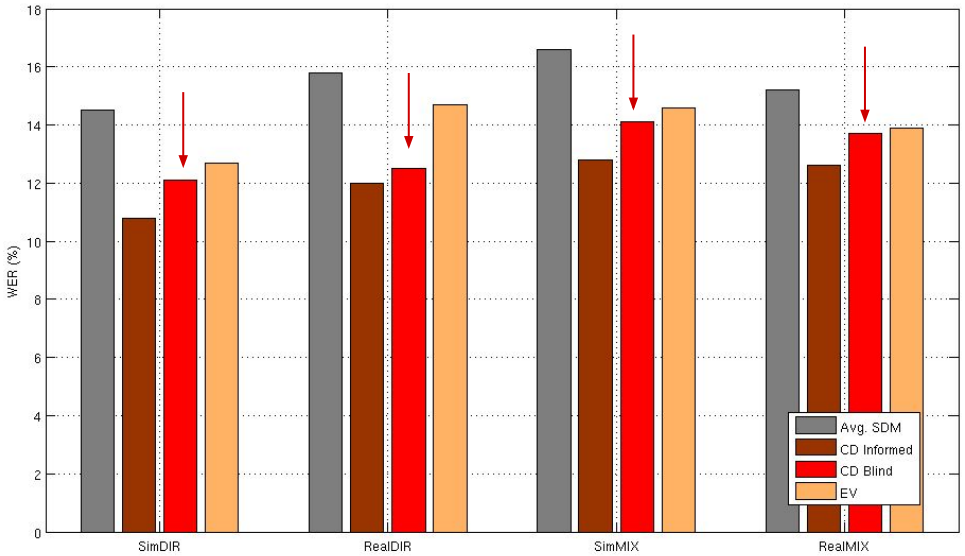
[Guerrero C., Tryfou G., Omologo M., INTERSPEECH 16]

CS: Experiments

WER [%] of the distributed microphones.

SDM	SimDIR	RealDIR	SimMIX	RealMIX
L1C	16.6	14.4	16.0	14.8
L2R	10.8	19.2	15.8	16.2
L3L	13.6	15.8	16.5	15.2
L4L	15.0	16.3	17.0	15.1
LA6	16.5	15.1	17.7	14.9
LD07	14.8	14.2	16.4	14.7
Avg	14.5	15.8	16.6	15.2

sdm: single distant mic



WER of CS methods. (Lower WER = better performance)

CS: Results

- CS validity for multi-microphone DSR
- Objective measures for CS
- CD-based CS as a relevant CS tool
- Potential improvement sources are identified for the proposed method.
- Benefit other fusion approaches (e.g., hypothesis combination).

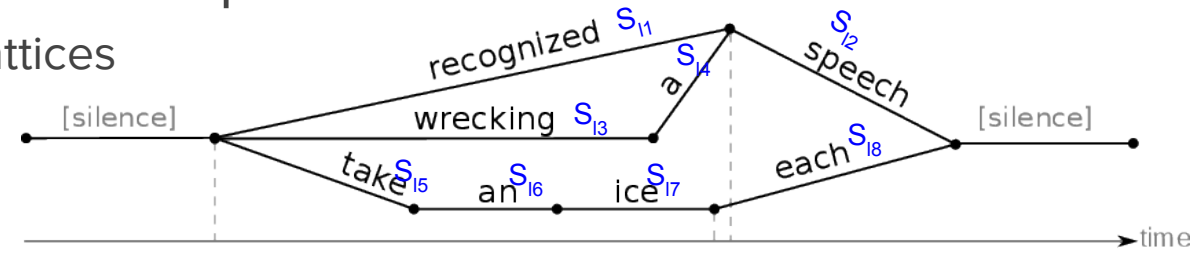
- Specific outcomes: CS framework, novel metrics, Publications: Guerrero C., Tryfou G., Omologo M. INTERSPEECH 16, under submission at the Computer, Speech and Language Journal.

Outline: Part II *“Hypothesis Combination”*

1. Basic notions on ASR
2. Hypothesis combination
3. Contribution: Multi-mic confusion network
4. Experiments
    - Effects of speaker, microphones
    - Hyp. comb. and other fusion methods
5. Results

ASR: Hypothesis space

- Hypothesis: resulting transcription
- Hypothesis space
  - Lattices



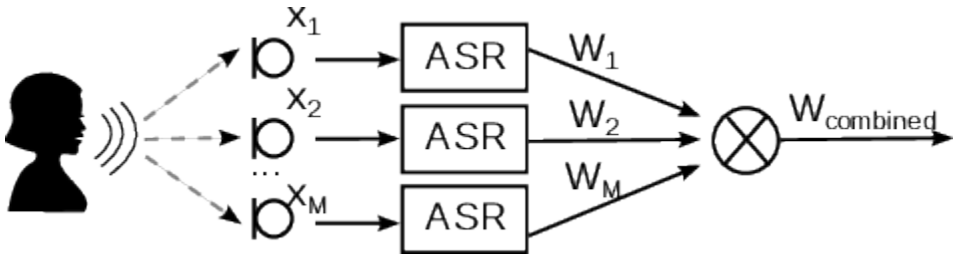
Hypotheses:

recognized speech	0.9
wrecking a speech	0.7
take an ice each	0.1

  - Confidence measures

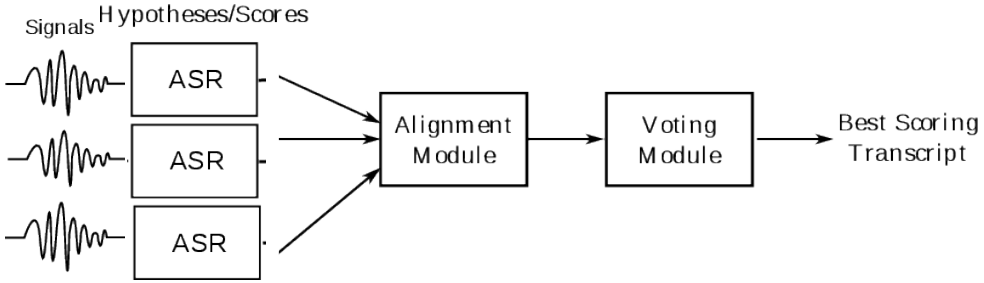
Hypothesis Combination (HypComb)

- Word-level processing
- Decoding of each channel is required

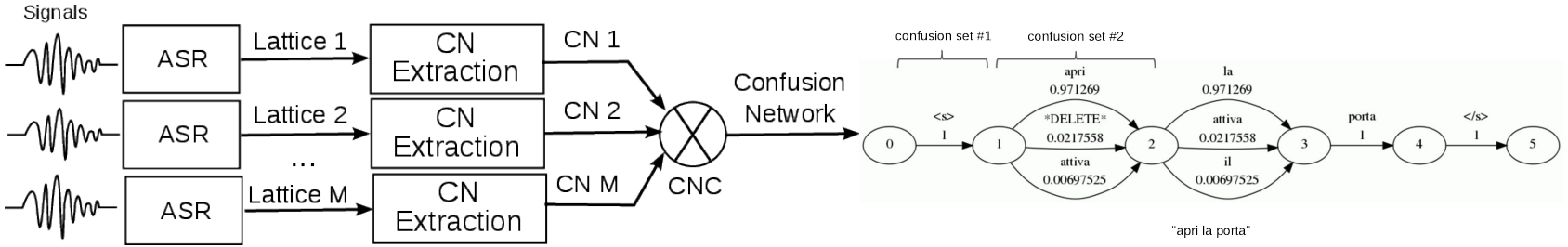


HypComb: Methods

ROVER [Fiscus, 1997]



CNC [Mangu, 2000, Evermann and Woodland, 2000]

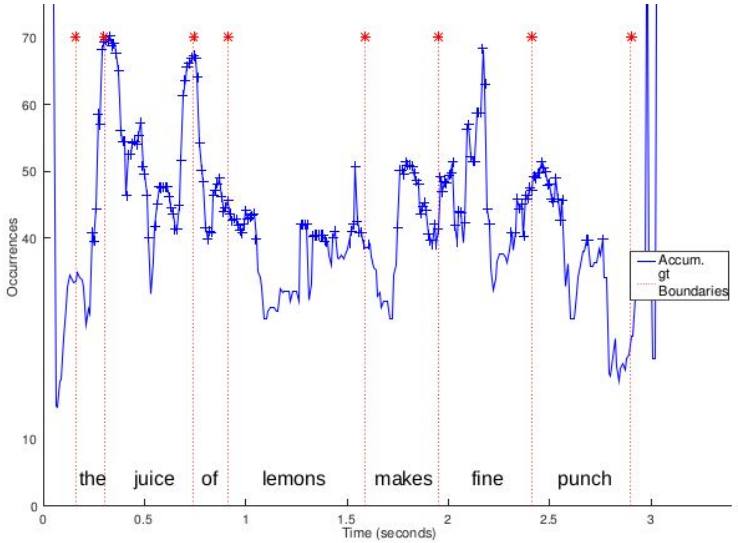


HypComb:

Analyzing multi-mic data:

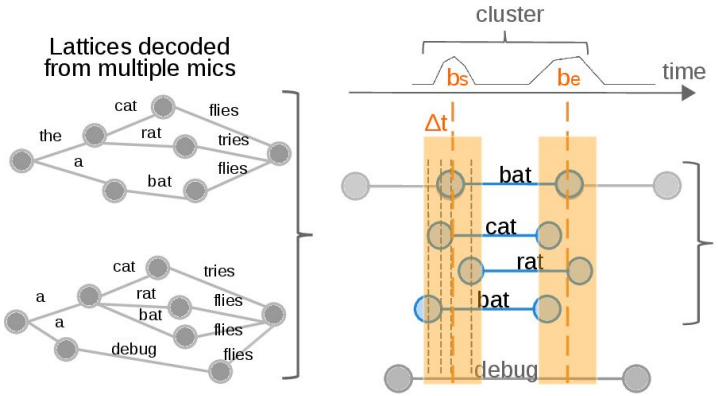
- Temporal agreement (word boundaries)

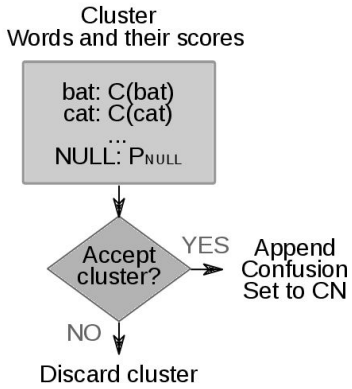
- Examine temporal segments
  - Link posterior probabilities
  - Words



HypComb Contribution: MMCN

- Word boundaries [Guerrero C., Omologo M.,, HSCMA 2014]
- Inter/intra mic scoring
- Extract a CN





Intra-mic Score

$$C\left([W_{lij}, B_i, B_{i+1}]\right) = \sum_{\substack{[w;\tau,t]: \\ [B_i-\Delta\leq\tau\leq B_i+\Delta], \\ [B_{i+1}-\Delta\leq t\leq B_{i+1}+\Delta]}} P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right)$$

Inter-mic Score

$$C'([W_{li}, B_i, B_{i+1}]) = \frac{1}{M}\sum_j C\left([W_{lij}, B_i, B_{i+1}]\right)$$

HypComb: Experiments

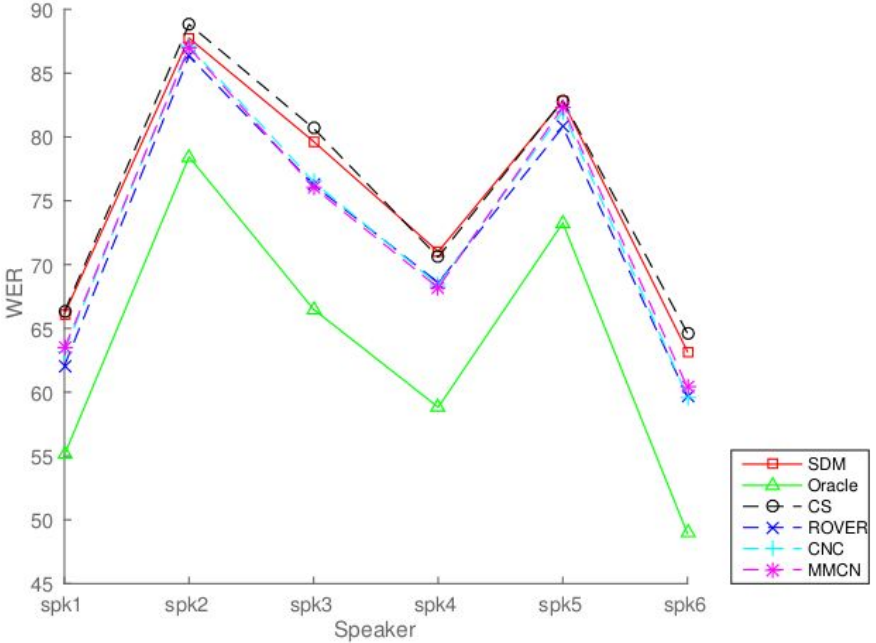
- Setting: DIRHA living room
- Study the effects of the language/acoustic models- on HypComb. Platforms/toolkits (HTK, Kaldi).
- Effect of microphone network composition (Are more mics helpful?)
- Observe variations per speaker.
- Performance of MMCN

HypComb: Experiments

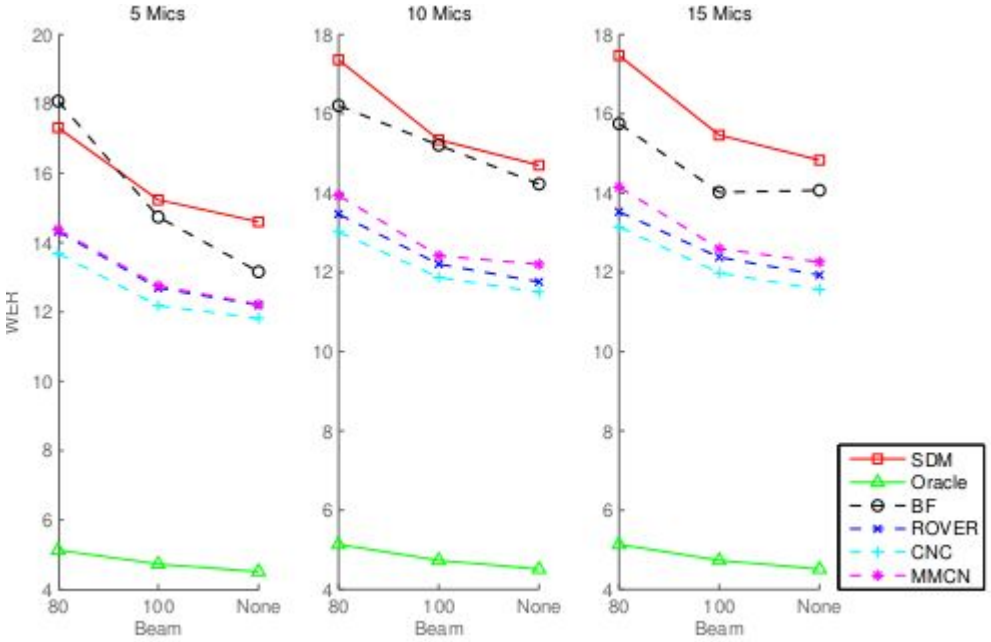
- Task: recognize continuous spoken utterances
- Tested on: Simulated and Real data (DIRHA)
  Acoustic models:trained on a contaminated dataset
- Full set of mics(15): mic-group combinations
- Methods oracle/ EV/ Beamforming / ROVER/ CNC/ MMCN

HypComb: Experiments

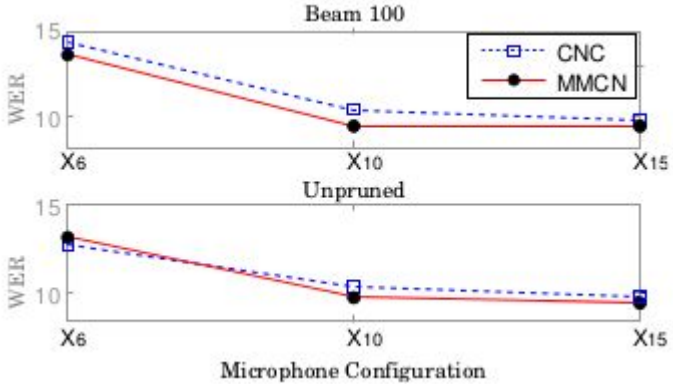
- Data: Phonetically rich sentences in English
- 3 female/ 3 male
- 0-gram grammar



HypComb: Experiments



Data: WSJ0-5k sub-set of the DIRHA-English



HypComb: Experiments

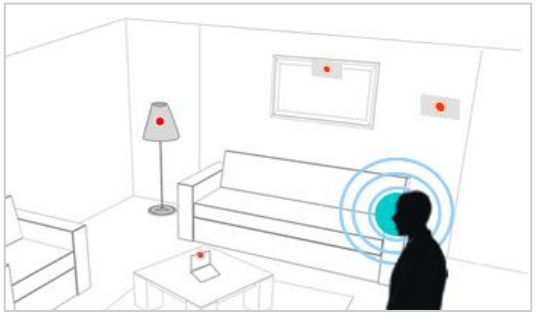
Performance of MMCN (Boundary identification):

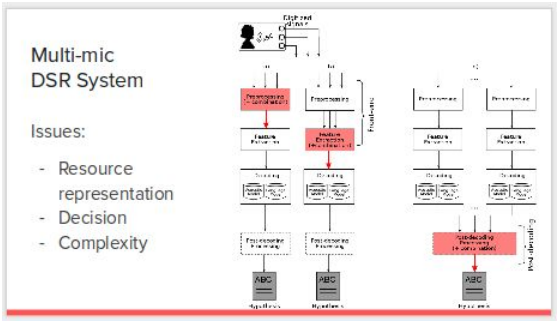
- Addition/ shifting/ loss of boundaries
- Addition & shifting: not critical effect on WER
  - Compensated by the Segment Validation stage
- Loss of boundary are more detrimental
  - Boundary identification in MMCN cautious policy
  - More likely to add extra boundaries
- MMCN improvements: boundary identification

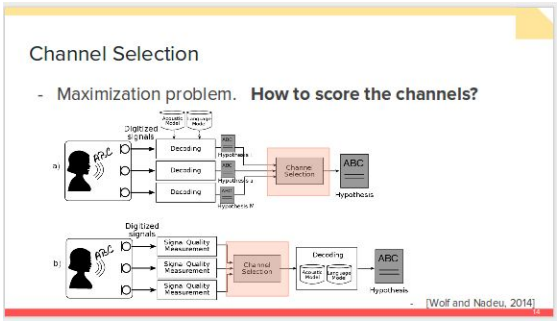
HypComb: Results

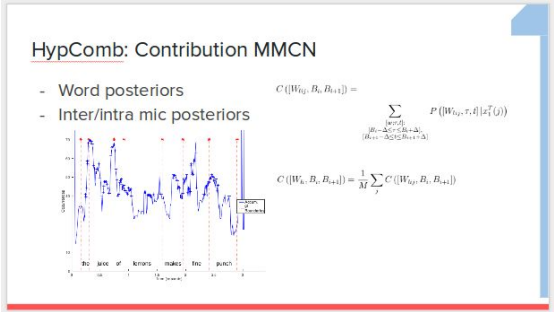
- Comparison: **MMCN vs state-of-the-art methods.**
- **Validity of Hypothesis Combination** for DSR.
- **Simple approach** to extract information for HypComb.
- MMCN, not dependent on **order of combination**.
- Implemented as an **extension of the standard SRILM** toolkit.
- Specific outcomes: MMCN - hypothesis combination method, extension of the SRILM toolkit at: https://github.com/cristinagf/mmcn, Guerrero C., Omologo M., HSCMA 14, Guerrero C., Omologo M., EUSIPCO 14, Guerrero C., Tryfou G.

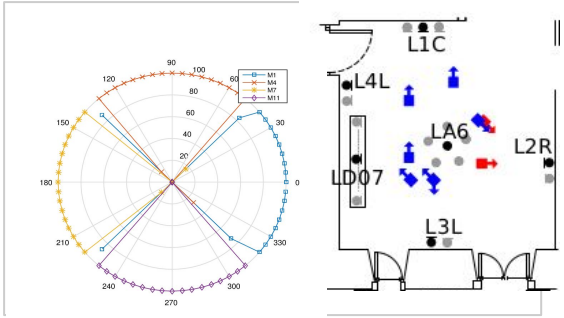
Summary

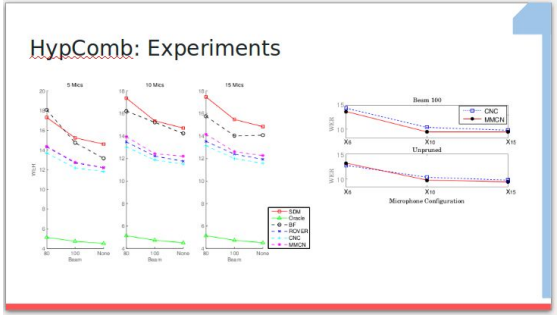



Multi-mic
DSR System
Issues:
- Resource representation
- Decision
- Complexity


Channel Selection
- Maximization problem. How to score the channels?
Digitized signals
Decoding
Decoding
Decoding
Channel Selection
ABC
hypothesis
Digitized signals
Signal Quality Measurement
Signal Quality Measurement
Signal Quality Measurement
Channel Selection
Decoding
ABC
hypothesis
[Wolf and Nadeu, 2014]


HypComb: Contribution MMCN
- Word posteriors
- Inter/intra mic posteriors


M1
M4
M7
M11
L1C
L4L
LA6
L2R
LD07
L3L


HypComb: Experiments
5 Mics
10 Mics
15 Mics
Beam 100
Unpruned
CNC
MMCN
SDM
Delay
BF
NOVAN
Microphone Configuration

Conclusions

- Framework and metrics exploiting CD for CS
- CS assessment methodology
- Post-decoding information fusion approach
- Verification and validation on synthetic and real material.
- Comparison of state-of-the-art information fusion approaches.

Future Work

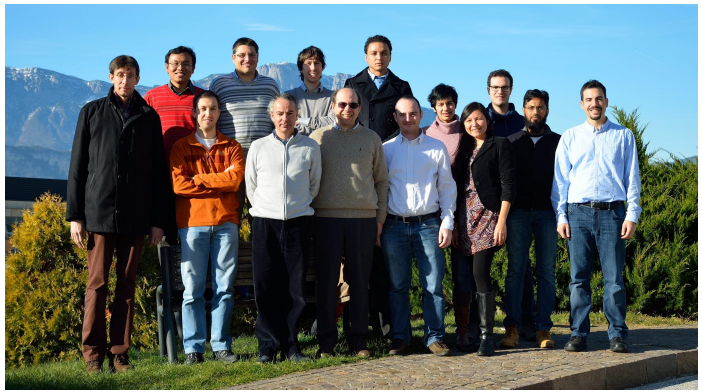
- Extend CS framework presented for CD-based to other objective signal quality measures.
- Incorporate other acoustic characteristics and conditions to the CS assessment.
- Design of integrative approaches. Combination of front-end post-decoding approaches for DSR
- For example: explore the channel scoring approaches for weighting hypothesis combination.

Acknowledgement

Prof. A. Abad, B. Demir, S. Squartini, H. Van Hamme

SHINE
SPEECH-ACOUSTIC SCENE ANALYSIS AND INTERPRETATION

Maurizio Omologo



Thank you