Luca Brayda

Robust Speech Recognition with Microphone Arrays

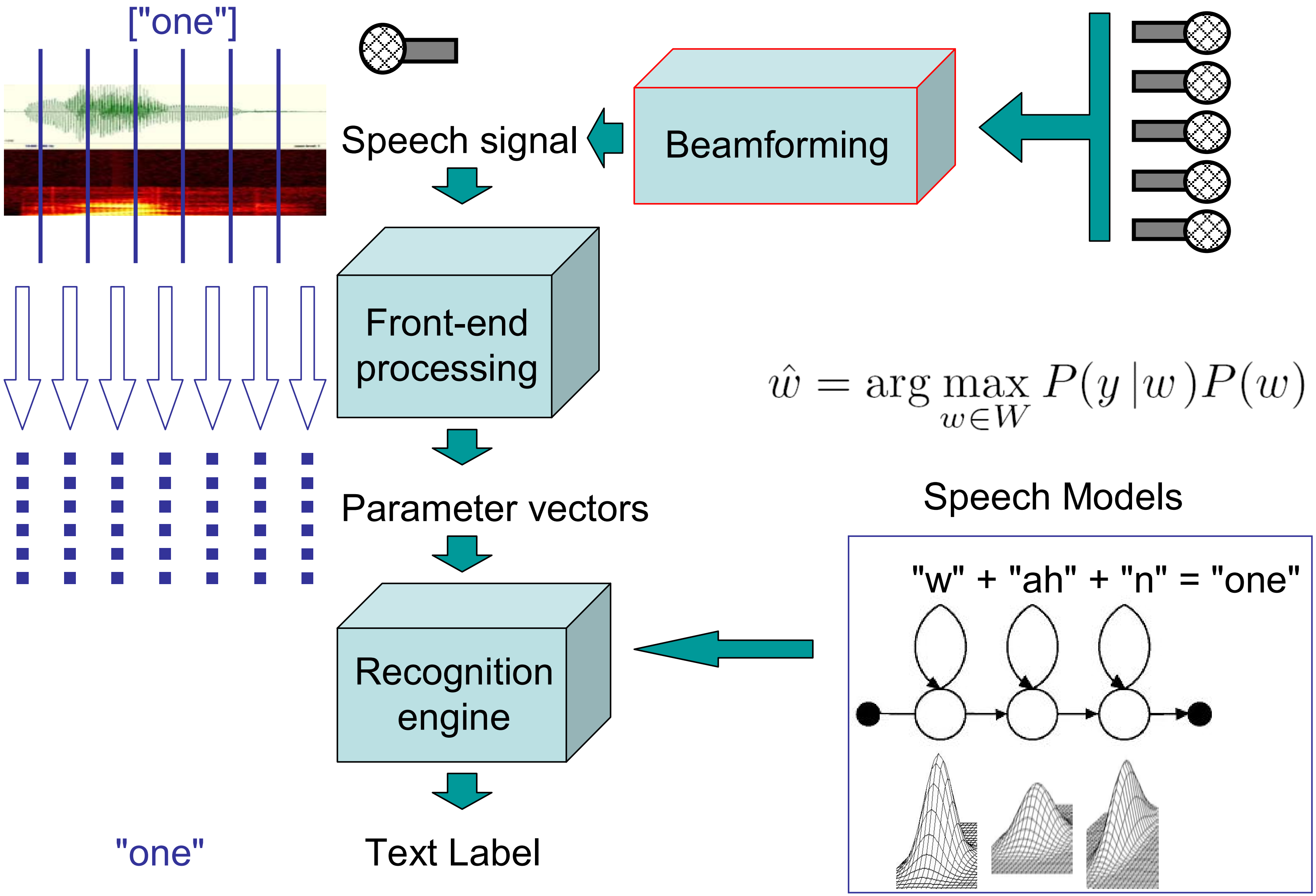
PhD advisor: Christian Wellekens

EURECOM
Sophia Antipolis

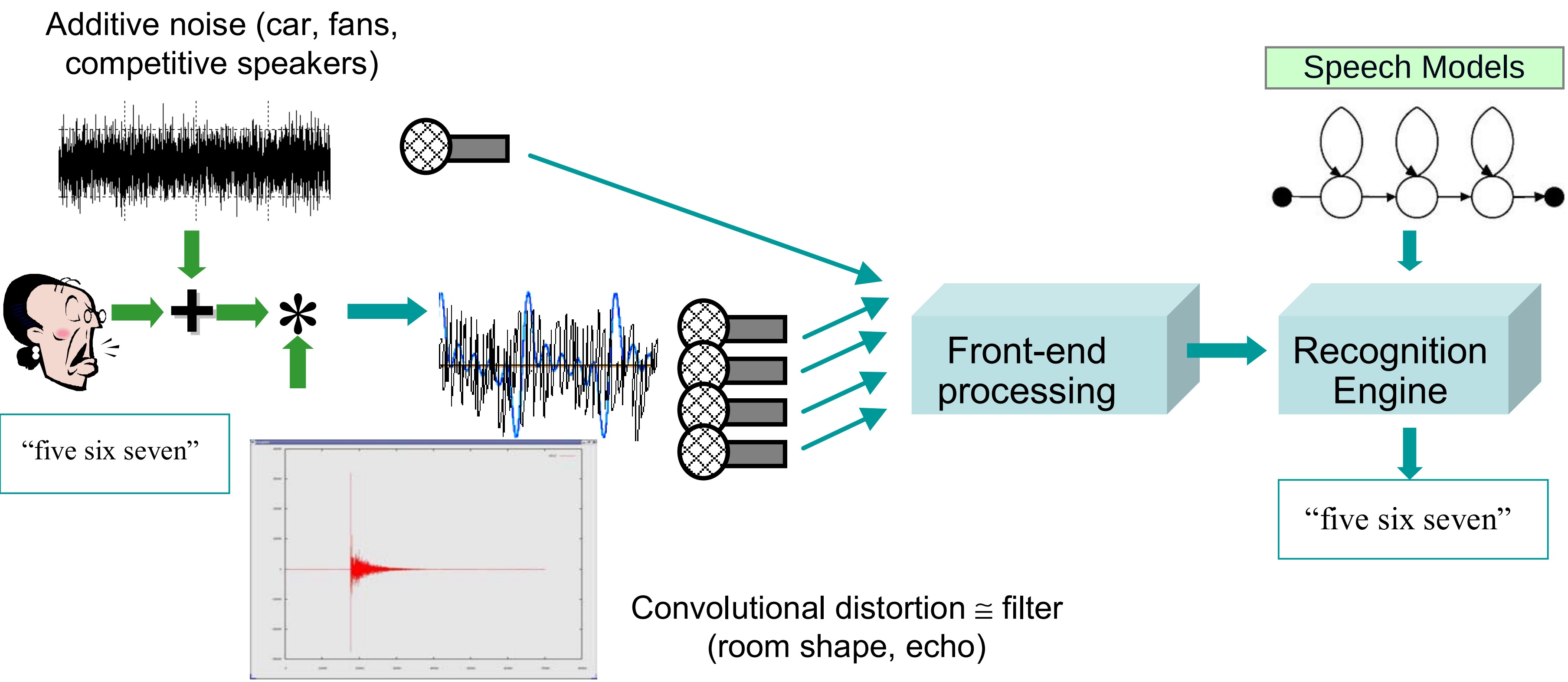
Outline

- Overview of ASR
- Likelihood-based beamforming
- N-best approach
- Ongoing work and Applications

Automatic Speech Recognition (ASR)



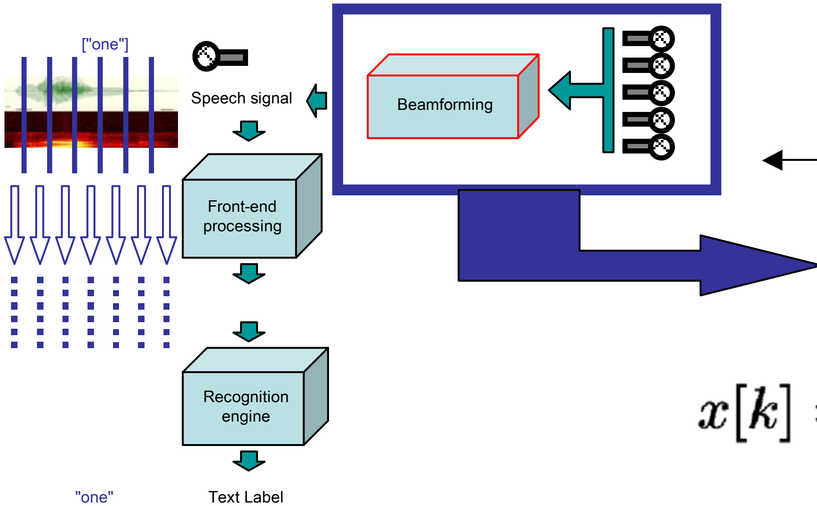
Environmental robustness in ASR

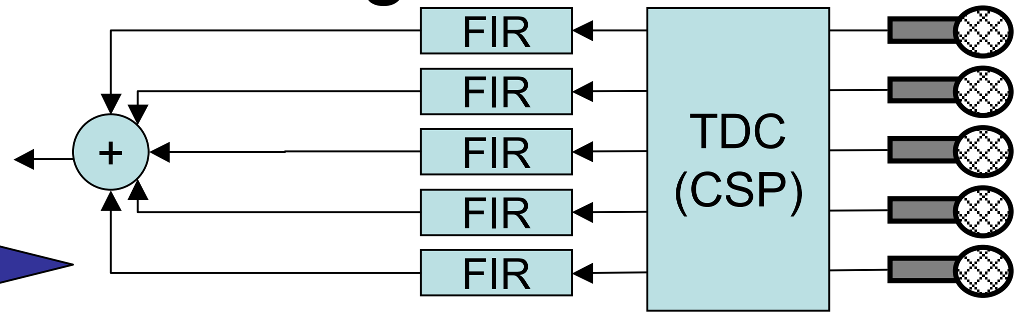


Purpose of this thesis: improve Speech Recognizers performances against background Additive Noise and Convolutional Distortions using Microphone arrays:

- Time-Frequency algorithms for single microphone can be extended and adapted thanks to the spatial dimension added by a microphone array.
- Rely as least as possible on noise estimation techniques (blind adaptation)

Beamforming





$$x[k] = \sum_{m=1}^{M} h_m[k] * s_m[k]$$

- Delay and Sum Beamforming is the simplest way of enhancing speech: FIR are set to [1,0…,0], or, alternatively, to [0,..,0, τ_m ,0…,0] if the TDC block is absent.
  - Useful to compensate for diffuse additive noise.
  - Does not compensate neither for directive noises nor for reverberation.

- If filters are not deltas then we deal with Filter and Sum Beamforming. Filter can be fixed or adaptive.

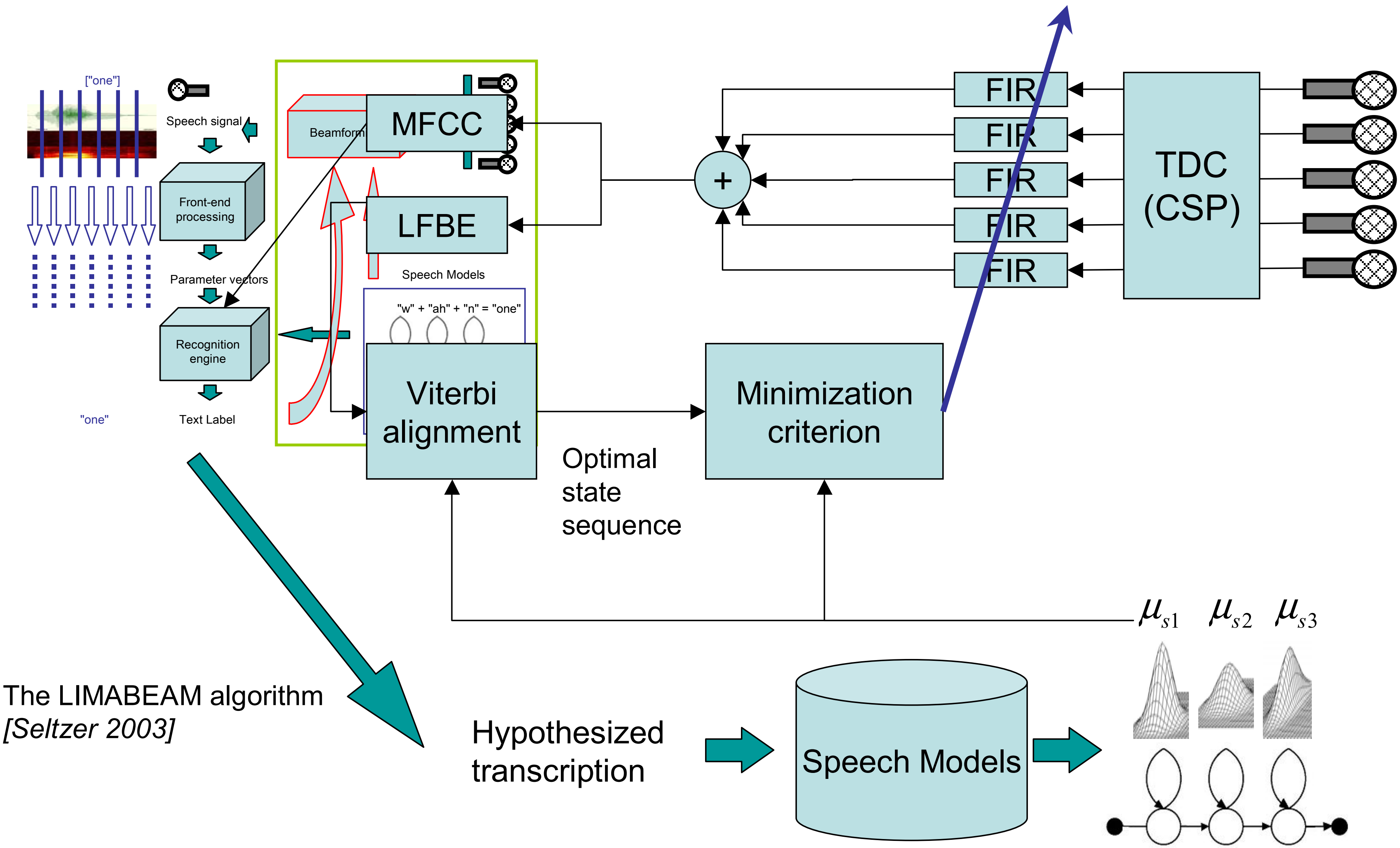
- More sophisticated methods exist to combat additive noise (Generalized Sidelobe Canceler, Superdirective Beamformer) or reverberation (Matched Filtering), but they adopt a criterion which maximizes the SNR (e.g. calculating an inverse filter of the room impulse response) .

  HMM-base speech recognizers do not act as human listeners (no SNR).

  We want an utterance to be better recognizable, not better audible.

  The criterion to maximize should be the same of the recognizer (likelihood)

Enhancement vs. Recognition: how to optimize FIRs?



The LIMABEAM algorithm
[Seltzer 2003]

SINGLE multi-variate gaussian model of "one"

How to get better than LIMABEAM?

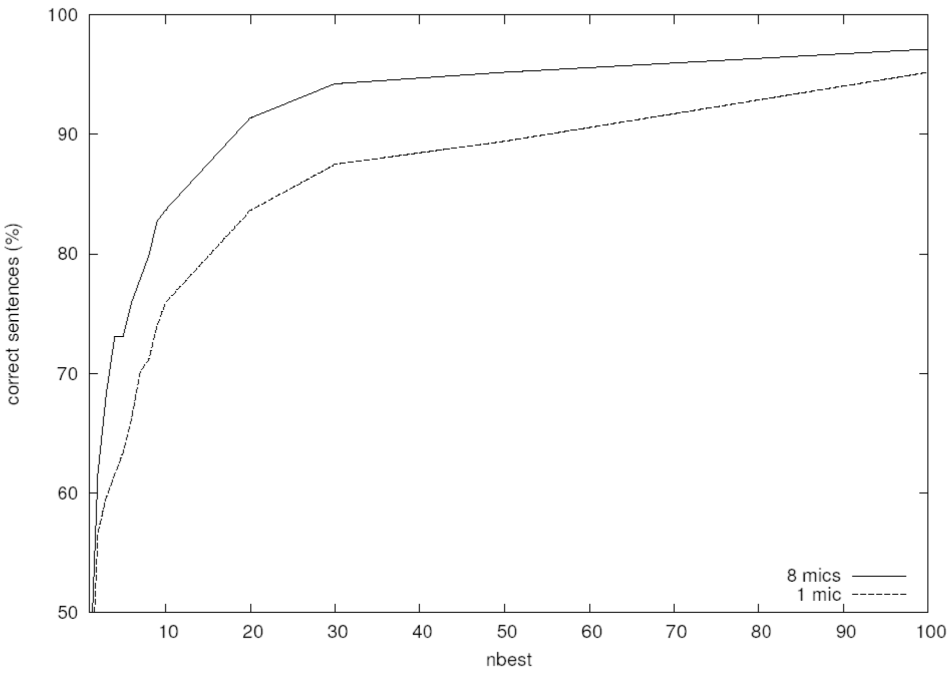
1) By looking closer to the algorithm, we realized that

- it is an adaptation algorithm: performance of optimization strongly depends on the transcription output of the **first** recognition step.
- if we skip the first step and directly provide the correct phrase (Oracle Limabeam), the algorithm NOT ALWAYS converges to a better solution (surprising). Mismatch Likelihood-Word Recogntion Rate.

- Providing a good alignment (from the RECOGNIZER point of view) should always improve performances.

2) Independently on the signal processing method, we found that the correct sentence is “pushed up” in the N-best list of recognized sentences if a microphone array is used.

- We propose to run N-best instances of Limabeam in parallel. After optimization each phrase will have a final acoustic score, which will **automatically re-rank** the N-best list. ML phrase will be chosen.



N-best Limabeam

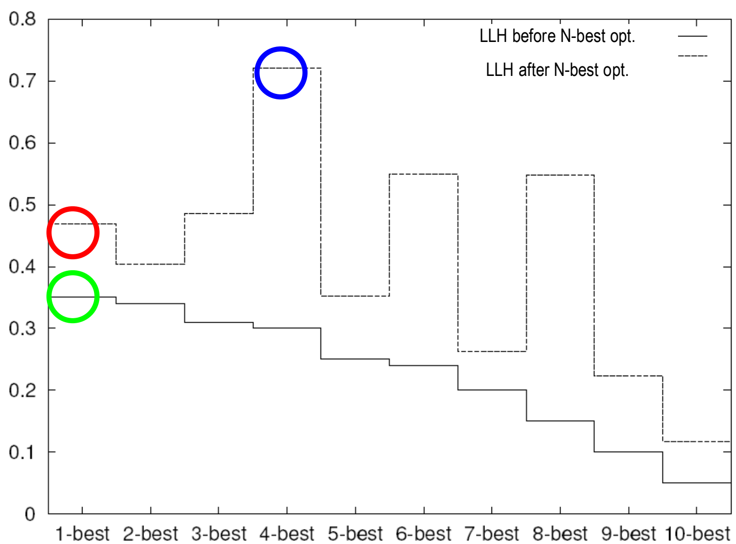
$$\mathbf{y}_L(\mathbf{h}) = \log_{10}\left(W \left| \mathrm{FFT}(\mathbf{x}(\mathbf{h}))\right|^2\right)$$

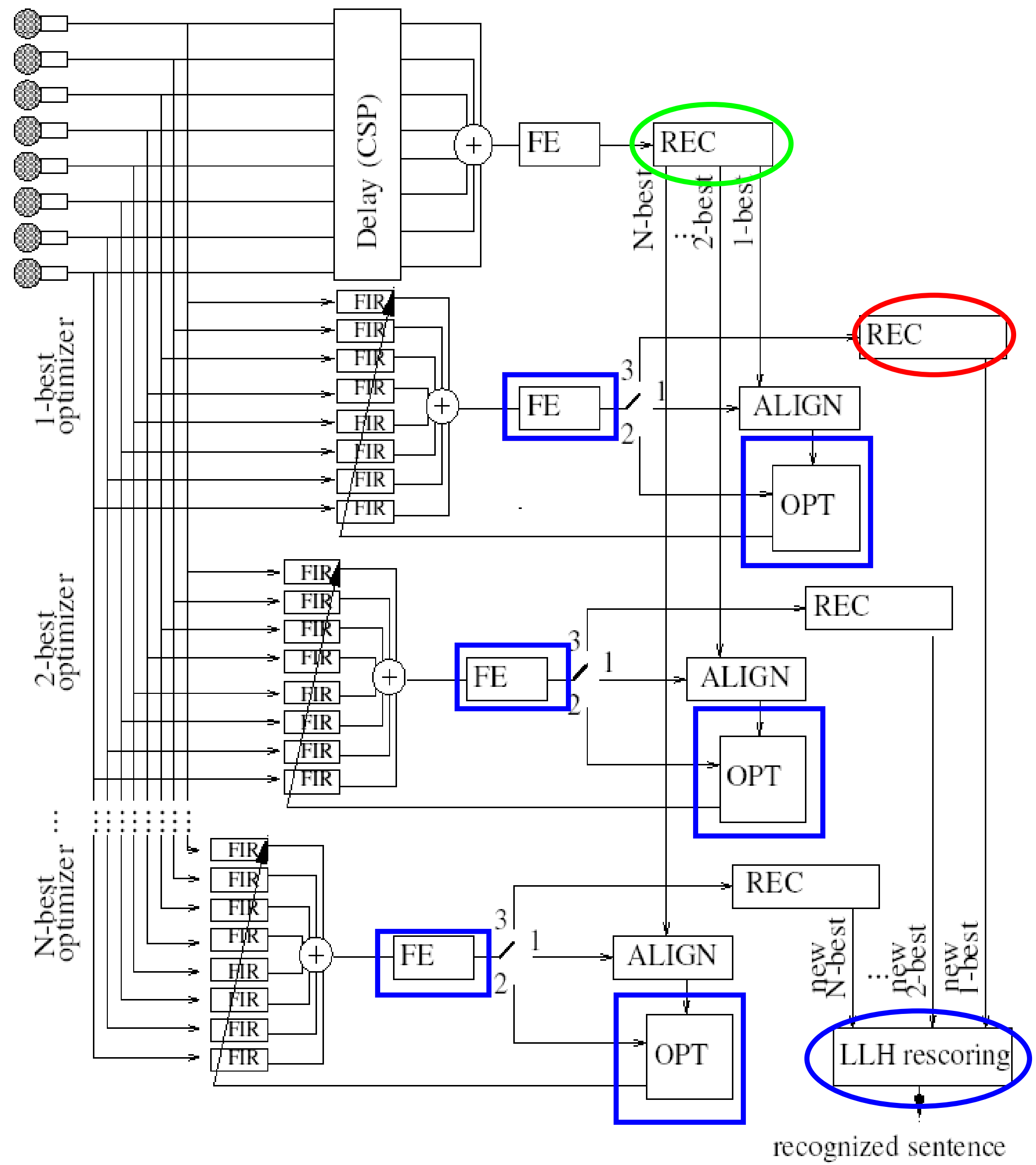
$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h}) \,|w\,)$$

$$\hat{\mathbf{h}}_n = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h}) \,|w_n\,)$$

$$\hat{n} = \arg\max_{n} P(\mathbf{y}_C(\hat{\mathbf{h}}_n) \,|\hat{w}_n\,)$$

$$\mathbf{y}_C(\mathbf{h}) = \mathrm{DCT}\left(\mathbf{y}_L(\mathbf{h})\right)$$





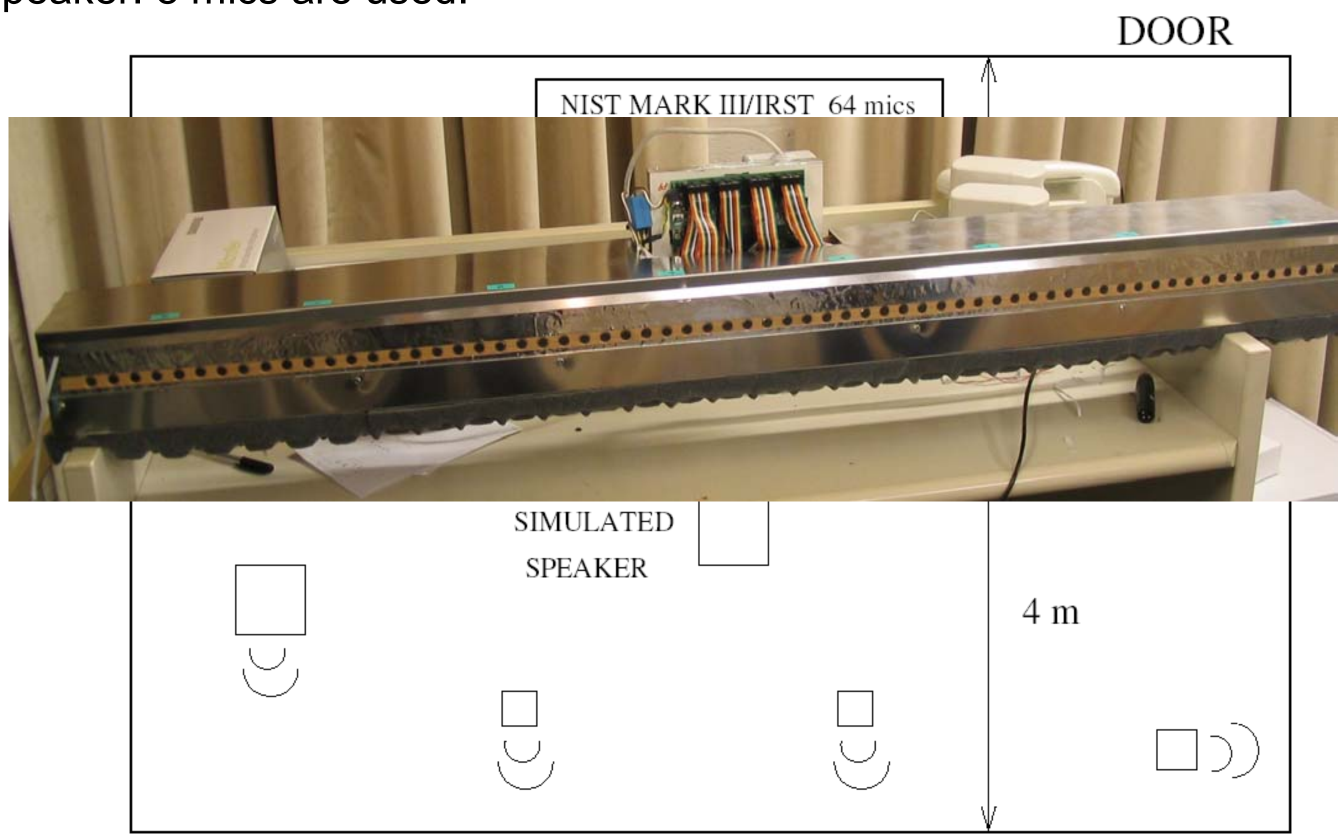
<u>The rank in the N-best list is automatically changed</u>

Environmental setup and Task

We analize performance of our N-best approach:

• with **simulated** data : real additive noise recorded from a computer fan is synthetically added to clean speech, simulating a 8-microphone array)

• in a **real** environment : real cockpit-like noise is spread from 8 speakers in a quasi-anechoic room (at ITC-IRST, Trento,Italy) T60=143 ms. Clean speech comes from a central high quality speaker. 8 mics are used.



MarkIII/IRST:

• 64 channels (8 used by now)

• data sampled @ 44100 kHz, 16 bit.

• partially **redesigned** by us

Recognition engine:

• HTK v 3.2.1

• flat language model

Task:

• English TI-digits (11)

• silence/pause models

Front-end:

• 39 MFCC (s+Δ+ ΔΔ)

• window size: 25 ms

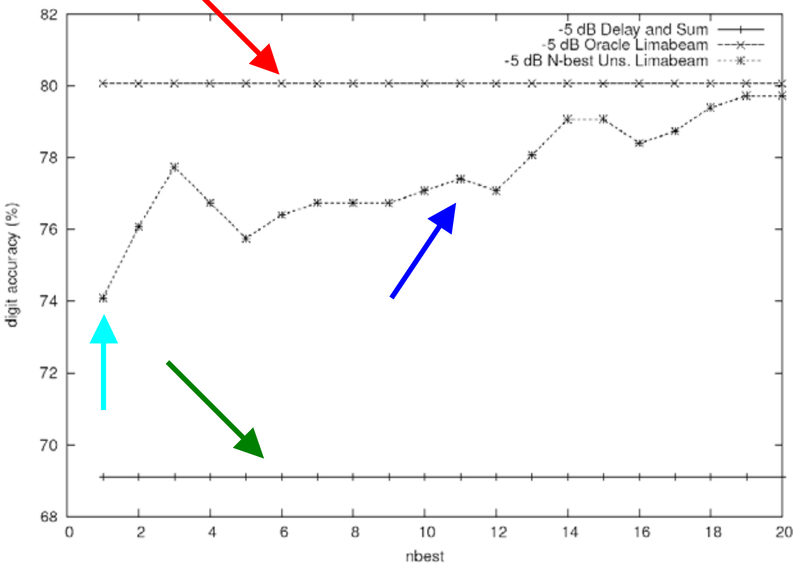
• frame rate: 100 fps

Back-end:

• word-level HMMs

• 1 or 3 multi-variate Gaussians per state

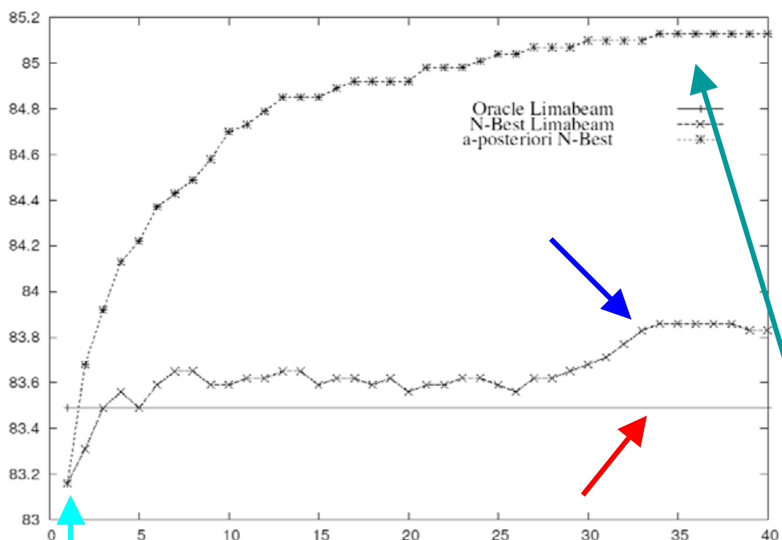
Experimental results



$$Accuracy = \frac{\#\,Correct - Ins}{Total\ \#}$$

-5 dB	D&S	Uns. Lim.	Oracle Lim.
1 ch	63.79%	66.11%(6%)	70.43%(18%)
8 ch	69.10%	74.09%(16%)	80.07%(35%)

	D&S	Uns.Lim.	N-best Lim.	Oracle Lim.
15dB	98.34%	98.34%	98.34% (1)	98.34%
5 dB	95.68%	96.35%	96.68% (9)	96.68%
-5 dB	69.10%	74.09%	79.73% (19)	80.07%



mic	1	9	17	25
Acc.	50.76%	57.26%	63.91%	61.46%
mic	33	41	49	57
Acc.	62.52%	64.21%	62.76%	52.69%

	D&S	Uns.Lim.	N-best Lim.(40)	Oracle Lim.	a-post(40)
Sup	-			X	X
Uns	-	X	X		X
Acc%(RI%)	80.74%	83.16%(12.5%)	83.83%(16%)	83.49%(14.2%)	85.13%(22.8%)

With a better criterion we could achieve this!

Ongoing work and Applications

• We presented a multi-microphone, multi-pass algorithm, which can be improved thanks to a multi-hypothesis approach.

• Ongoing work is focusing on:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|w) \;\Rightarrow\; \hat{\mathbf{h}} = \arg\max_{\mathbf{h}} \sum_{i=1}^{N} \phi_n P(\mathbf{y}_L(\mathbf{h_n})\,|w_n)$$

➢Modifying the optimization criterion [implemented, testing]

➢directing the microphone arrays towards multiple reflections of the speech signal on the walls (exploiting multipath) [submitted to ICSLP 2006]

➢ designing off-line ML FIR filters which work well in very reverberant environments (T60> 600 ms) [implemented, testing]

• ASR is already on the market for close-talk applications (dictation, reservations by phone), where performance are higher.

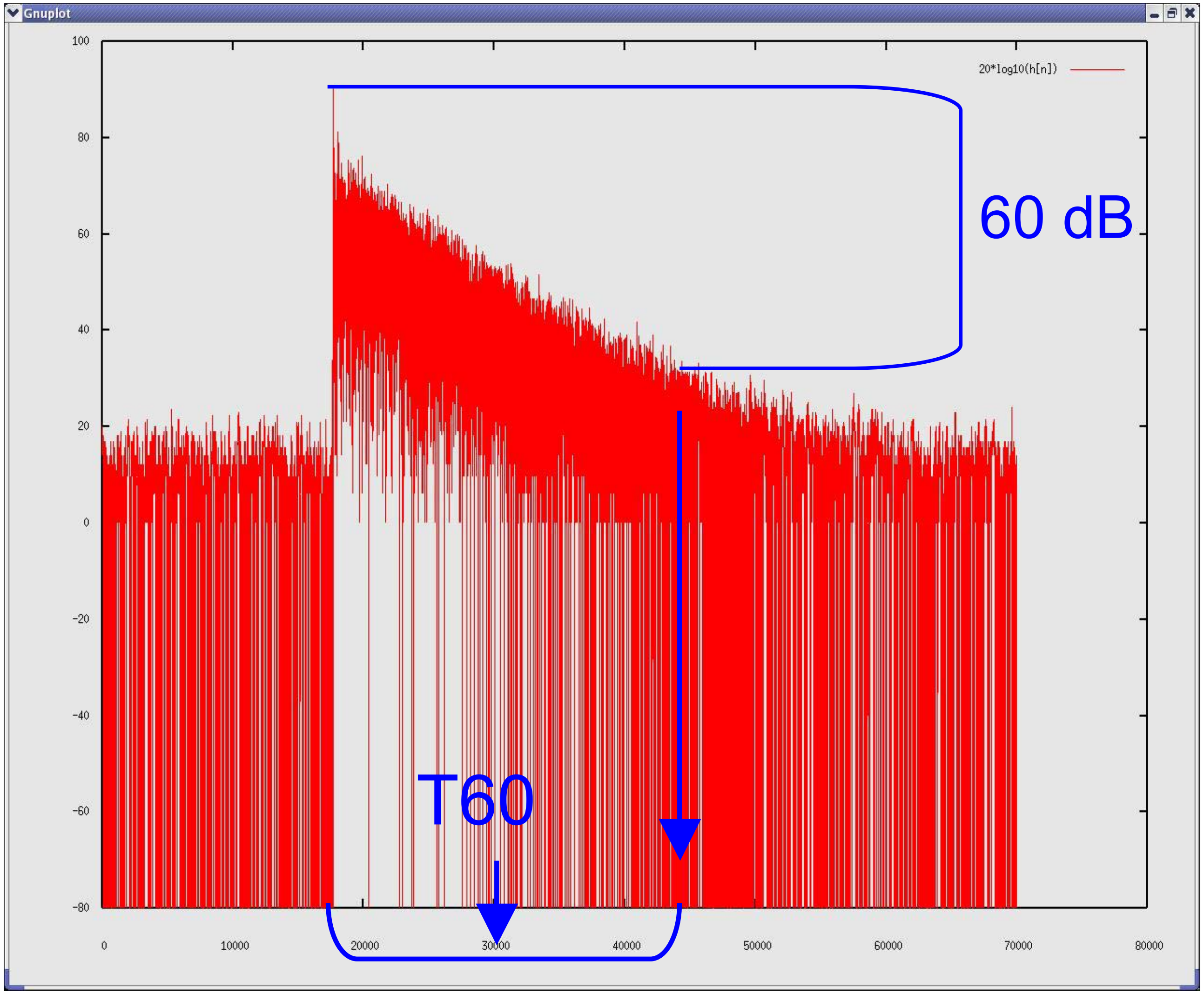
• Noise and Echo- robust algorithm allow **Distant-talking** ASR to be used in **automatic meeting transcription** (Parliament), voice-driven **medical reporting**.

• **Hands-free** ASR allow to develop applications to make easier **in-car** human-computer interaction (voice commands, navigation), **domotics** (no more TV remote control?), voice-based videogames, **deaf** people (speech-to-text on a display) and **blind** people (speech-to-text + text-to-speech) assistance. Definitely useful.

Thank you for you your attention!

Questions?

Appendix A: T60



Appendix B: Matched Filtering

$$x(t) = \sum_{n=1}^{N} s(t) * h_n(t) * h_n(-t)$$

SIMO ➡ SISO

per mic FIR filter

$$\boxed{SNR_{D\&S} = \frac{N}{K-1}}$$

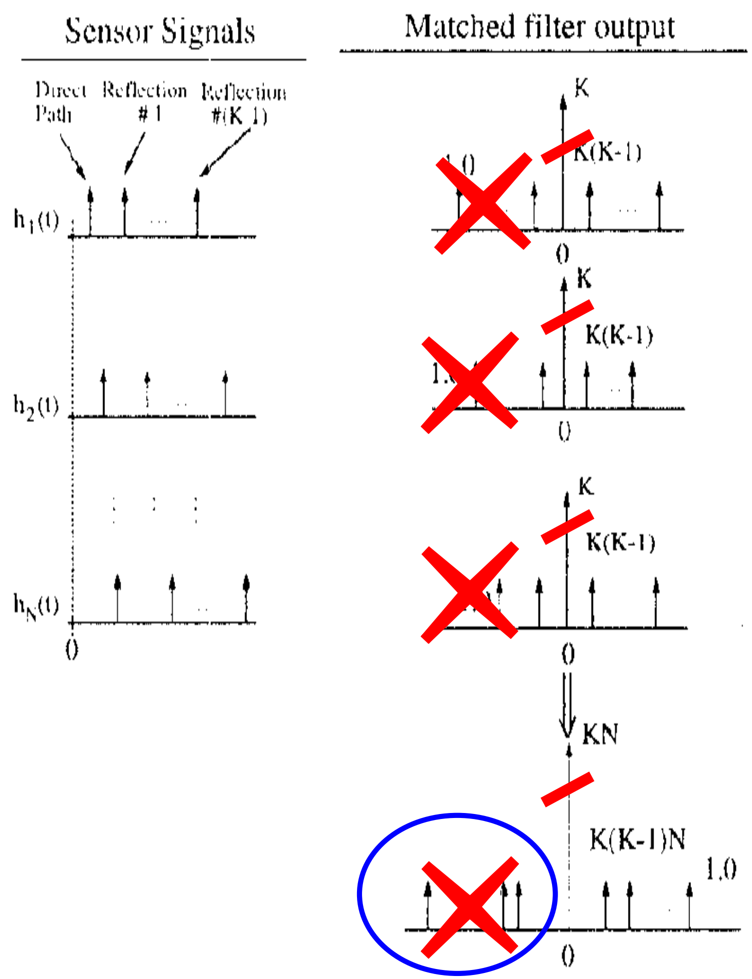
$$\boxed{SNR_{MF} = \frac{KN}{K-1}}$$

D&S:

- Reduces the output power for directions other than that of steering location by means of destructive interference.
- Applies a low-pass filter (while low frequency resolution is important for ASR).
- Wrong inter-channel delay estimates lead active beamformers to imperfect steering.

MF:

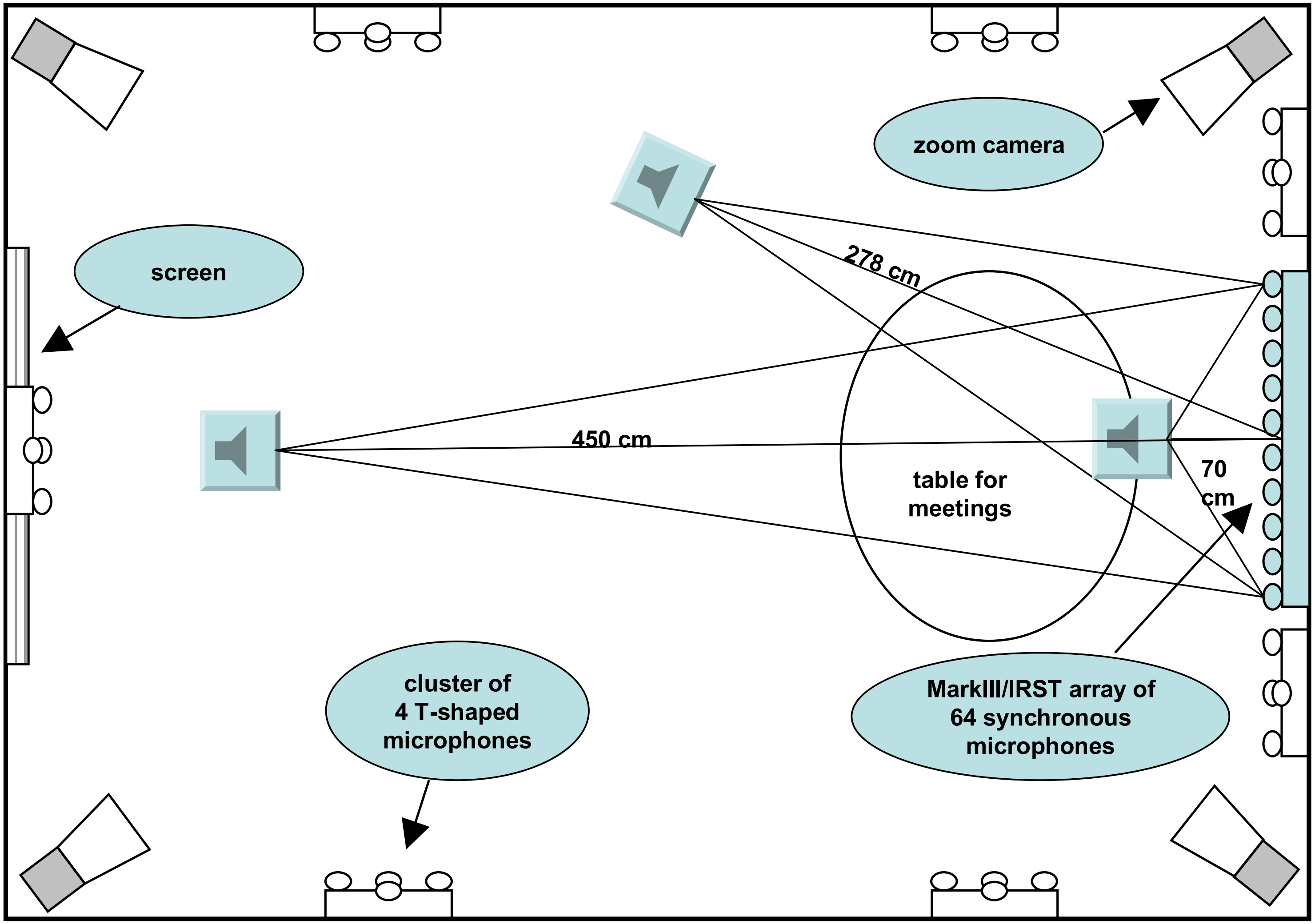
- Increases much more the SNR, but introduces an anti-causal effect which generates an "early echo", This artifact is NOT taken into account by HMMs trained with clean speech



Output of N matched filters for K multipath

These methods introduce artifacts affecting a human listener differently from a recognizer.

Apendix C: The microphone network at IRST



Experiments reported here deal with:

- Speaker in the furthest (and most challenging) position form the array (seminar-like config.)
- Additive noise coming from the right at different SNRs
- Waveforms sampled at 44100 Hz, 24 bits by the MarkIII array

Dataflow of > 8 MB/s

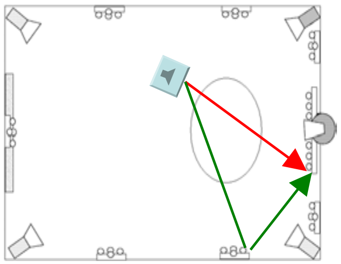
- Speech processing on parallel CPUs
- Big storage requirements

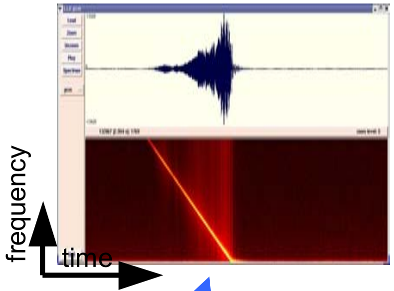
The CHIL room is:

- 600 x 470 x 300 cm
- used for lectures and meetings
- equipped with more than 100 microphones
- a very reverberant environment (T60=600 ms)

- suitable to test ASR in a real environment.
- useful when coupled with the IRST anechoic chamber to test algorithms (**and instruments!** we'll see the Appendix if we have time) in a more quiet and controlled environment.

Appendix D:Room Transfer Function measurement





We chose to measure room impulse responses with CHIRP (aka Time Streched Pulses) signals because:

- Simple signals, better than an utterance because their autocorrelation is a delta
- A real delta would cause dynamics, physical-breaking problems.
- Chirps have a flat frequency response

→ energy distributed → accurate measure.

We also have results (not shown here) when simulating the multipath via Image Method[Allen, Berkley '79]

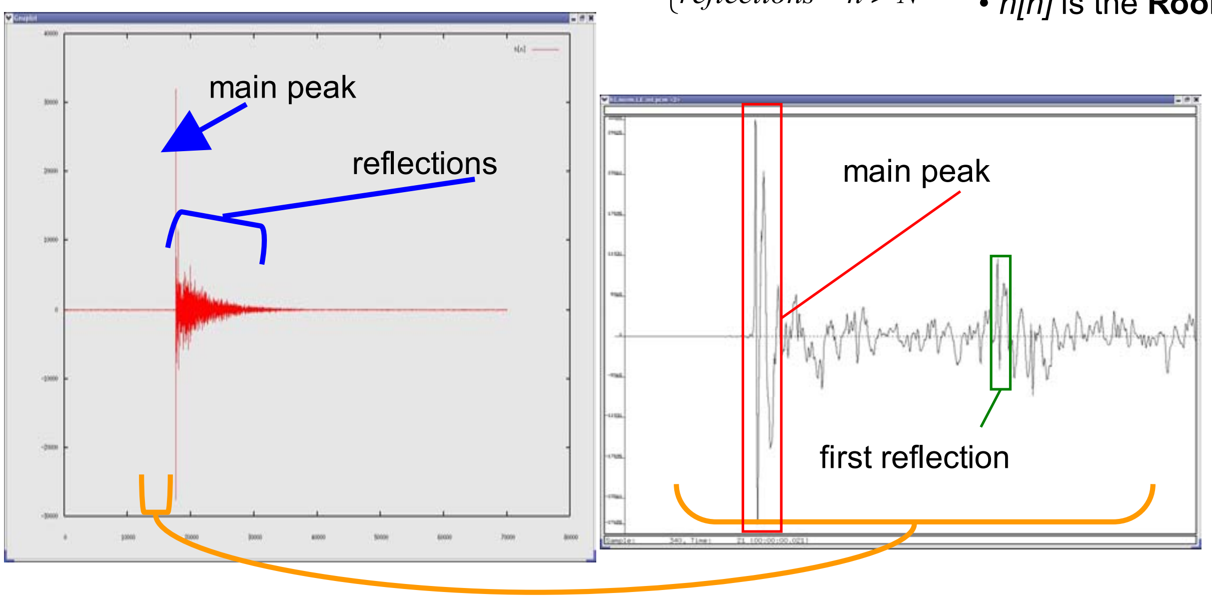
$$x[n] = \sum_{k=0}^{2N-1} chirp(k-n)\, chirp(k)$$

$$h[n] = \sum_{k=0}^{2N-1} chirp(k-n)\, revchirp(k)$$

$$x[n] = \begin{cases} \delta[n] & for\ n = N \\ 0 & elsewhere \end{cases}$$

$$h[n] = \begin{cases} 0 & n < N \\ \delta[n] & n = N \\ reflections & n > N \end{cases}$$

- *h[n]* characterizes the multipath propagation inside the room from a SINGLE source to a SINGLE microphone -> 64 IR have to be collected
- *h[n]* allows to create realistic models for far-microphone signals acquired from real talkers.
- *h[n]* is the **Room Transfer Function**



- 44 kHz clean chirp signal [chirp(k)]
- 44 kHz reverberated chirp signal [revchirp(k)]
- Room IR at 4,5 m from the array [h(n)]