AUDITORY MODELS FOR SPEECH RECOGNITION

Fernando S. Perdigão

Instituto de Telecomunicações - Pólo de Coimbra
Dept. Eng. Electrotécnica - Pólo II da Univ. Coimbra
3030 COIMBRA - PORTUGAL
fp@co.it.pt

Research on Auditory Models

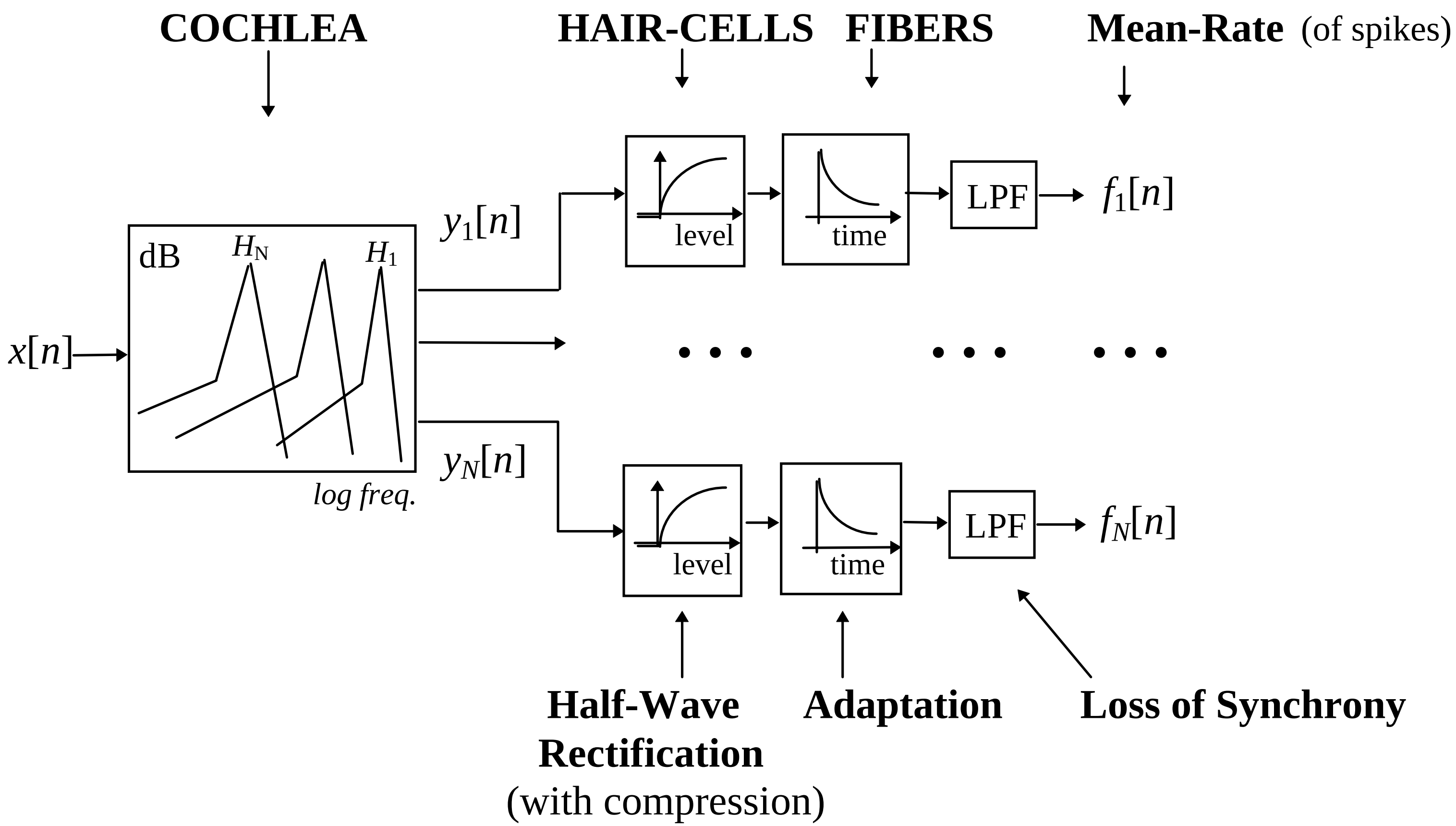
MOTIVATION:

- Auditory Models have shown to be superior in recognition tasks when environment degrades (additive noise + linear filtering).
- There is not a deep understanding of their functioning.

OBJECTIVE:

- Verify if there are advantages in the characterization of speech signals using models of the auditory periphery.
- Compare Auditory Models with other speech analysis methods.

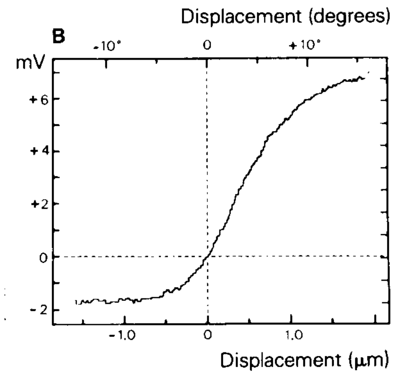
MAIN CHARACTERISTICS OF PERIPHERAL AUDITORY PROCESSING



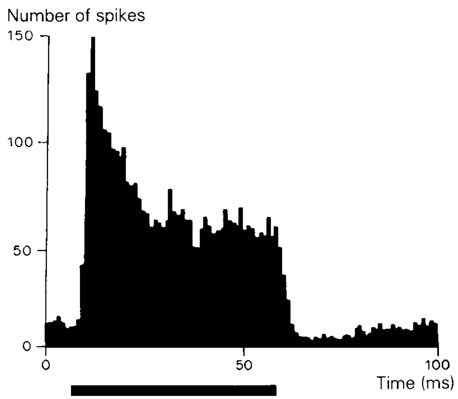
INNER-HAIR CELLS AND NERVE FIBERS

Main Characteristics:

- Half-Wave Rectification (IHC transduction is directional)
- Auditory Fibers Firing-Rate
  - Spontaneous and Saturation values
  - Threshold of excitation
  - Adaptation

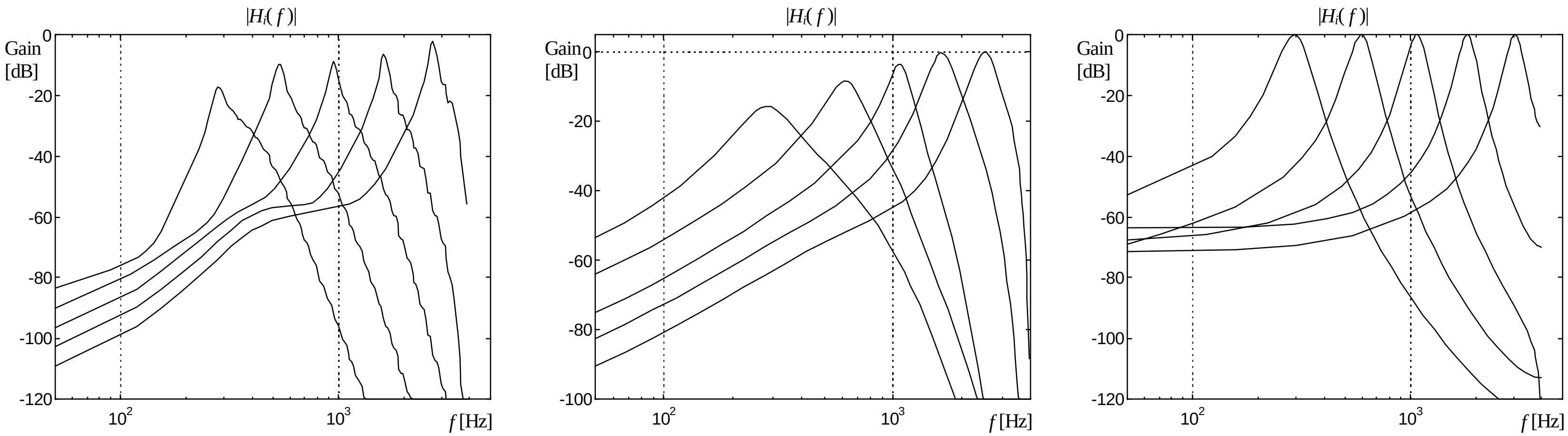


IHC transduction



Time-Histogram of Discharges

COCHLEAR RESPONSES
Modeled with a Filter-Bank

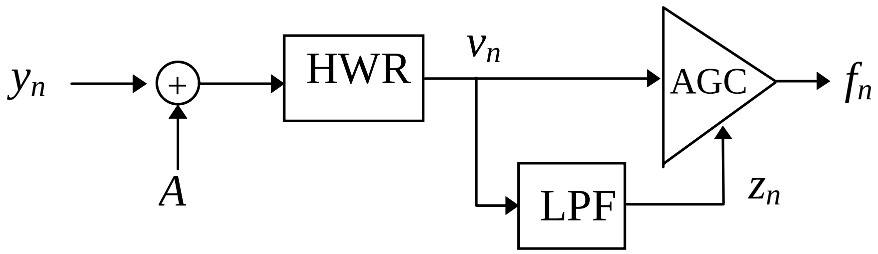


a) Cochlear Model b) Seneff - Stage I c) Gamma-Tone

Frequency responses of 3 filter banks

Block Diagram of the IHC/Synapse Model

(Martens-Immerseel's Model)



HWR : Half-Wave Rectification (IHC transduction is directional)

AGC: Automatic Gain Control (as a function of q_n) : $f_n = \dfrac{v_n f_{sat}}{\left(B + z_n^{\frac{1}{2}}\right)^2}$

LPF: Low-Pass Filter

Problems with IHC/Synapse Models:

- Due to non-linearities they have to be simulated in time, on a sample-by-sample basis. This demands a high computational load.
- The output mean-rate is then decimated to have a feature representation every 10ms.

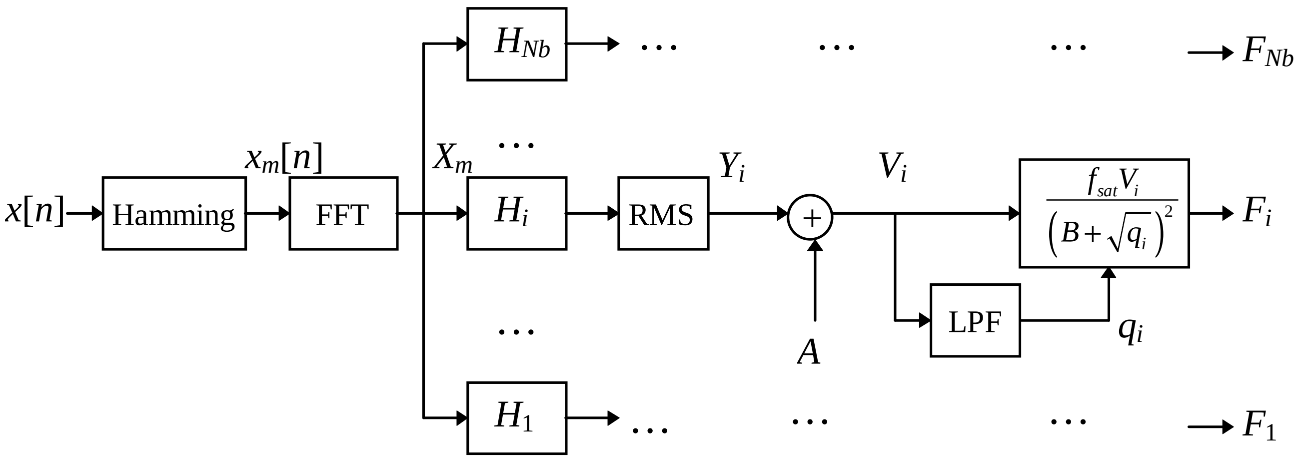
But, the study carried out shows that:

Adaptation can be reasonably modeled in terms of the short-term envelope of the energy (or RMS value) of the sub-band signals.

⇒ *Functional Model of Adaptation*

⇒ Energies are computed in frequency domain (using the FFT).
⇒ Adaptation is modeled with RMS values.

Functional Model of Adaptation



$$Y_i[m] = \frac{1}{N}\sqrt{\sum_{k=0}^{N-1}\left|X_m(k)H_i(k)\right|^2}$$ (Square root of short-term energy)

H_i: filter i of the filter bank (Gamma-Tone).
A: rate threshold
F_i: output firing rate.

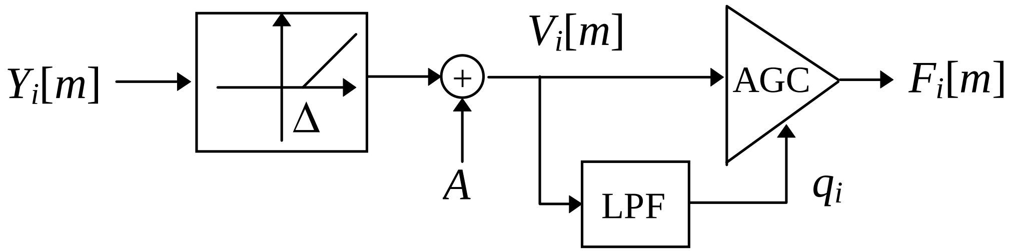
Adaptation Model operates with a sampling frequency of 100Hz (frame-rate).

NOISE SUPPRESSION

- Mean-Rate representation is <u>very sensitive to noise</u>.
  (noise increases mean-rates and adapts the response due to speech).

- [Vereecken & Martens, Eurospeech'95] proposed a **Center-Clipper** in front of the IHC model.
- An analogous noise suppression technique was used, based on:

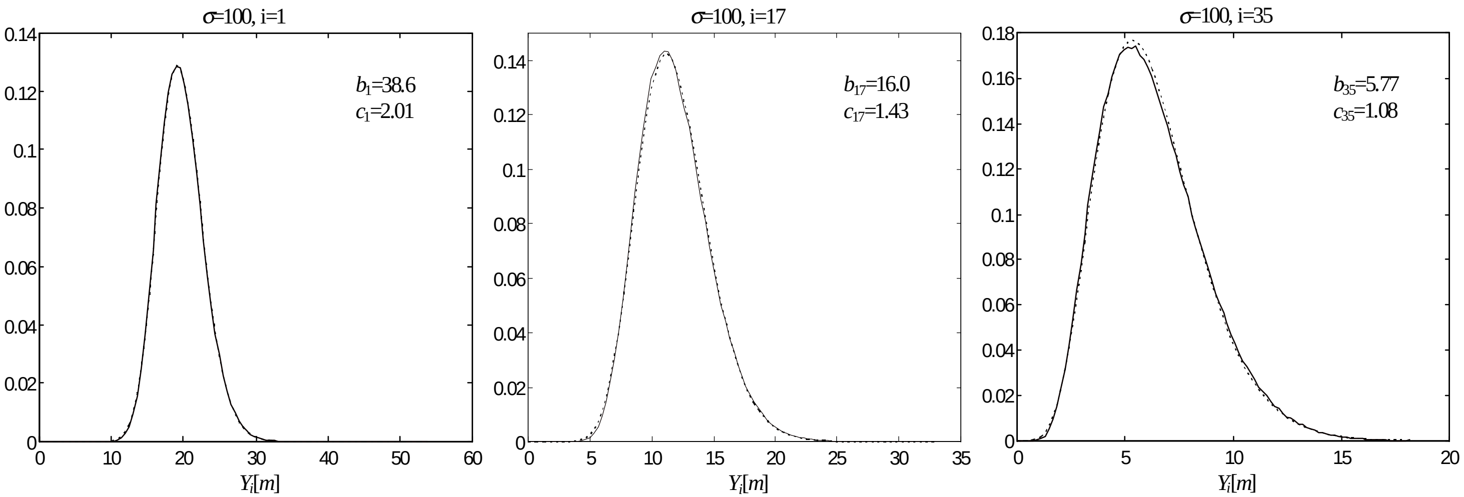
$$V_i[m] = A + \max(Y_i[m] - \Delta, 0)$$



- Δ is a function of the noise level, σ^2.
- kind of a spectral subtraction with a noise floor.

- The level Δ is chosen in order to keep the mean of V_i, $E(V_i[m]) = (1+\varepsilon)A$ (almost constant), and is estimated during speech pauses.
- The knowledge of the pdf of Y_i is needed.

- An empirical study showed that RMS values, Y_i, follow approximately a **gamma** distribution: $f_x(x) = \dfrac{c^b}{\Gamma(b)} x^{b-1} e^{-cx} \quad , x > 0$.



- Assuming a **gamma** pdf for Y_i, the non-linear function Δ, as a function of σ^2 , is calculated and the proper value of Δ is updated.

Theoretical Study

- Assuming: $\mathbf{x} \sim N\left(\mathbf{0}, \mathbf{C}_{\mathbf{x}}\right)$ (vector of N Gaussian random variables)

  Then:

$$Q_i = \frac{1}{N}\mathbf{x}^T \mathbf{B}_i \mathbf{x}$$

($\mathbf{B}_i$ is a circulant positive semidefinite matrix)

$$Y_i = \sqrt{Q_i}$$

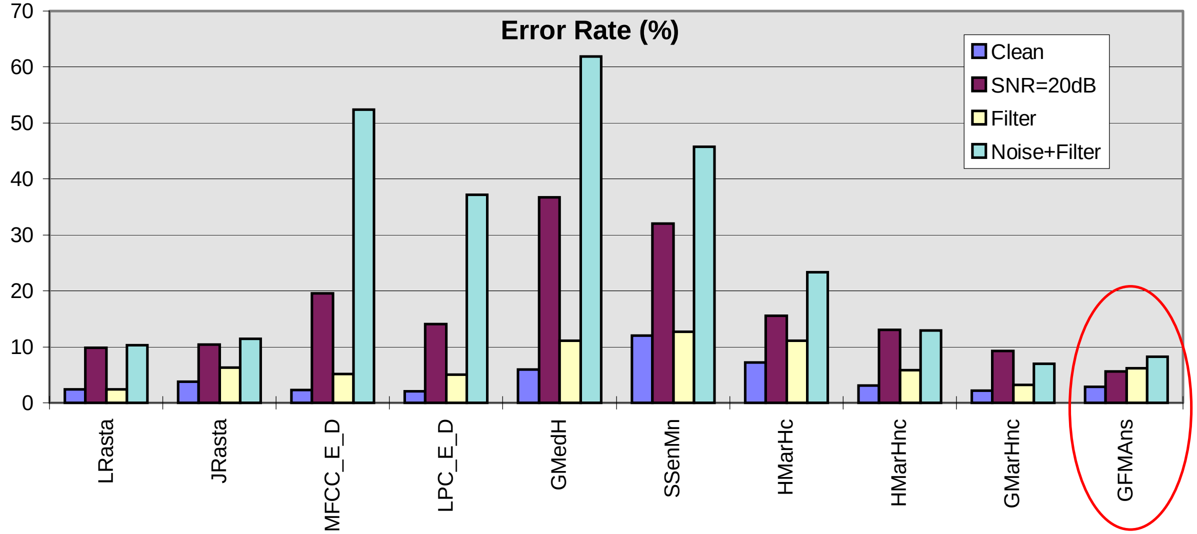
- 1st and 2nd statistics of Q_i:

$$\mu_{Q_i} = E\left\{Q_i\right\} = \frac{1}{N}\mathrm{tr}(\mathbf{B}_i \mathbf{C}_{\mathbf{x}}) \quad ; \quad \sigma_{Q_i}^2 = \mathrm{var}(Q_i) = \frac{2}{N^2}\mathrm{tr}\left((\mathbf{B}_i \mathbf{C}_{\mathbf{x}})^2\right)$$

- The pdf of Q_i has not a simple definition (nor Y_i).
- However Y_i can be approximated by a Gamma distribution.
- If $\mathbf{C}_{\mathbf{x}} = \sigma^2 \mathbf{I}$ then the variance of $\log(Y_i)$ (assuming Gamma pdf), does not depend on σ^2. This can be used to extend the work [Ephraim & Rahim, 99] to MFCCs.

ISOLATED-DIGIT EXPERIMENTS

Database: Telephone-speech digits, ≈800 speakers, 4200 digits (2100 for training).
Recognizer: CDHMM, 7 states, mixture with 6 components, diagonal covariance matrices. Silence model.



Label	*Clean*	*SNR=20dB*	*Filter*	*Noise+Filter*
L-Rasta	97.57	90.22	97.55	89.66
J-Rasta	96.27	89.59	93.71	88.56
MFCC_E_D	97.67	80.49	94.81	47.57
LPC_E_D	97.95	85.90	94.96	62.77
GMedH	94.00	63,33	88.96	38.16
SSenMn	88.01	68.00	87.30	54.25
HMarHc	92.84	84.43	88.88	76.64
HMarHnc	96.95	86.93	94.22	87.13
GMarHnc	97.85	90.76	96.79	93.01
GFMAns	97.19	94.45	93.77	91.73

G: Gamma-tone filter-bank
H: Cochlear model filter-bank
S: Seneff's filter-bank

Functional Model

Pre-Processing: signal normalization (n)
Post-Processing: clipping (c) or spectral subtraction (s)
Distortion: adding white noise (SNR=20dB) and/or filtering the speech signals

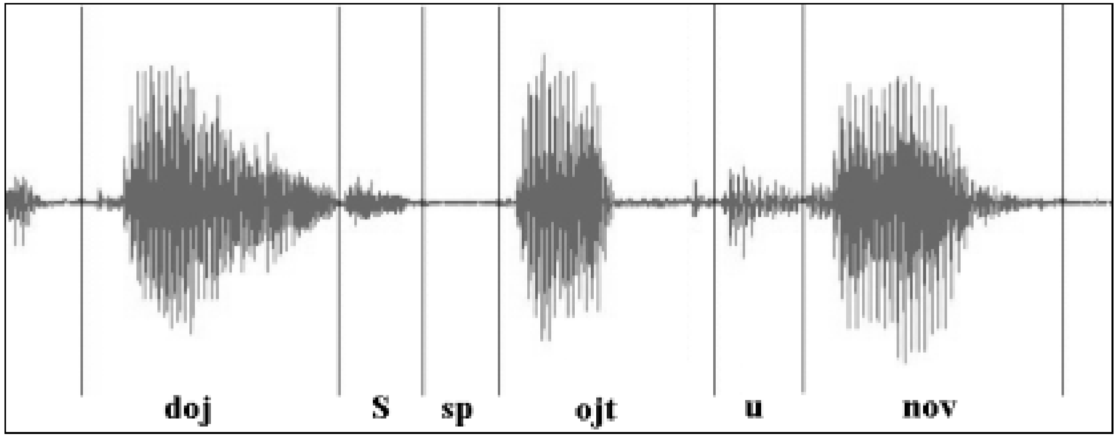
Recognition Experiments with Connected-Digits

(Elisabete Cordeiro, Jorge Rato, João Duque & Fernando Perdigão)

- Instead of using whole-word models or phoneme models...
  we used "syllable-like" models (phones).
- 5 noise/silence models
- Manual annotation of 100 sentences for model initialization.

Digits	SAMPA transcription		
/um/	u~		
/dois/	doj	S	
/três/	tre	S	
/quatro/	kwa	tru	
/cinco/	s	i~k	u
/seis/	s	6j	S
/sete/	s	Et	
/oito/	ojt	u	
/nove/	nOv		
/zero/	zEr	u	

Example of annotation



Experiments with Connected-Digits

- In order to take into account coarticulation between digits, triphones were used.
- Only 34 models were generated including monophones, diphones and triphones (with tied states).

Database: Connect-digits set from TELEFALA (sentences with 9 connected digits)
Training set: 1690 files
Test set: 849 files
Scores: 96% correct words (32 mixtures, 2 reestimations)

```
------------------------ Overall Results --------------------------
SENT: %Correct=74.68 [H=634, S=215, N=849]
WORD: %Corr=96.07, Acc=95.37 [H=7341, D=54, S=246, I=54, N=7641]
-------------------------------------------------------------------
```

- Need to improve results.

Recognition of Word-Commands for TV/VTR sets
(Experiments with the TIMIT Database)

(Gonçalo Pereira, Paulo Melanda & Fernando Perdigão)

- Task: Recognition of 37 word commands, e.g. /play/, /record/, /stop/ ...
  using sub-word models.

- Database: TIMIT

	Total	Train	Restricted test	Complete test
#sentences	6300	4620	192	1344
#distinct texts	2342	1718	192	624
#distinct words	6099	4891	912	2371
#phonemes	45	45	45	45

- Results for TIMIT: 46.6%
- Results for the 37 words: ≅ 94%

```
------------------------ Overall Results ------------------
SENT: %Correct=18.18 [H=2, S=9, N=11]
WORD: %Corr=93.61, Acc=93.37 [H=381, D=1, S=25, I=1, N=407]
===========================================================
```

- Only 44 recordings (33 for retrain, 11 for test) with 37 words each
- Portuguese accent.

CONCLUSIONS

- Auditory Models produce a rich representation of speech signals. However, Mean-Rate representation is very sensitive to noise and level variation in signals. Normalization and noise compensation is needed.

- The Functional Model of Adaptation works as well as or better than models operating in the time domain. It is almost as efficient as MFCC analysis.

- The REC project was very important to get experience on speech recognition systems. We intend to continue to research this area, specially on acoustic analysis for robust recognition.