Effects of Noise on a Speaker-Adaptive Statistical Speech Synthesis System

Jose Moreno
02.04.2014

GTH

- Madrid, Spain
- Speech Technology Group, ETSI. Telecomunicación, UPM

Chapter 1

Introduction

4th EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY
Madrid, Spain September 1995
adidas

Why?

- Obtaining data is not so easy...
- ...but many sources could be available
- GlottHMM
- R. Karhila, U. Remes, and M. Kurimo, Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments, Signal Processing, IEEE Journal of, Selected Topics in vol. PP, no. 99, pp. 11, 2013

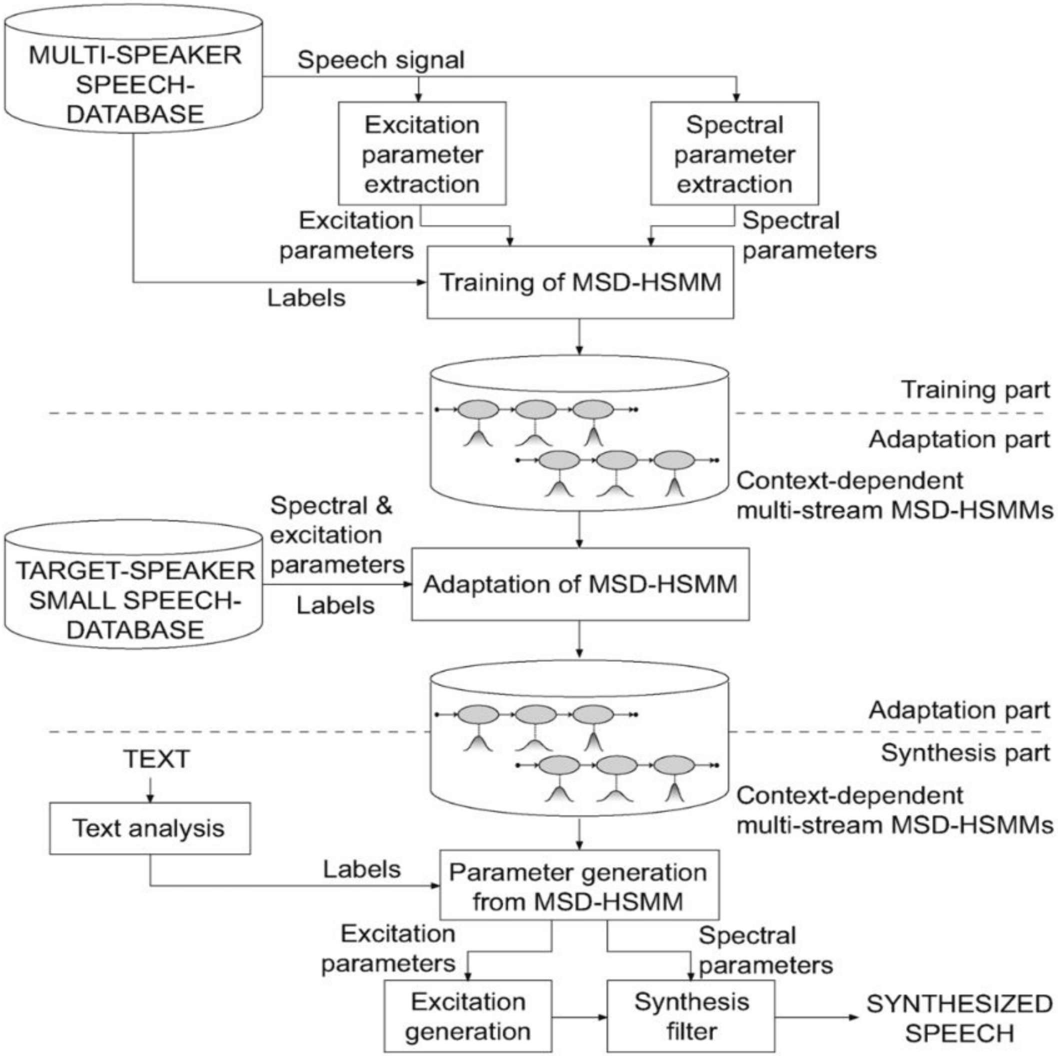
What?

- GlottHMM-based
  - Statistical
- Speaker-Adaptive
- Text-to-Speech (TTS)
- Noisy data
- Comparison to STRAIGHT

General Structure

Chapter 2

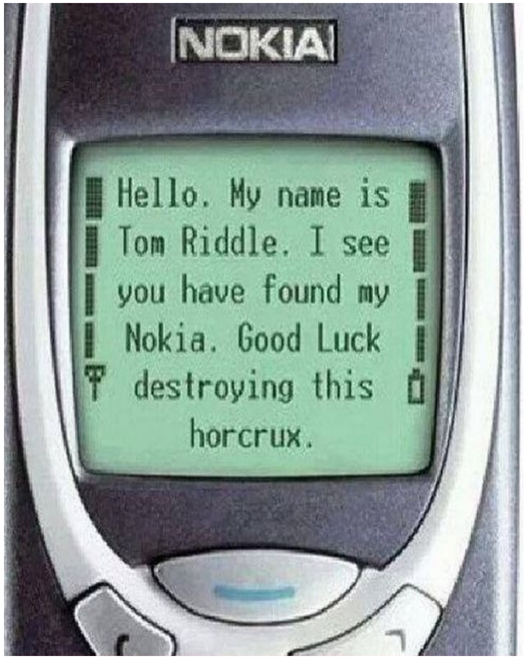
TTS System



Jose Moreno
02.04.2014

Labeling

- Full-context labels
- TTU/Nokia Finnish front-end
- Based on a set of pronuntiation (stress) rules from a “Finnish from foreigners” book
- Nice black box


NOKIA
Hello. My name is Tom Riddle. I see you have found my Nokia. Good Luck destroying this horcrux.

Feature extraction

- 30 LSF components
- 10 LSF source
- 5 HNR componentes
- F0
- 14 Aurora components
  - ETSI advanced front-end
- Antti Suni's scripts
  - Dynamic features
  - Global Variance (GV)

Average voice model

- Finnish PERSO corpus
  - 26 phonemes
- 20 male voices
- Context-dependent MSD-HSMM
  - EMIME 2010 Blizzard entry (modified)
  - SAT
  - 3 reclustering iterations

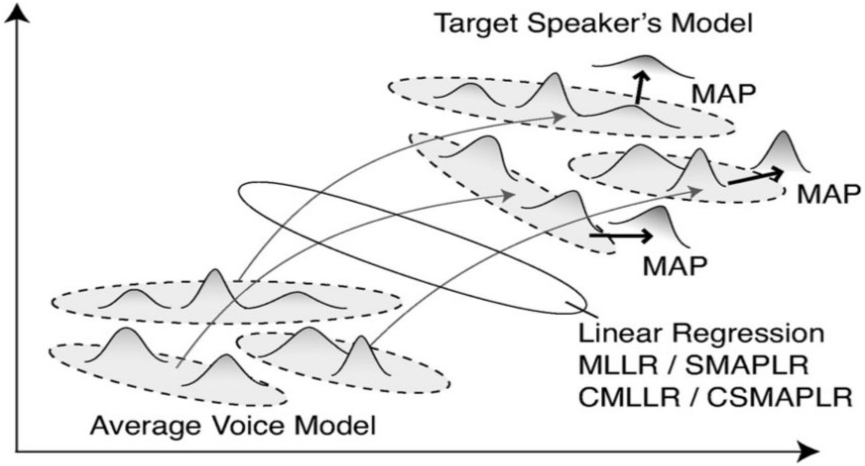
Adaptation (I)

- EMIME corpus
- NOISEX-92 corpus
  - Babble (20, 10 and 5 dB)
  - Factory (10 and 5 dB)
  - Machine gun (0 dB)
- Speech enhancement

Adaptation (II)

- Constrained Structural Maximum A Posterior Lineal Regression (CSMAPLR)
  - CMMLR, SMAP, MAP
- Combined algorithm:
  - 2 rounds of CSMAPLR
  - MAP

- Realign labels
- 64 leaf nodes regression trees
- Global Variance (GV)

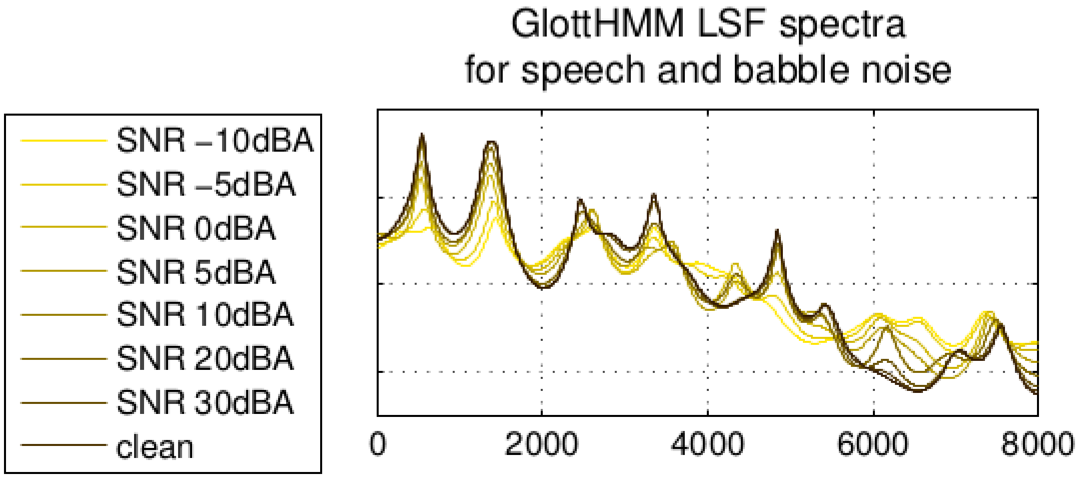


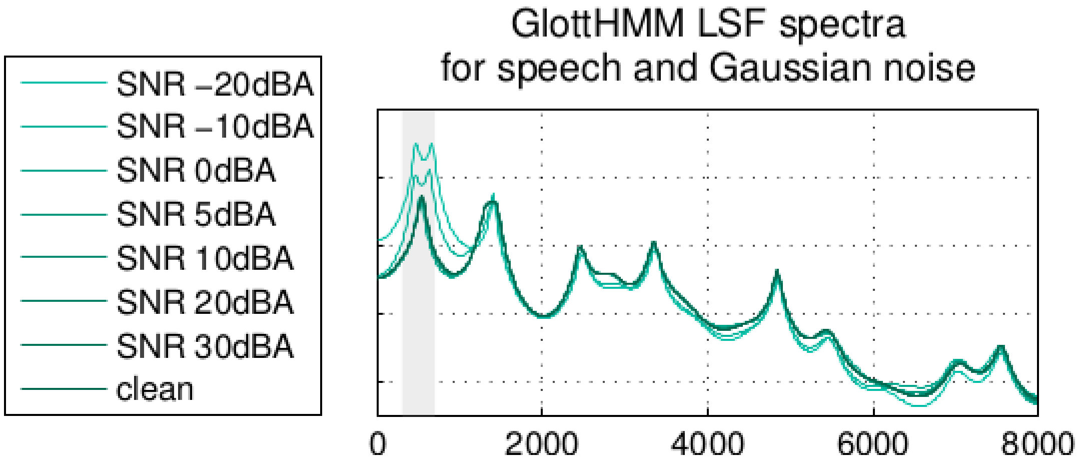
Jose Moreno
02.04.2014

Effects on features: LSF spectra

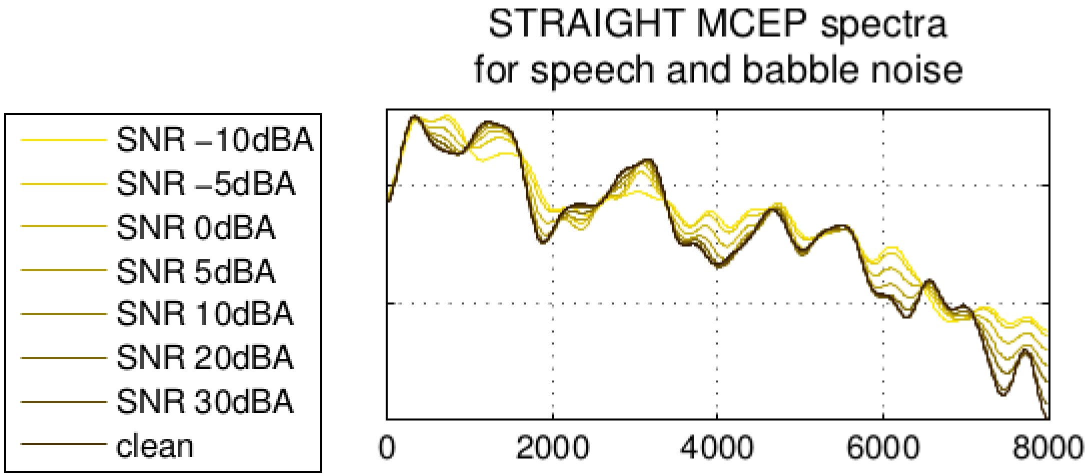
Chapter 3

Effects of noise

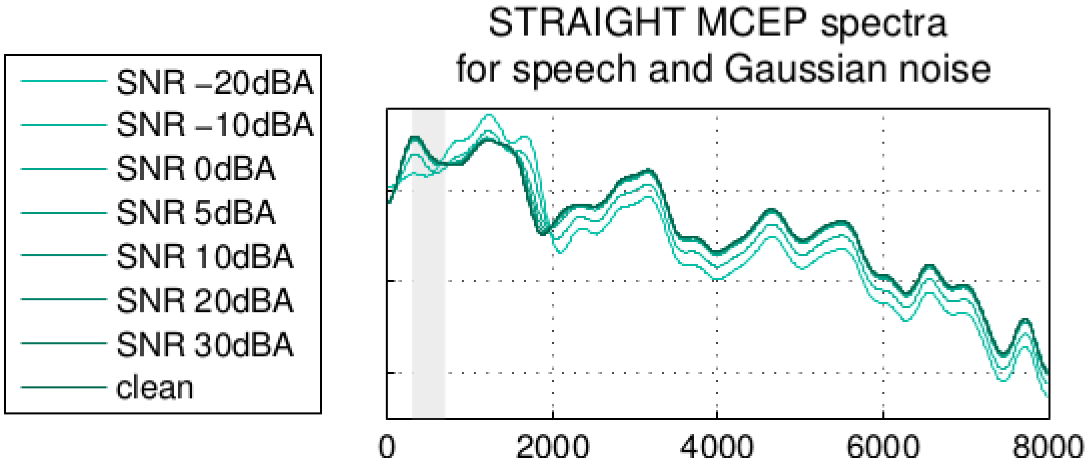

GlottHMM LSF spectra
for speech and babble noise


GlottHMM LSF spectra
for speech and Gaussian noise

Effects on features: STRAIGHT spectra

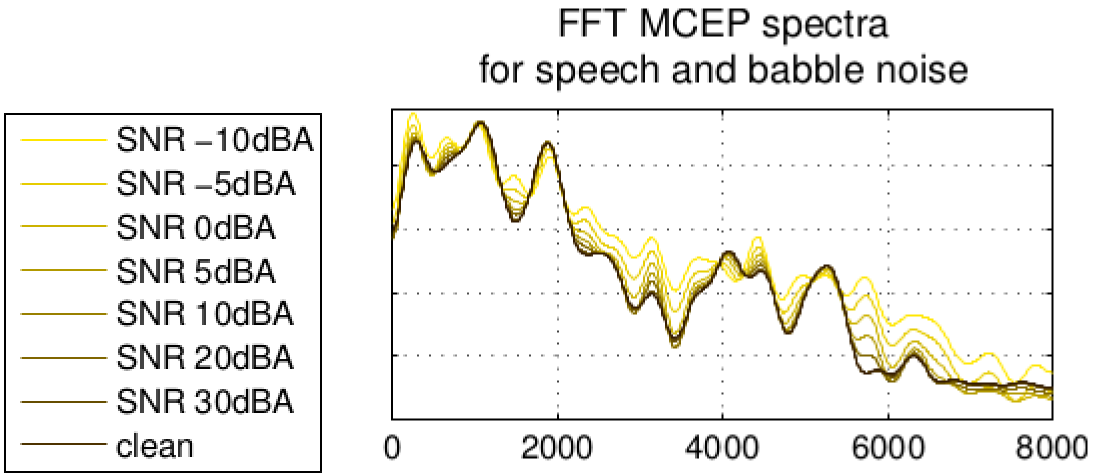


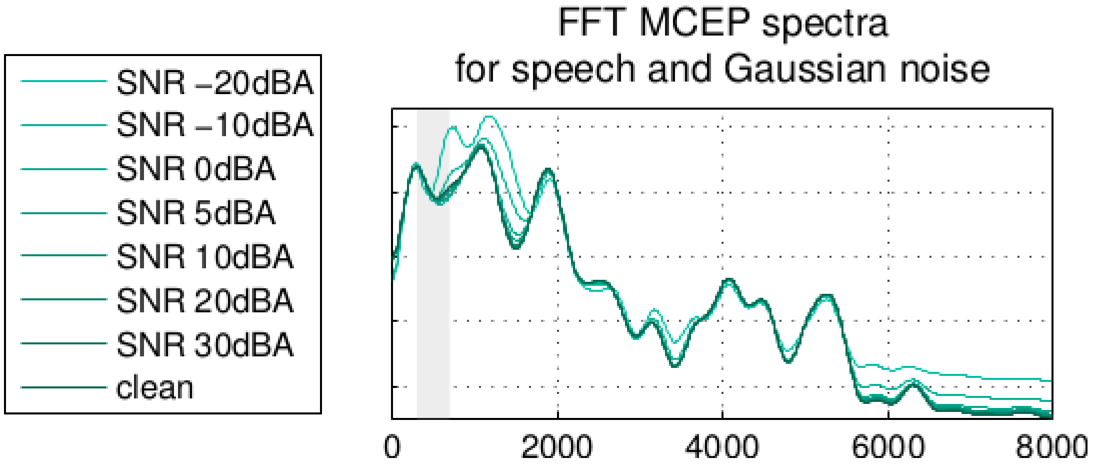
STRAIGHT MCEP spectra for speech and babble noise



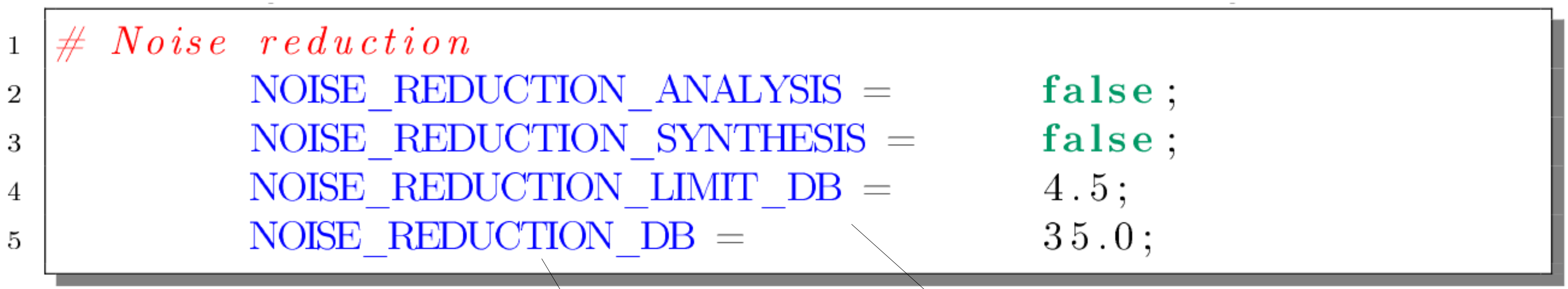
STRAIGHT MCEP spectra for speech and Gaussian noise

Effects on features: FFT spectra





GlottHMM noise reduction (I)



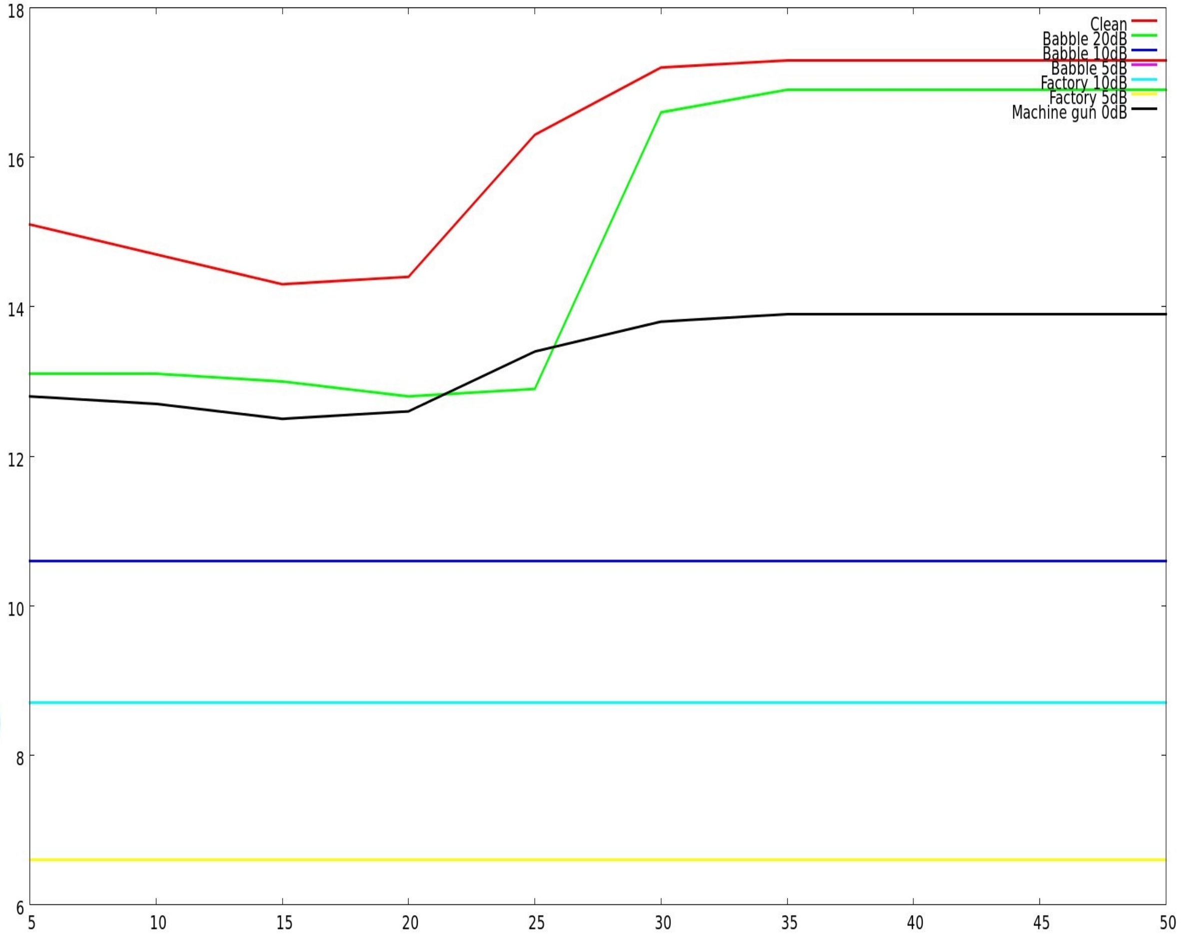
```
# Noise reduction
        NOISE_REDUCTION_ANALYSIS =        false;
        NOISE_REDUCTION_SYNTHESIS =       false;
        NOISE_REDUCTION_LIMIT_DB =        4.5;
        NOISE_REDUCTION_DB =              35.0;
```

Gain reduction

Minimum energy frame

GlottHMM noise reduction (II)

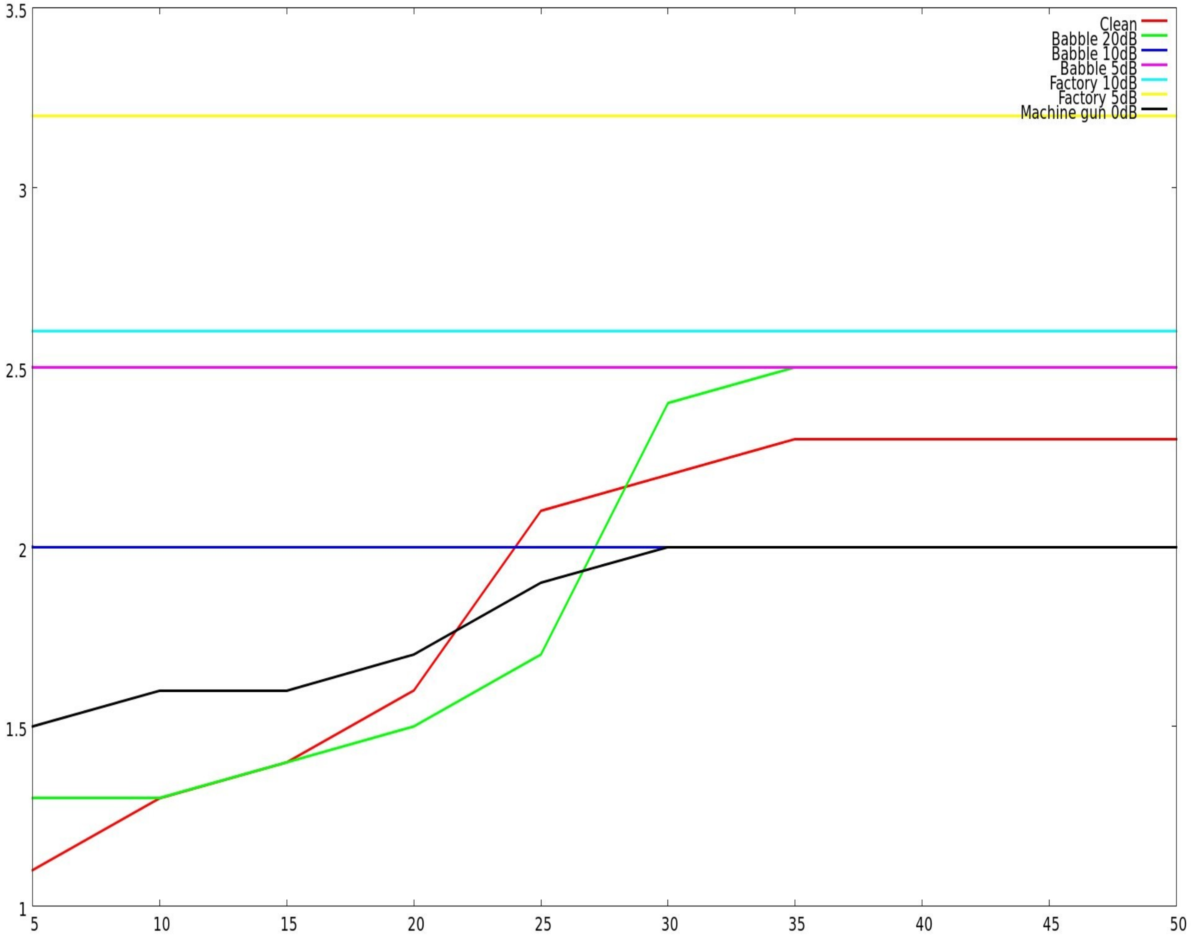
SNR
NOISE RED. LIMIT = 4.5
NOISE RED. = 5 – 50



Jose Moreno
02.04.2014

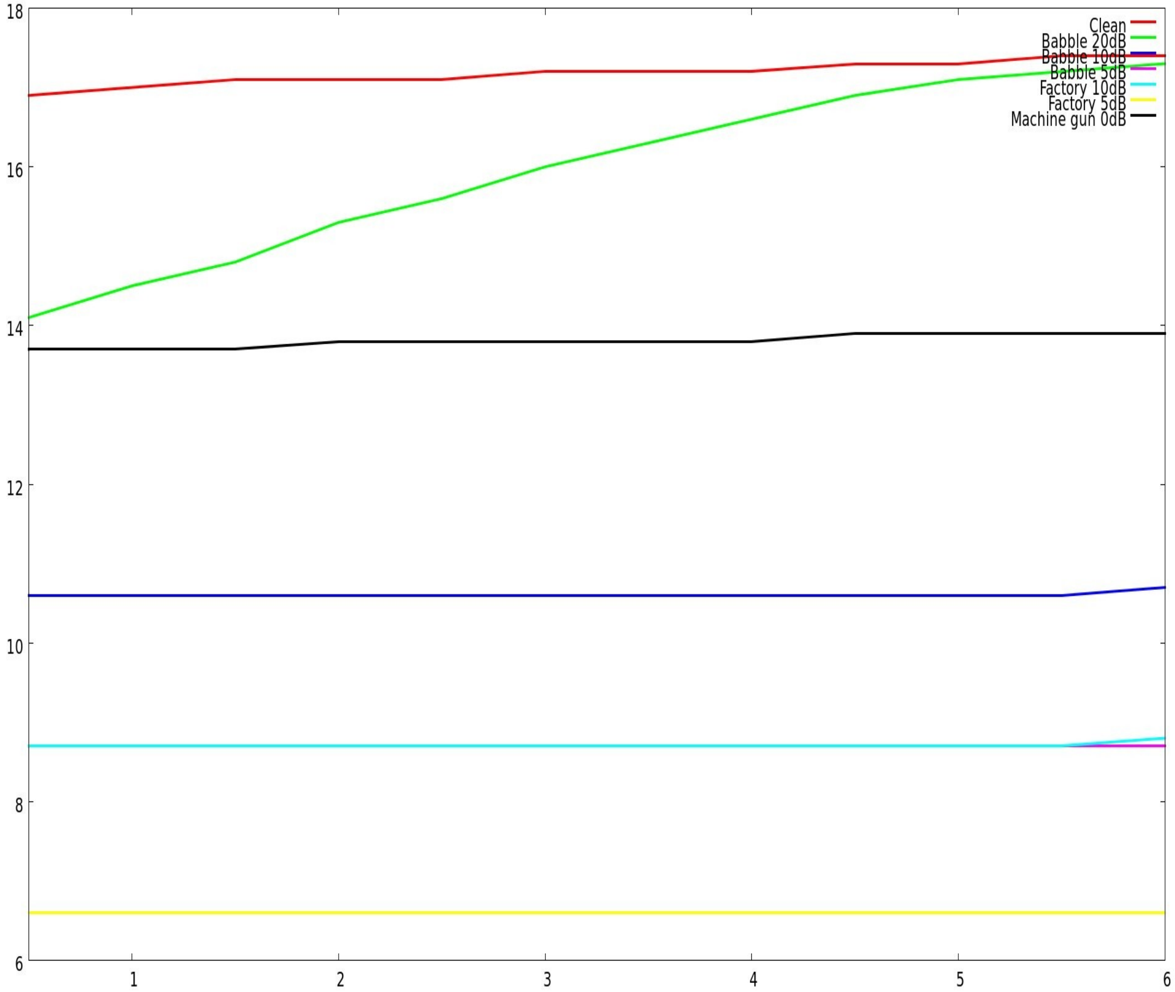
GlottHMM noise reduction (III)

MCD
NOISE RED. LIMIT = 4.5
NOISE RED. = 5 – 50



GlottHMM noise reduction (IV)

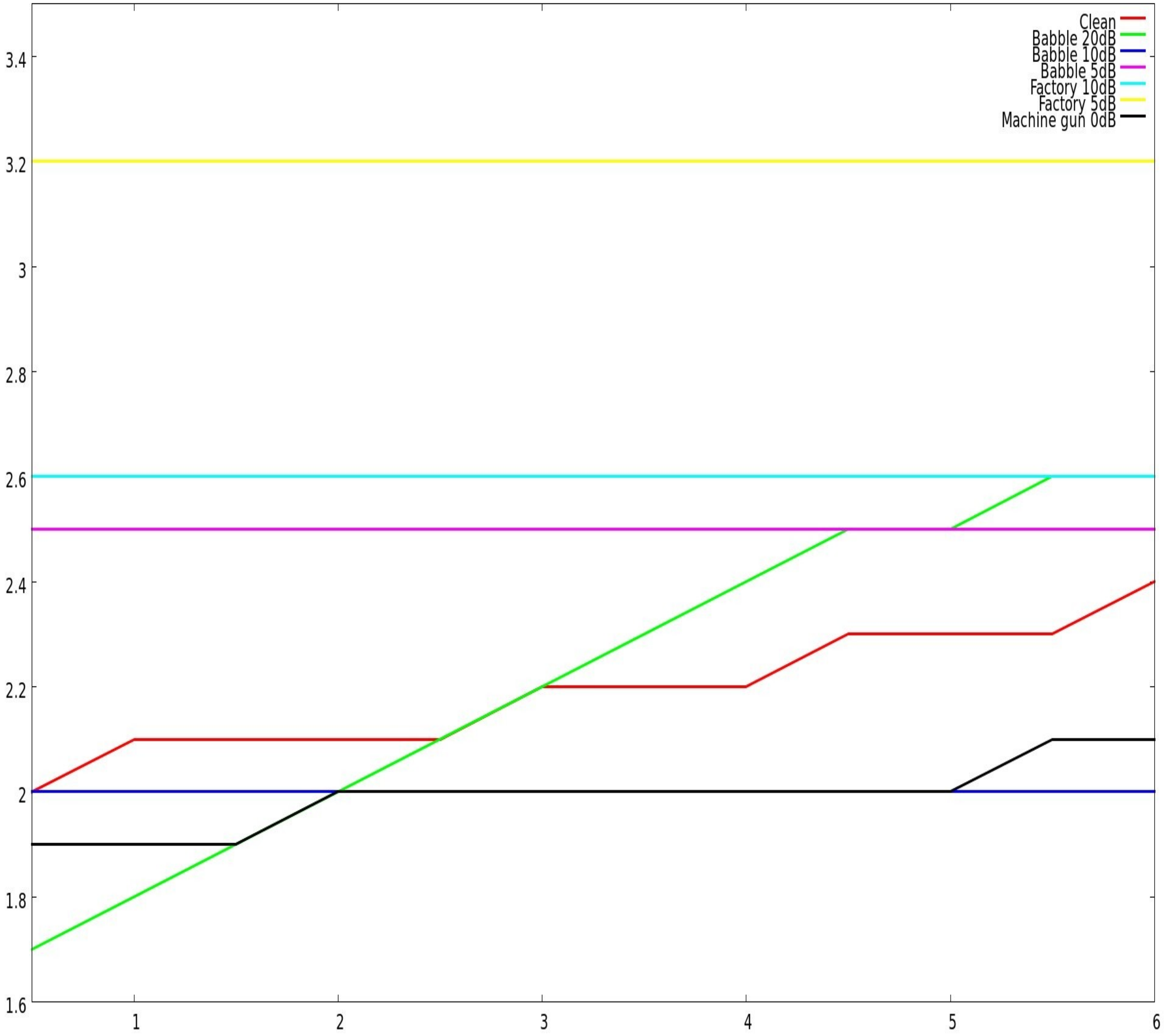
SNR
NOISE RED. LIMIT = 0.5 – 6
NOISE RED. = 35



Jose Moreno
02.04.2014

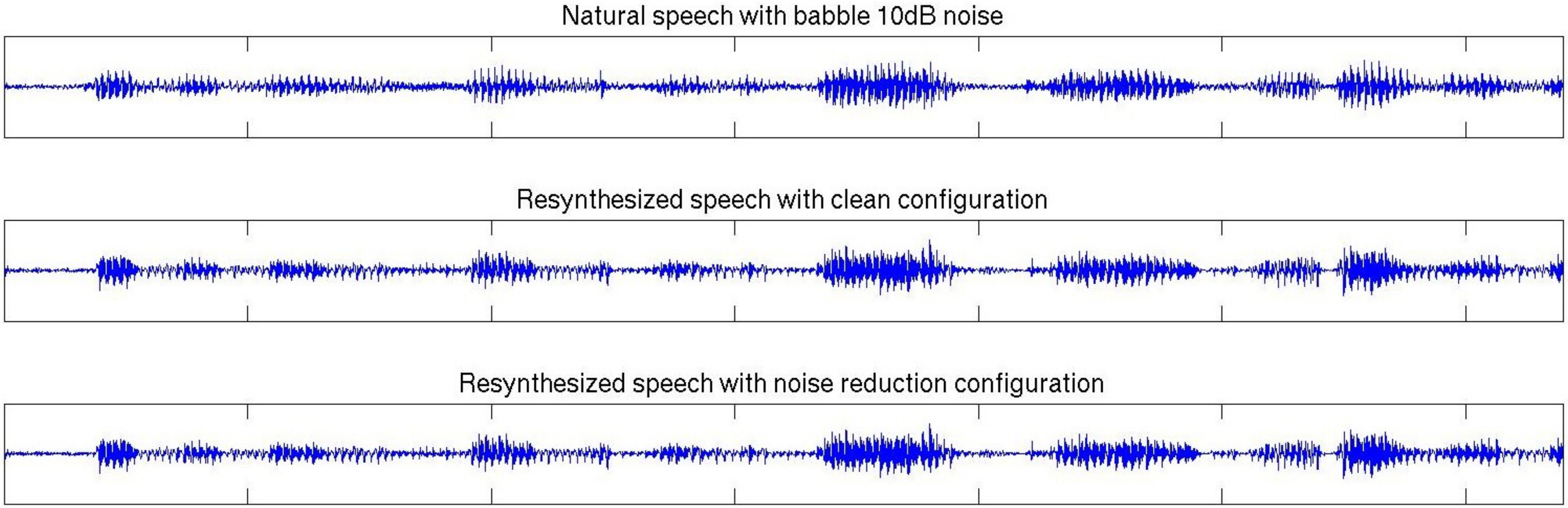
GlottHMM noise reduction (V)

MCD
NOISE RED. LIMIT = 0.5 – 6
NOISE RED. = 35



Jose Moreno
02.04.2014

GlottHMM noise reduction (VI)

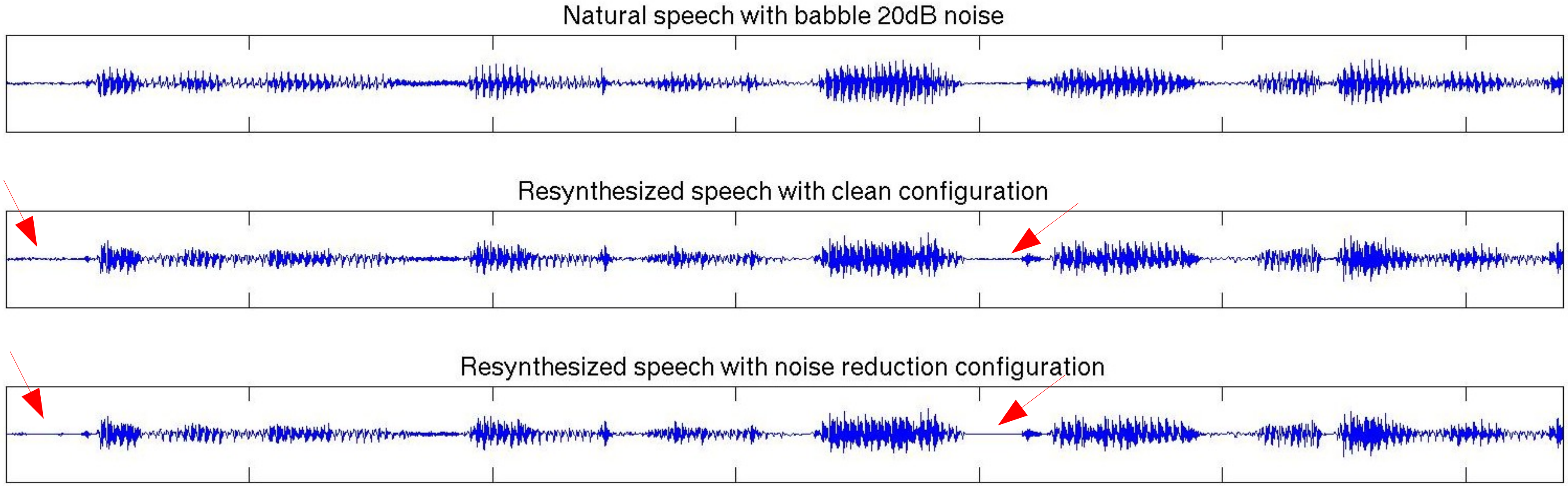


Natural speech with babble 10dB noise

Resynthesized speech with clean configuration

Resynthesized speech with noise reduction configuration

GlottHMM noise reduction (VII)

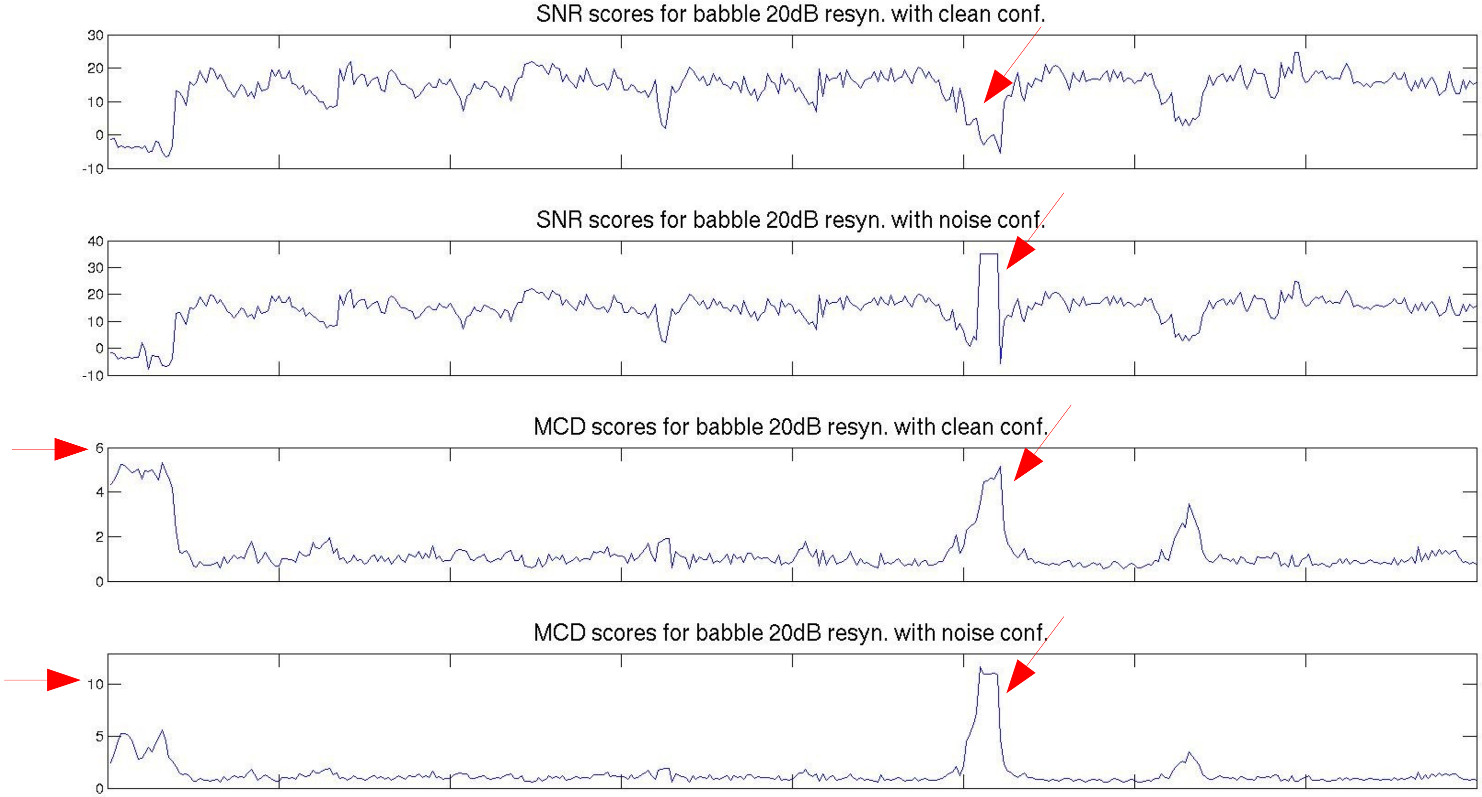


Natural speech with babble 20dB noise

Resynthesized speech with clean configuration

Resynthesized speech with noise reduction configuration

GlottHMM noise reduction (VIII)



SNR scores for babble 20dB resyn. with clean conf.

SNR scores for babble 20dB resyn. with noise conf.

MCD scores for babble 20dB resyn. with clean conf.

MCD scores for babble 20dB resyn. with noise conf.

Objective

Chapter 4

Evaluation

$$fwS = \frac{10}{M}\sum_{m=1}^{M}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{X(j,m)^2}{((X(j,m)-\hat{X}(j,m))^2}}{\sum_{j=1}^{K} W(j,m)}$$

$$MCD = \frac{1}{M}\sum_{m=1}^{M}\sqrt{2\sum_{d=0}^{D-1}(c(d,m)-\hat{c}(d,m))^2}$$

Subjective

- Listening test
  - AB test
    - Binomial test
      - Acumulative probability ($p \leq 0.05$)
    - GlottHMM vs STRAIGHT
    - Compare different GlottHMM configurations
- Mean Opinion Scores (MOS)
  - Naturalness
  - Similarity
  - Background quality

Objective measures

Chapter 5

Results

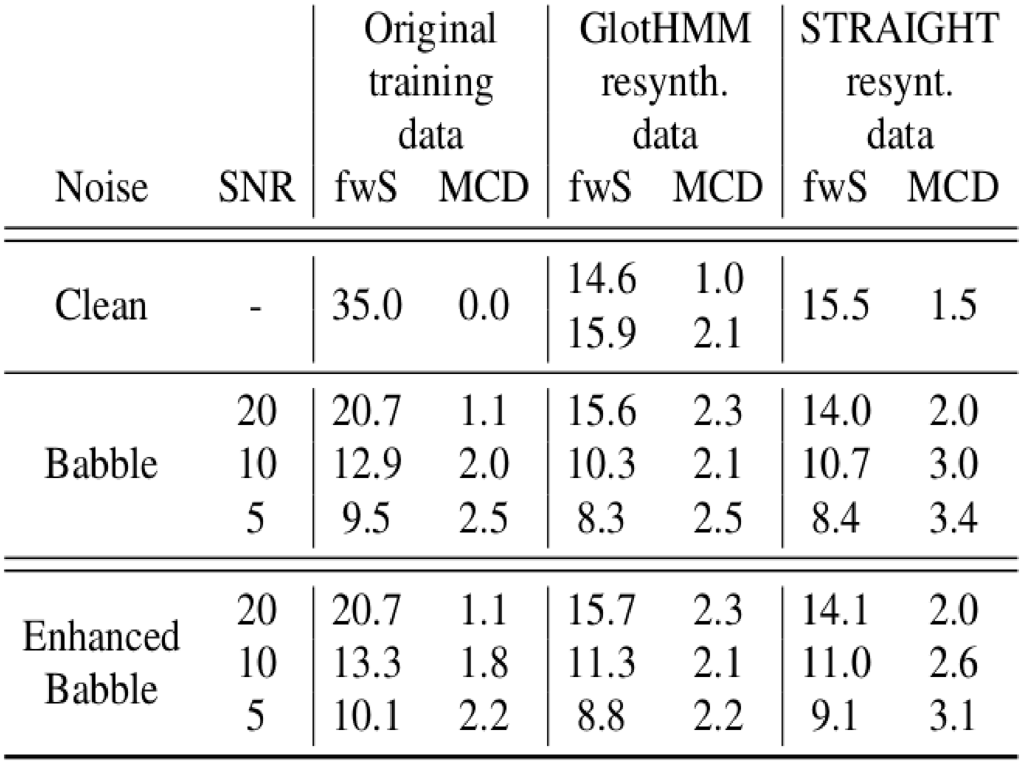
Noise	SNR	MCD
Clean	9.0	1.8
Babble 20	10.6	3.0
Babble 10	7.5	2.7
Babble 5	6.3	2.6
Factory 10	6.8	3.0
Factory 5	5.3	3.2
Machine gun	9.3	2.7
Enhanced babble 20	10.8	3.0
Enhanced babble 10	8.4	2.8
Enhanced babble 5	6.9	2.8
Ehanced factory 10	8.7	3.2
Enhanced factory 5	7	3.3

Objective measures (II)

Using external F0

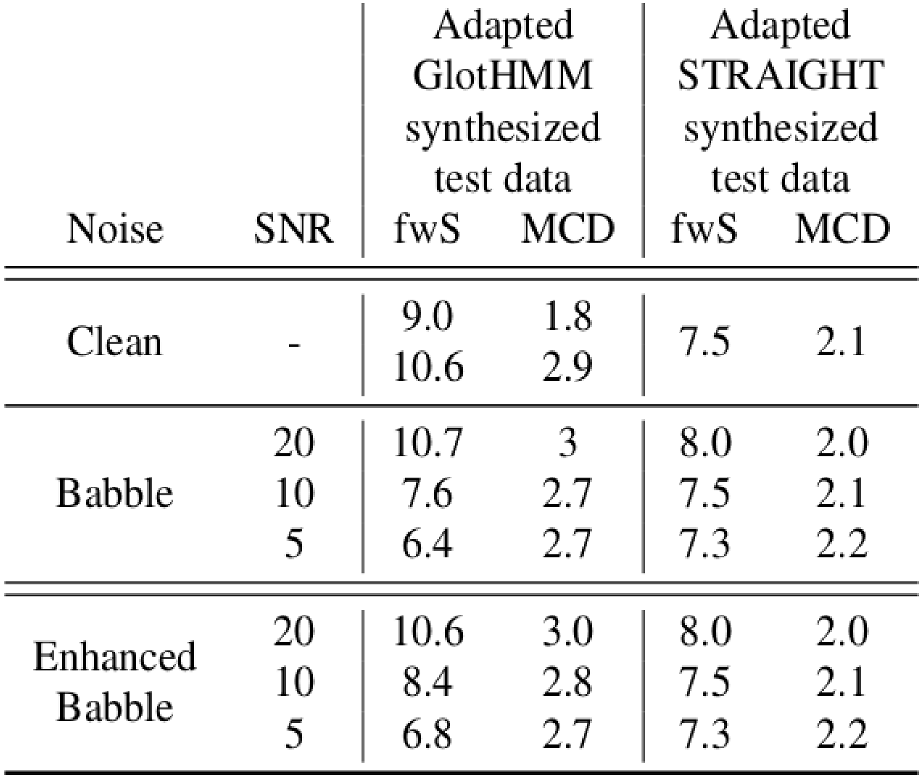
Noise	SNR	MCD
Babble 20	10.7	3.0
Babble 10	7.6	2.7
Babble 5	6.4	2.7
Factory 10	6.9	2.9
Factory 5	5.5	3.2
Machine gun	9.4	2.7
Enhanced babble 20	10.6	3.0
Enhanced babble 10	8.4	2.8
Enhanced babble 5	6.8	2.7
Ehanced factory 10	8.7	3.2
Enhanced factory 5	7.1	3.3

Objective measures: GlottHMM vs STRAIGHT (I)



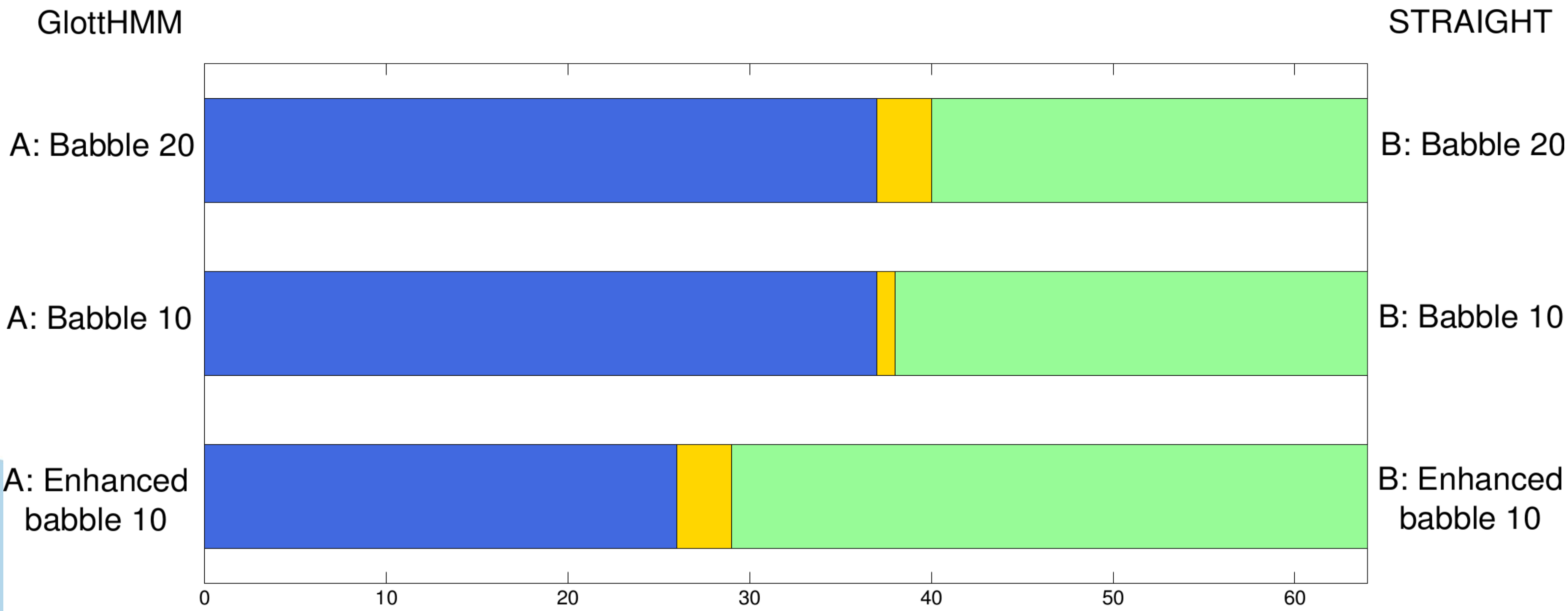
		Original training data		GlotHMM resynth. data		STRAIGHT resynt. data	
Noise	SNR	fwS	MCD	fwS	MCD	fwS	MCD
Clean	-	35.0	0.0	14.6 15.9	1.0 2.1	15.5	1.5
Babble	20	20.7	1.1	15.6	2.3	14.0	2.0
	10	12.9	2.0	10.3	2.1	10.7	3.0
	5	9.5	2.5	8.3	2.5	8.4	3.4
Enhanced Babble	20	20.7	1.1	15.7	2.3	14.1	2.0
	10	13.3	1.8	11.3	2.1	11.0	2.6
	5	10.1	2.2	8.8	2.2	9.1	3.1

Objective measures: GlottHMM vs STRAIGHT (II)



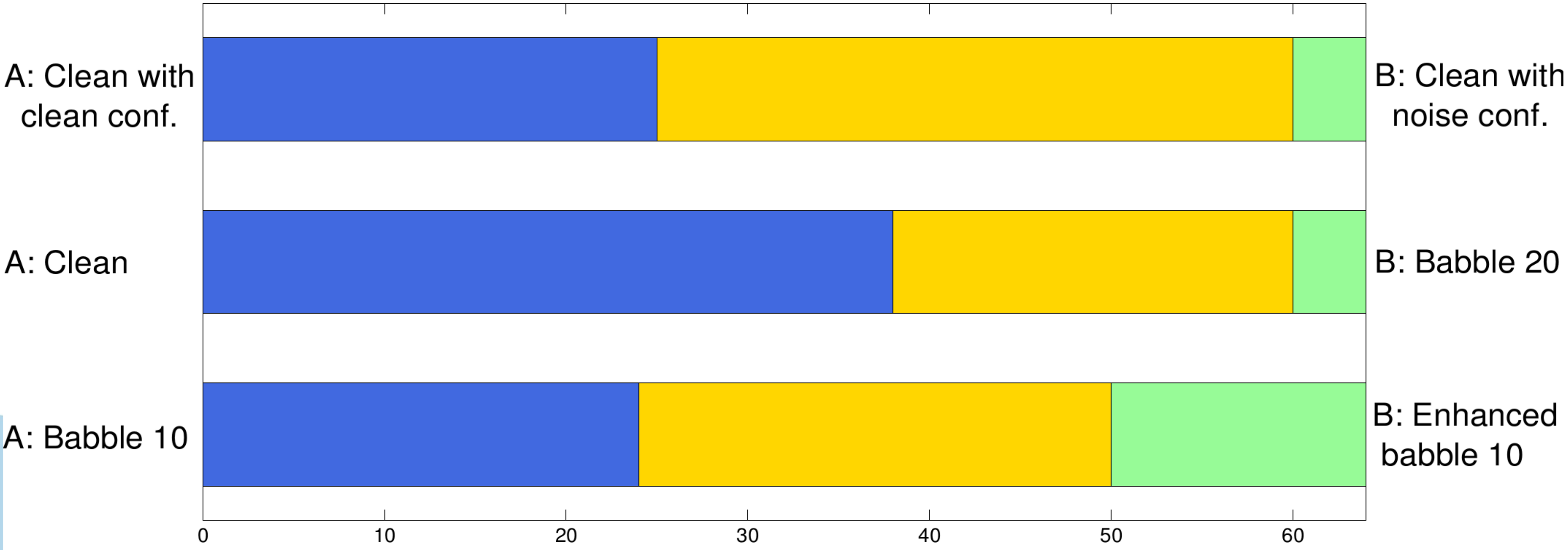
Noise	SNR	Adapted GlotHMM synthesized test data		Adapted STRAIGHT synthesized test data	
		fwS	MCD	fwS	MCD
Clean	-	9.0	1.8	7.5	2.1
		10.6	2.9		
Babble	20	10.7	3	8.0	2.0
	10	7.6	2.7	7.5	2.1
	5	6.4	2.7	7.3	2.2
Enhanced Babble	20	10.6	3.0	8.0	2.0
	10	8.4	2.8	7.5	2.1
	5	6.8	2.7	7.3	2.2

Subjective test (I)



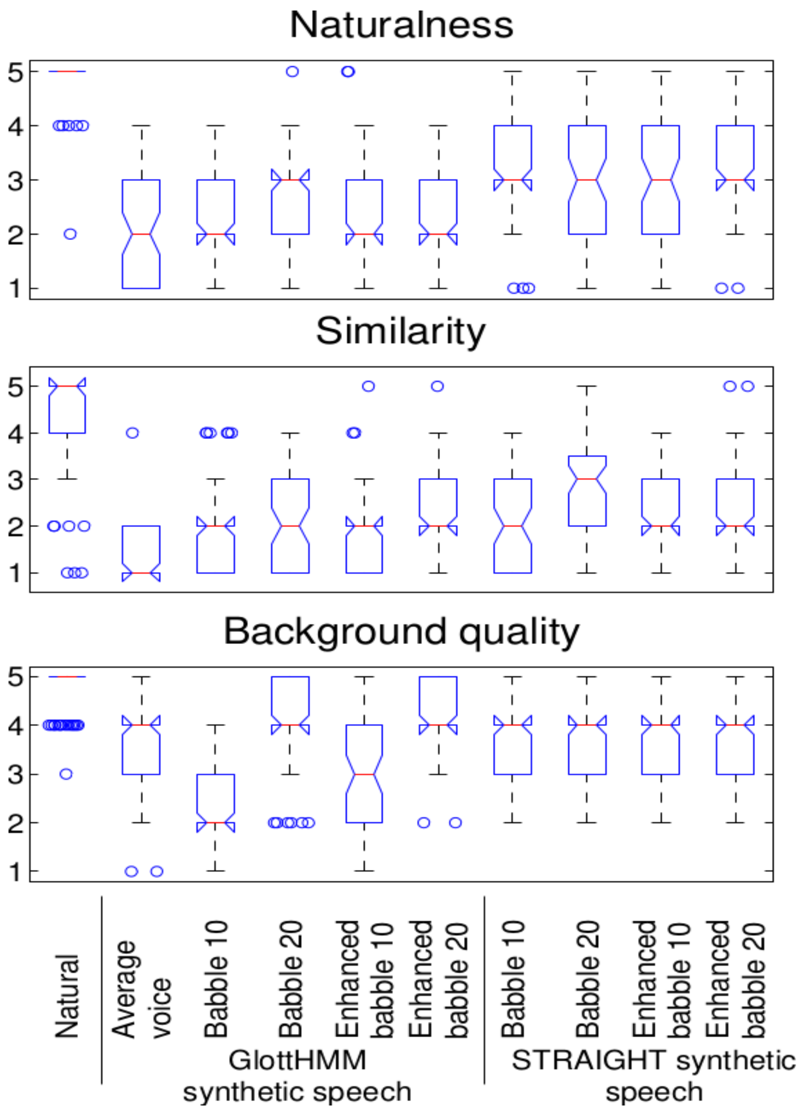
P values:
1st: P = 0.0516 (GlottHMM)
2nd: P = 0.0838 (GlottHMM)
3rd: P = 0.1292 (STRAIGHT)

Subjective test (II)



P values:
1st: P = 0.0043 (Sample A)
2nd: P << 0.05 (Sample A)
3rd: P = 0.1056 (Sample A)

Subjective test (III)



Jose Moreno
02.04.2014

Discussion

Chapter 6

Conclusions

- STRAIGHT slightly better MOS scores but no preference displayed
- Some factors can bias the test:
  - Stream dimension → different clustering thresholds / model realignment
- Contradictory objective measures

Conclusions

- Objective measures need further investigation
- STRAIGHT slighty higher rated in naturalness
- Very small differences in similarity
- GlottHMM more susceptible to degradation in more sever noise conditions
- Preference test show no significant differences

References

Chapter 7

References

- R. Karhila, U. Remes, and M. Kurimo, “Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments”, Signal Processing, IEEE Journal of, Selected Topics in vol. PP, no. 99, pp. 11, 2013
- J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm”, Audio, Speech, and Language Processing, IEEE Transactions on vol. 17, no. 1, pp. 6683, 2009.
- Reima Karhila

Questions?

Thanks!
It may be the moment for questions and samples

Jose Moreno
02.04.2014