Sinusoidal Models for Highly Intelligible Text-to-Speech Synthesis

… or …

Ingelligibility enhancement using a harmonic vocoder

Daniel Erro - derro@aholab.ehu.es

ikerbasque
Basque Foundation for Science

Bilbao, Spain

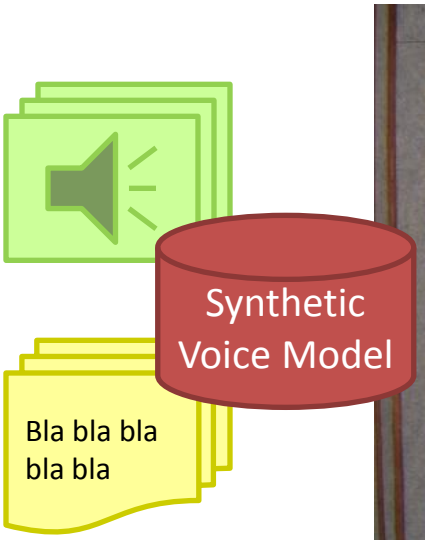
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Outline

- Introduction
- Modifications
- Experiments
- Conclusions

Introduction

- Recording a speech synthesis database

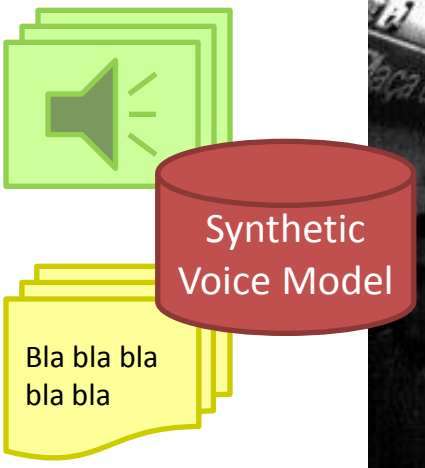


Introduction

- Recording a speech synthesis database

Introduction

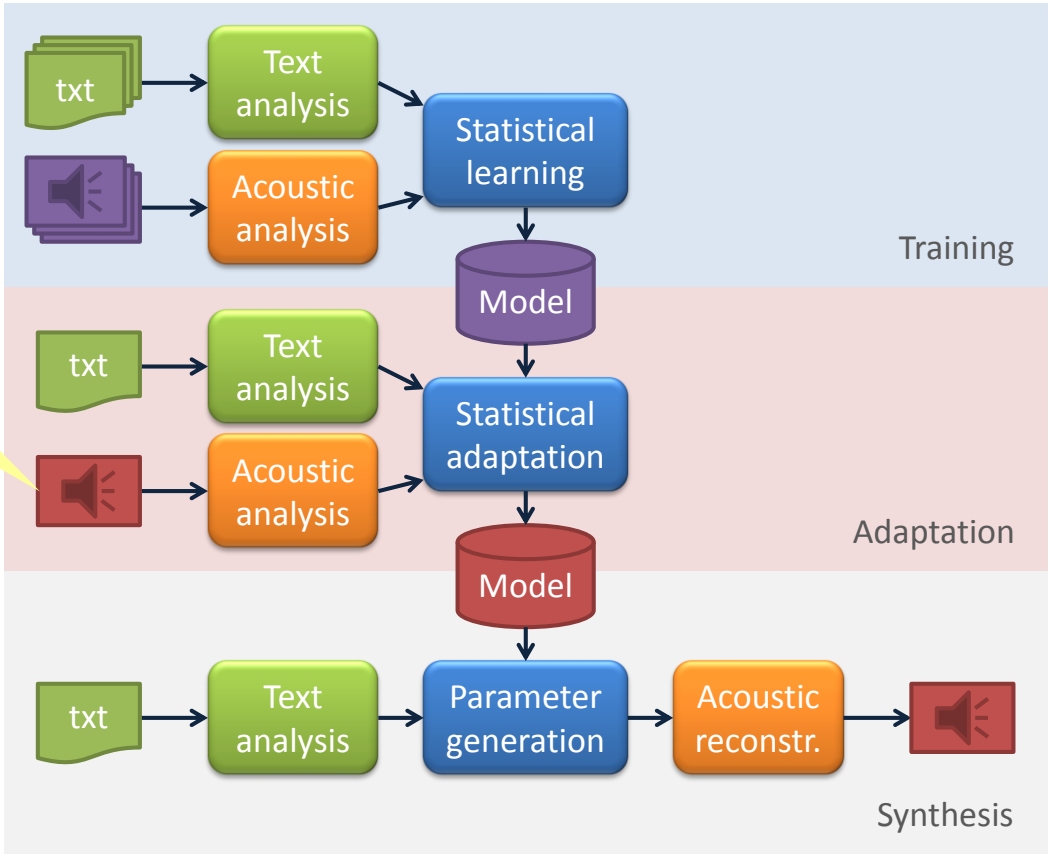
- Recording a speech synthesis database



Introduction

- Possible solution: speaker adaptation

Ad-hoc recordings (€) of Lombard speech, different noise types, not applicable to other speakers, etc

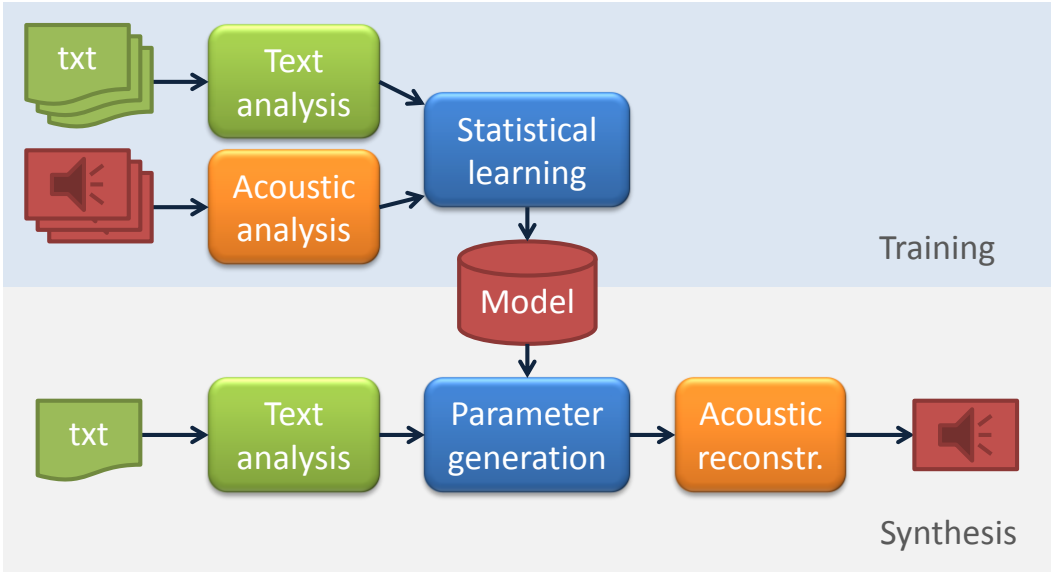


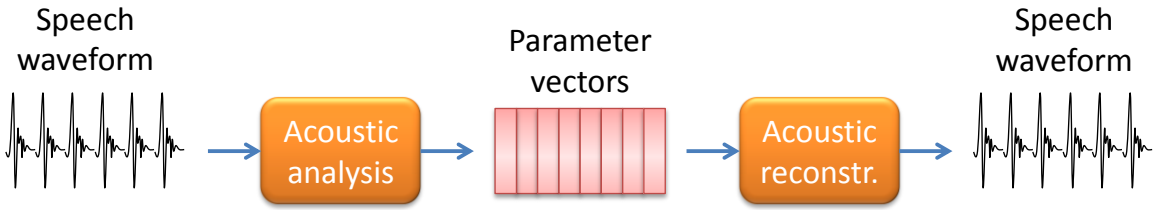
Introduction

Previously…

Introduction

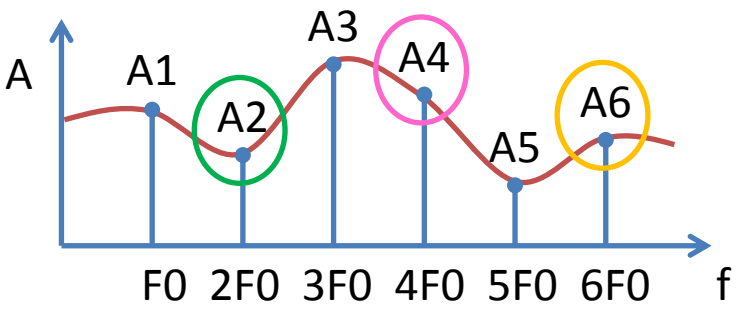
- Role of sinusoidal models in TTS:
  - Statistical parametric speech synthesis: **vocoders**

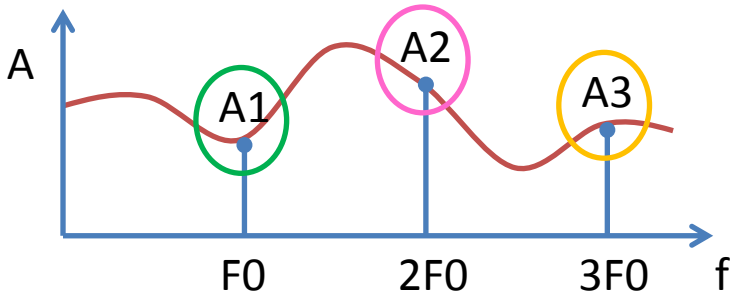




Introduction

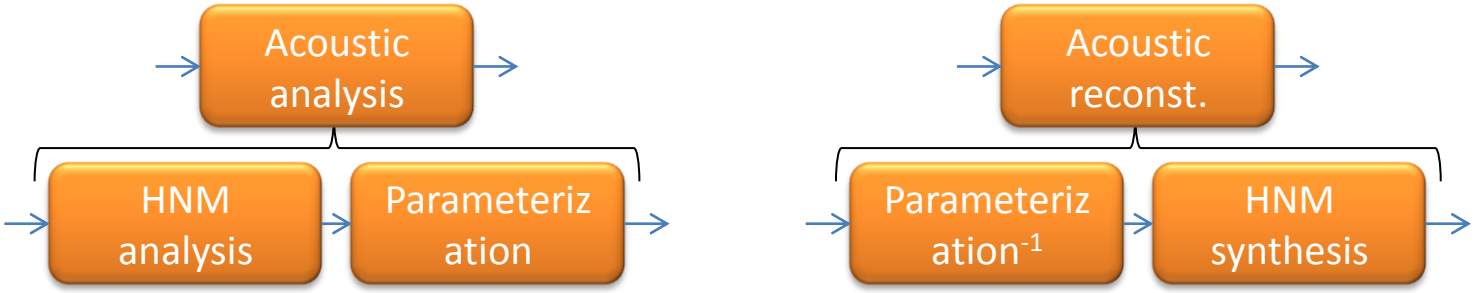
- Sinusoids(+noise) based vocoder?
  - HQ resynthesis and modification, but…
  - Variable dimension
  - Not very tractable, complicated dependencies





Introduction

- Sinusoids(+noise) based vocoder?
  - Use them as an intermediate stage between waveforms and parameters
    - Sinusoidal frequencies → logF0
    - Sinusoidal amplitudes → MCEP, MGC, LSF…
    - Sinusoidal phases → RPS, PD… or nothing!
    - Noise → HNR, MVF…



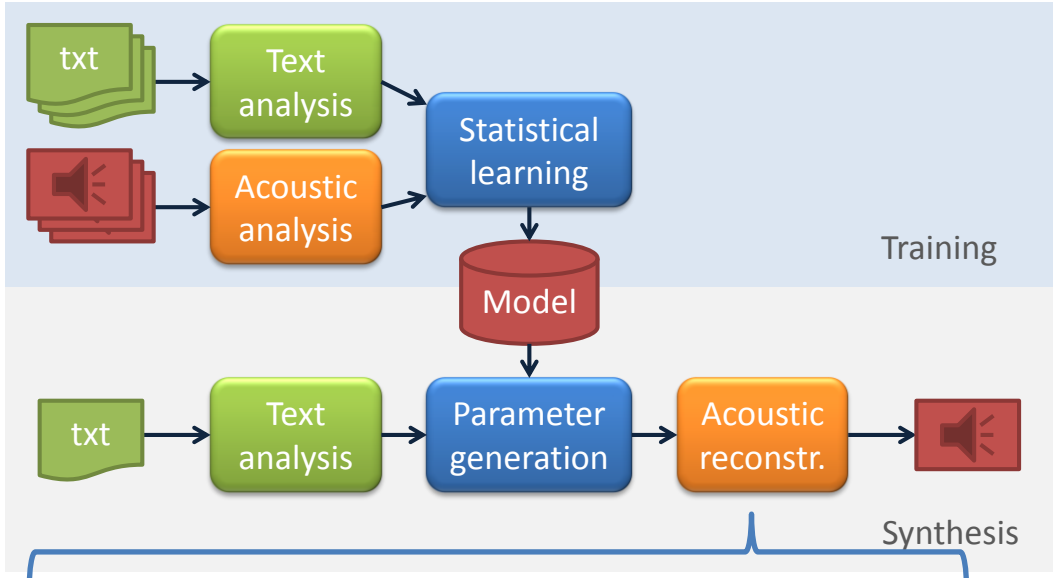
Introduction

- Sinusoids(+noise) based vocoder?
  - Use them as an intermediate stage between waveforms and parameters
    - Sinusoidal frequencies → logF0
    - Sinusoidal amplitudes → MCEP, MGC, LSF…
    - Sinusoidal phases → RPS, PD… or nothing!
    - Noise → HNR, MVF…
  - Enables intermediate modifications

Introduction

- Possible solution (ii): train the voice normally, then modify it during synthesis

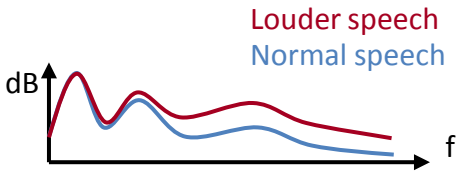
Cheap, fast, easy to develop, works well for natural speech, and goes beyond Lombard speech



Flexible signal processing framework and some expert knowledge

Introduction

- What should we modify? (Cooke et al., 2014)
  - Studies about Lombard effect and clear vs spontaneous speech
    - Spectral tilt variations
    - Smaller contrast between vowels and consonants/transients
    - Higher F0 mean and range (?)
    - Lower speaking rate (?)
    - Expanded vowel space, more articulated speech
    - …



It depends on the voice/database, the environment…

Introduction

- Noise-dependent?
  - Adapt the modification depth to noise level to preserve quality
  - Listening and predicting future noise → more sophisticated algorithms and computational requirements
- Noise-independent?
  - It has been shown to work well when the goal is intelligibility
  - Not listening, not predicting → easy to implement, rapid

Outline

- ~~Introduction~~
- Modifications
- Experiments
- Conclusions

Modifications

- Slightly modified version of the vocoder
  - In noise, with intelligibility as goal, subtle quality improvements are no longer necessary
    - No explicit MVF analysis (remember hands-on session!)
  - Energy-related operations
    - Harmonic model, without noise

$$e \propto A_1^2 + A_2^2 + A_3^2 + \cdots$$

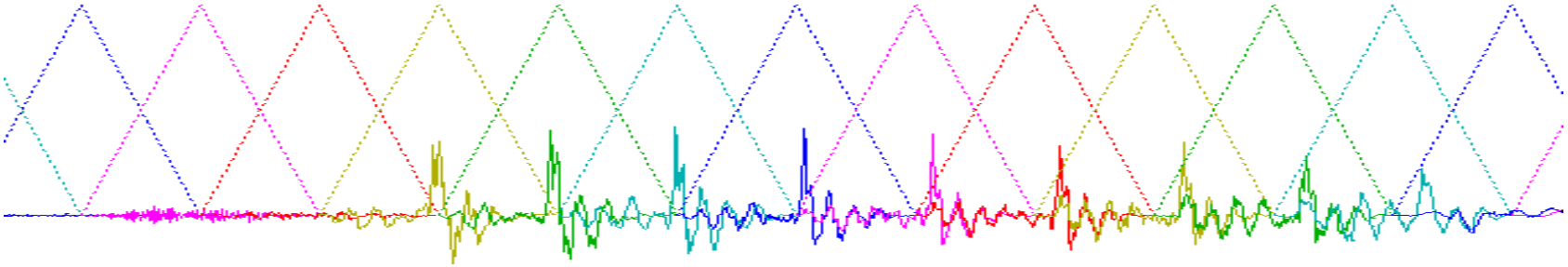
Modifications

- Slightly modified version of the vocoder

Triang win

Frame k

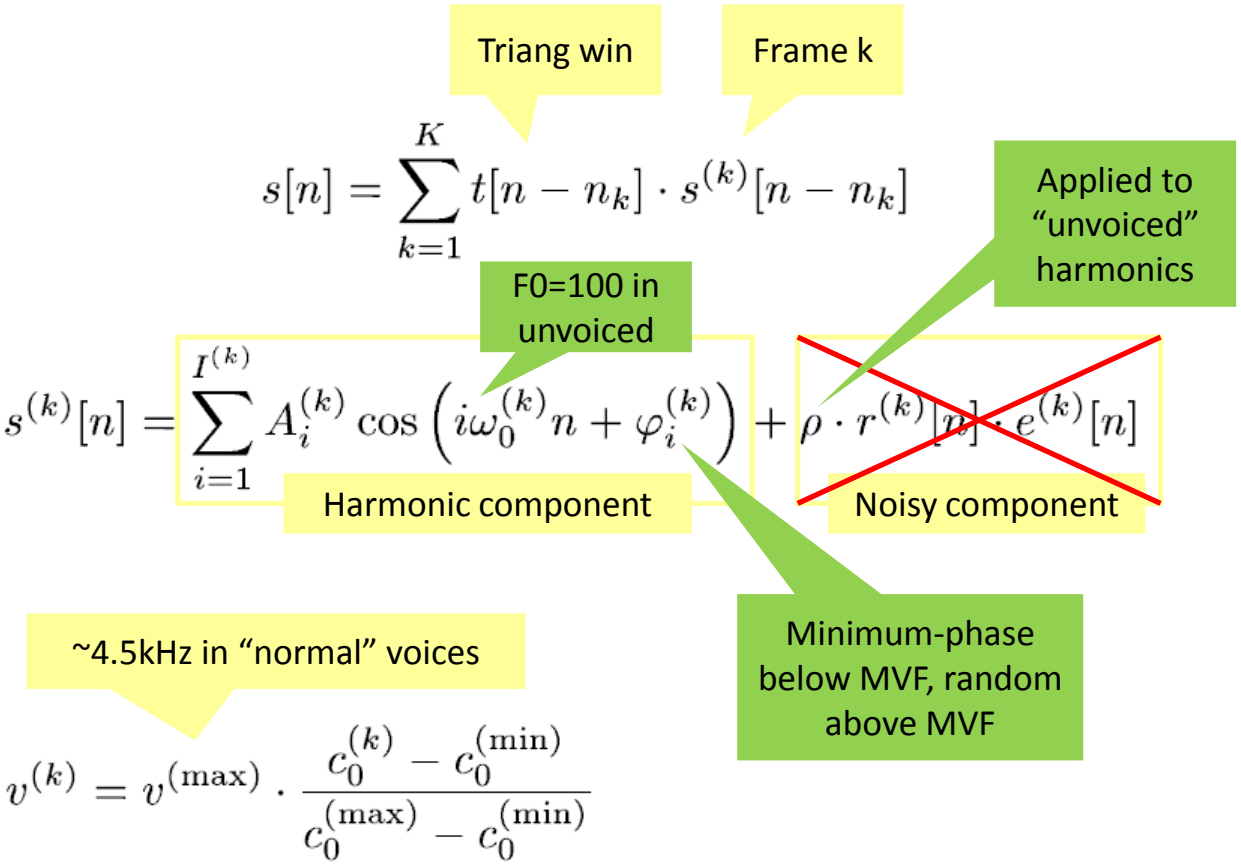
$$s[n] = \sum_{k=1}^{K} t[n - n_k] \cdot s^{(k)}[n - n_k]$$



Not pitch-synchronous but constant frame length of 10ms (from nk-5ms to nk+5ms)

Modifications

- Slightly modified version of the vocoder

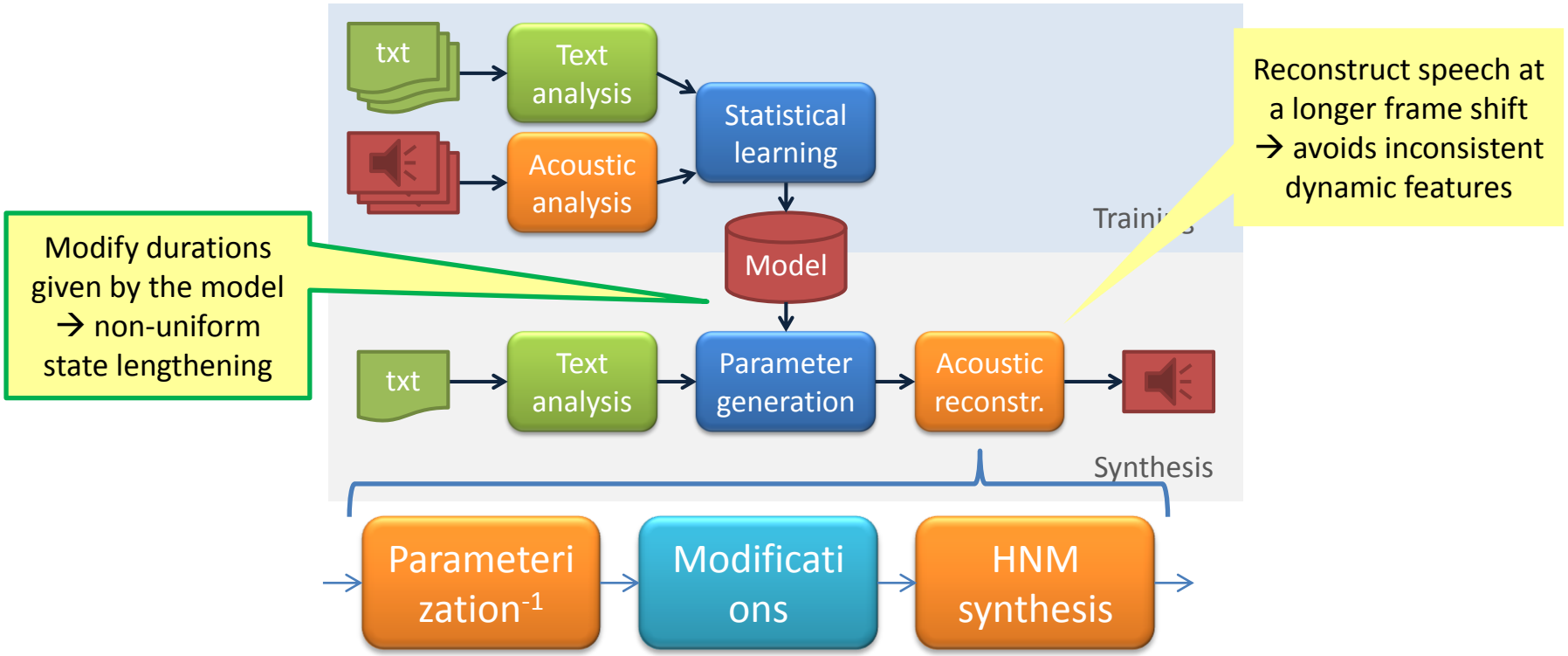


Modifications

- Modification #1: uniform lengthening
  - Clear speech is slower than casual speech (longer pauses?)
  - It has been shown to make synthetic speech more intelligible in various types of noise (Valentini-Botinhao, 2014)

Modifications

- Modification #1: uniform lengthening



Modifications

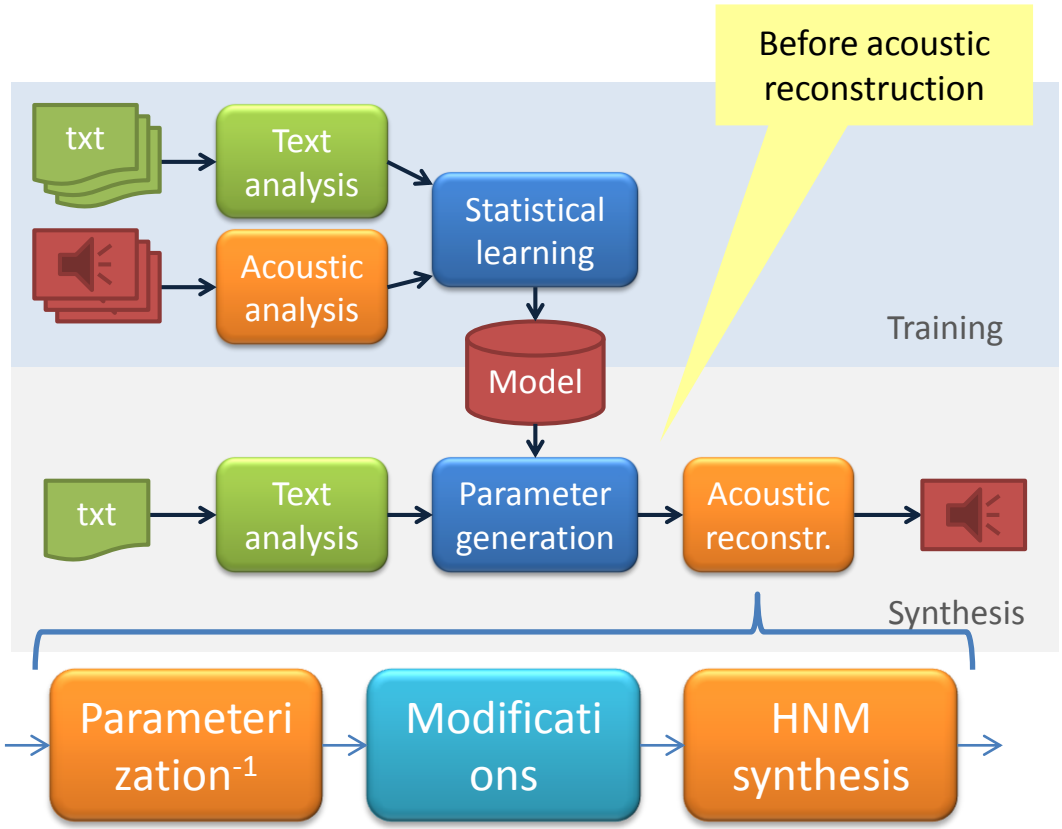
- Modification #1: uniform lengthening
  - Calculate phone durations from models
  - Multiply these phone durations by factor 1.2
  - Force new durations at input so that the states within each phone are lengthened in a non-uniform way

Modifications

- Modification #2: mean F0 level and range
  - Some people speak in a larger F0, some do not
  - It is known that F0 modifications do not improve intelligibility by themselves, BUT…
  - Louder speech → higher sub-glottal pressure → more rapid glottal fold vibration → higher F0!

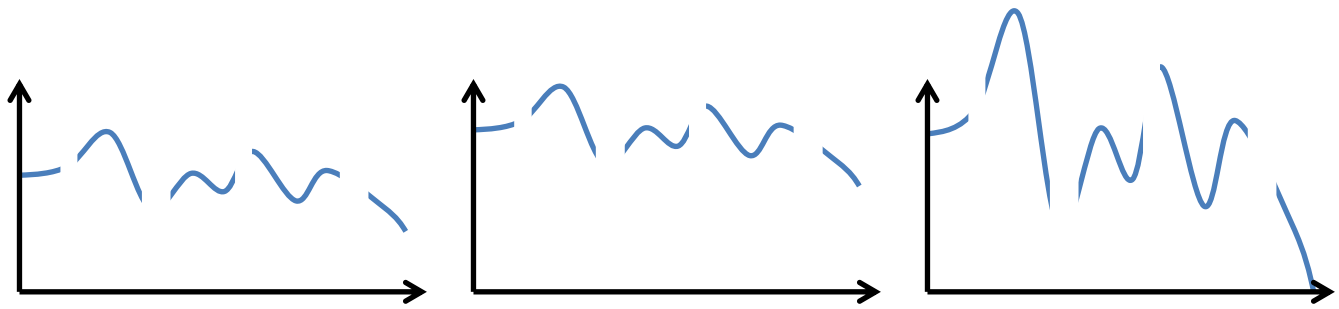
Modifications

- Modification #2: mean F0 level and range



Modifications

- Modification #2: mean F0 level and range
  - Generate parameters as usual
  - Sum log(1.2) to logF0 trajectory
  - Multiply utterance-level variance by1.5



Modifications

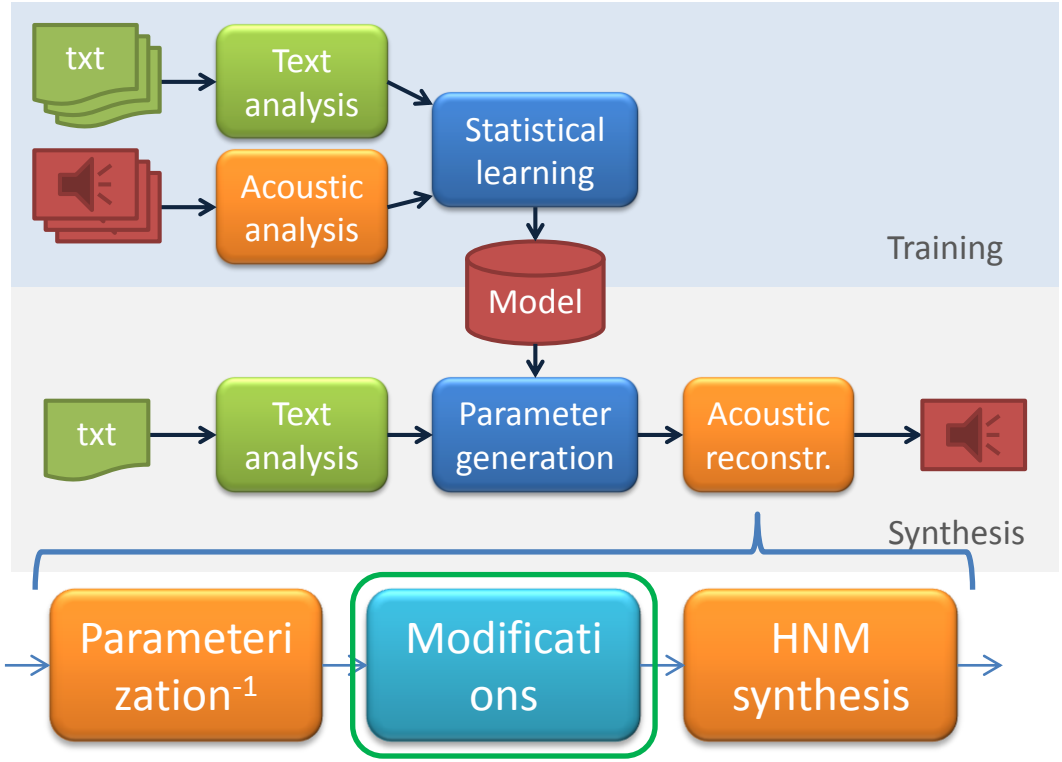
- Modification #3: redistribute energy over time
  - Reduce contrast betwen vowels and consonants without altering global SNR → steal E from “rich” frames and give it to “poor” frames

Modifications

- Modification #3: redistribute energy over time
  - Reduce contrast betwen vowels and consonants without altering global SNR → steal E from “rich” frames and give it to “poor” frames
  - Audio engineering solution: Dynamic Range Compression (DRC)

Modifications

- Modification #3: redistribute energy over time



Modifications

- Modification #3: redistribute energy over time

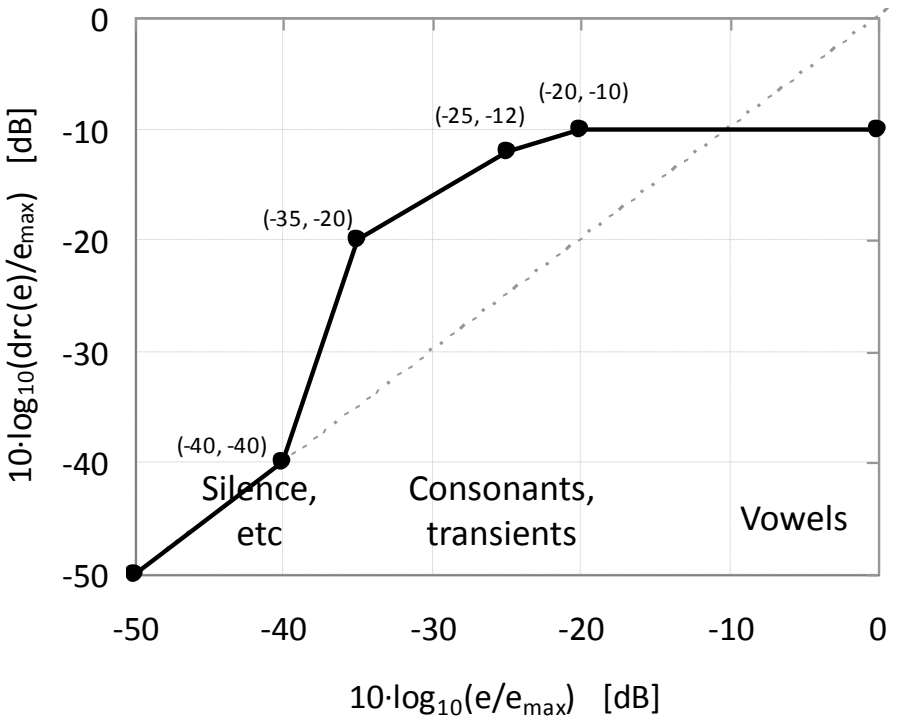
Local energy (Parseval)

$$e^{(k)} = \sum_{i=1}^{I^{(k)}} A_i^{(k)2}$$

$$\hat{e}^{(k)} = \gamma \cdot \mathrm{drc}\left(e^{(k)}\right)$$

$$\gamma = \frac{\sum_{k=1}^{K} e^{(k)}}{\sum_{k=1}^{K} \mathrm{drc}\left(e^{(k)}\right)}$$

Preserve total signal power

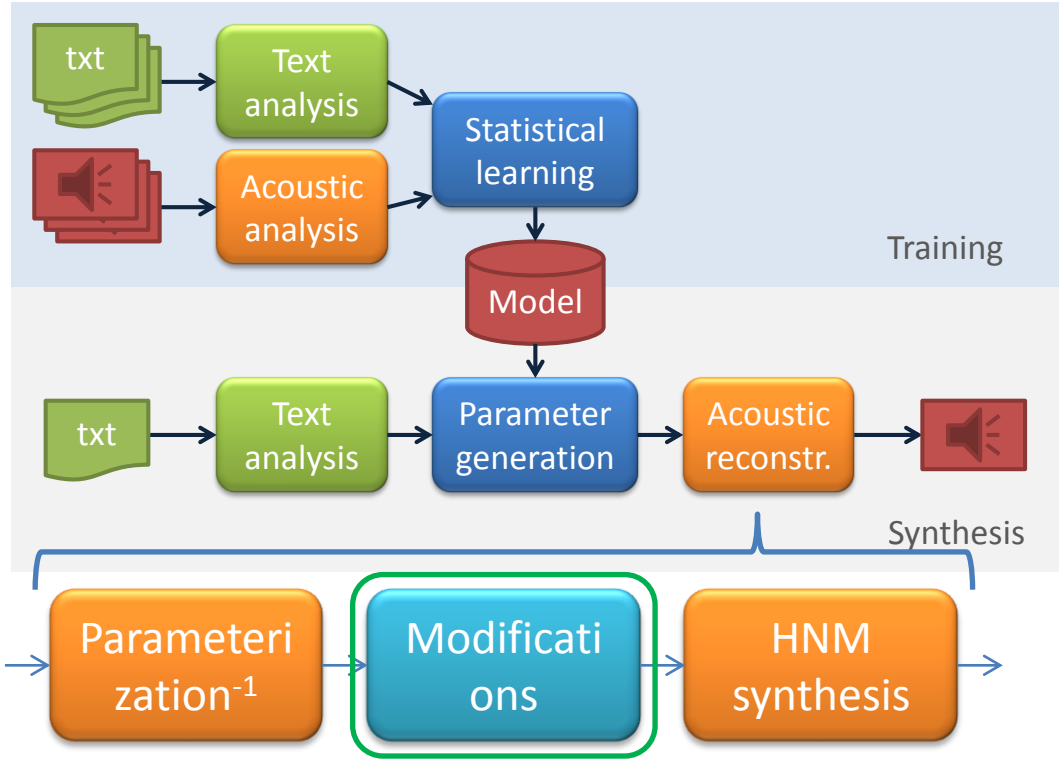


Modifications

- Modification #4: formant sharpening
  - Clear speech: narrower formants
  - Successful for hearing impaired
  - Unclear improvements in this context
  - Easy to implement in A domain

Modifications

- Modification #4: formant sharpening



Modifications

- Modification #4: formant sharpening

Multiplicative factor for each Ai

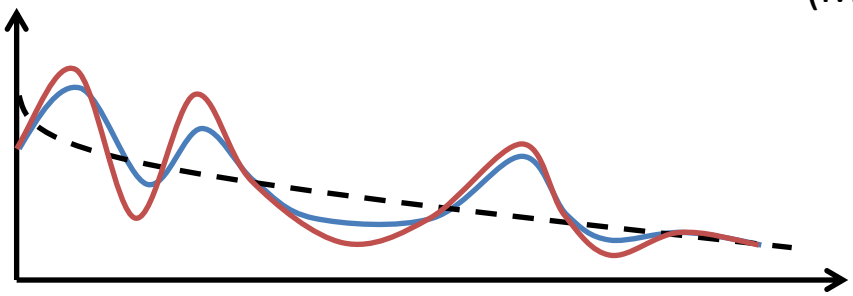
~0.25

$$P_i = \left[A_i \sqrt{\frac{I \cdot (r_0^2 - 2r_1 r_0 \cos(i\omega_0) + r_1^2)}{r_0 (r_0^2 - r_1^2)}} \right]^{\beta p_v}$$

Local probability of voicing (MSD weight)

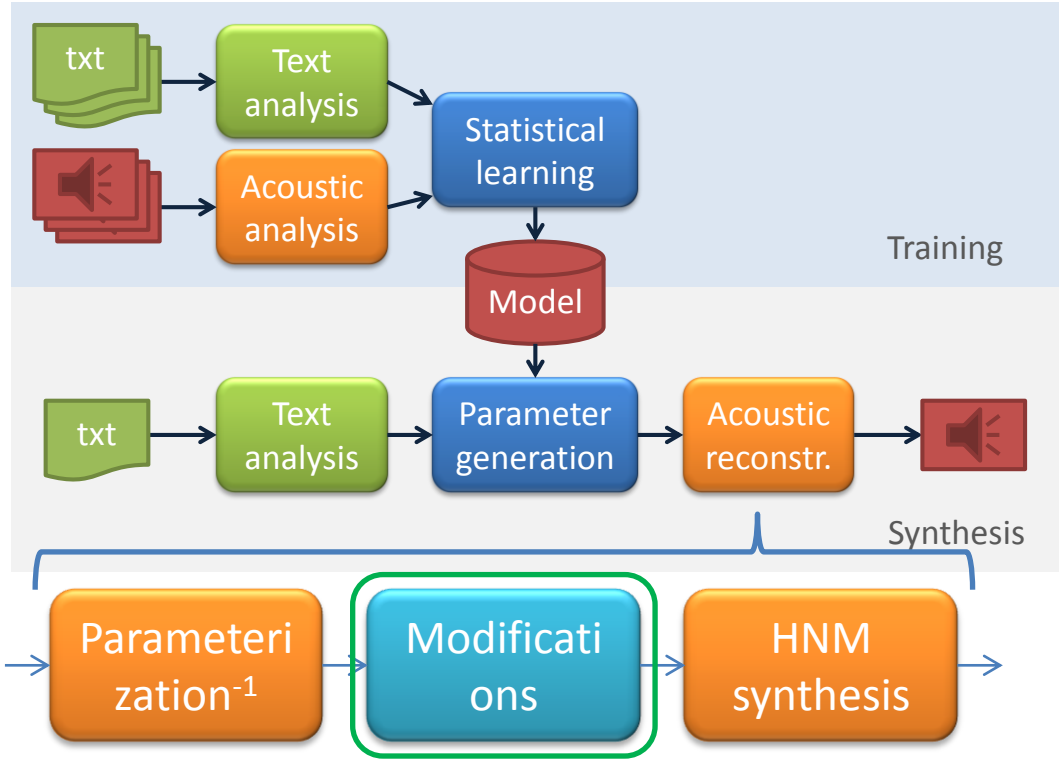
$$r_n = \sum_{m=1}^{I} A_m^2 \cos(m\omega_0 n)$$

(McAulay & Quatieri, 1995)



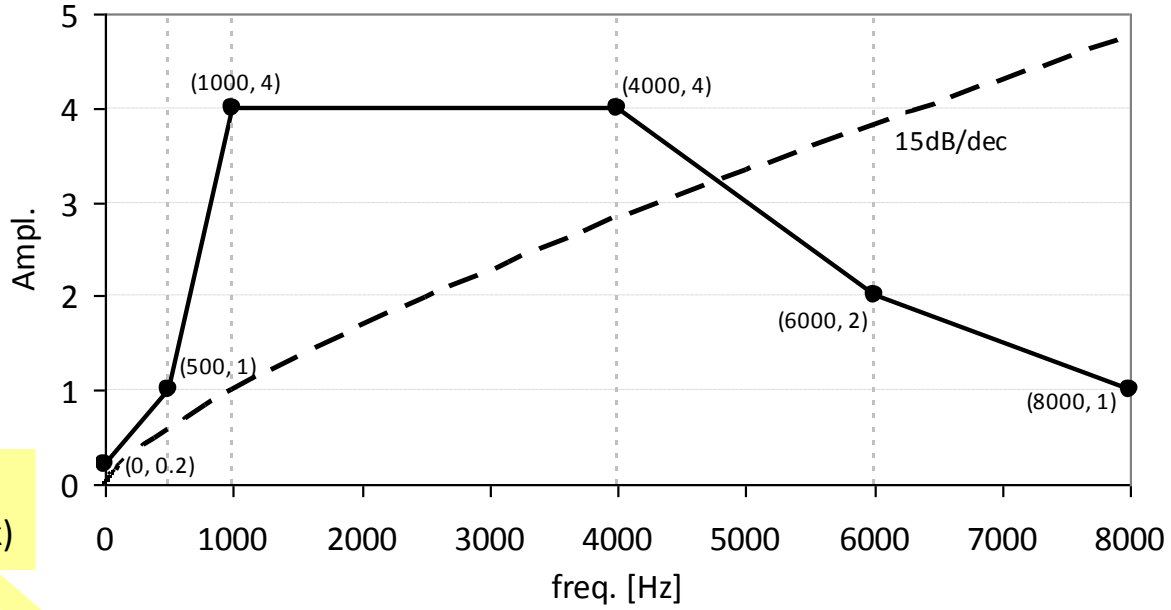
Modifications

- Modification #5: amplify mid-frequencies



Modifications

- Modification #5: mid-freq enhancement



Preserve energy at (k)

$$\hat{A}_i^{(k)} = \eta \cdot P_i^{(k)} H_i^{(k)} A_i^{(k)}$$

Formant sharpening

Mid-f filter

$$\eta = \sqrt{\frac{\hat{e}^{(k)}}{\sum_{m=1}^{I^{(k)}} \left(P_m^{(k)} H_m^{(k)} A_m^{(k)} \right)^2}}$$

Energy after DRC

Modifications

- Modification #1: uniform lengthening
- Modification #2: mean F0 level and range (F0)
- Modification #3: redistribute energy over time (DRC)
- Modification #4: formant sharpening (PF)
- Modification #5: mid-freq enhancement (SS)

Outline

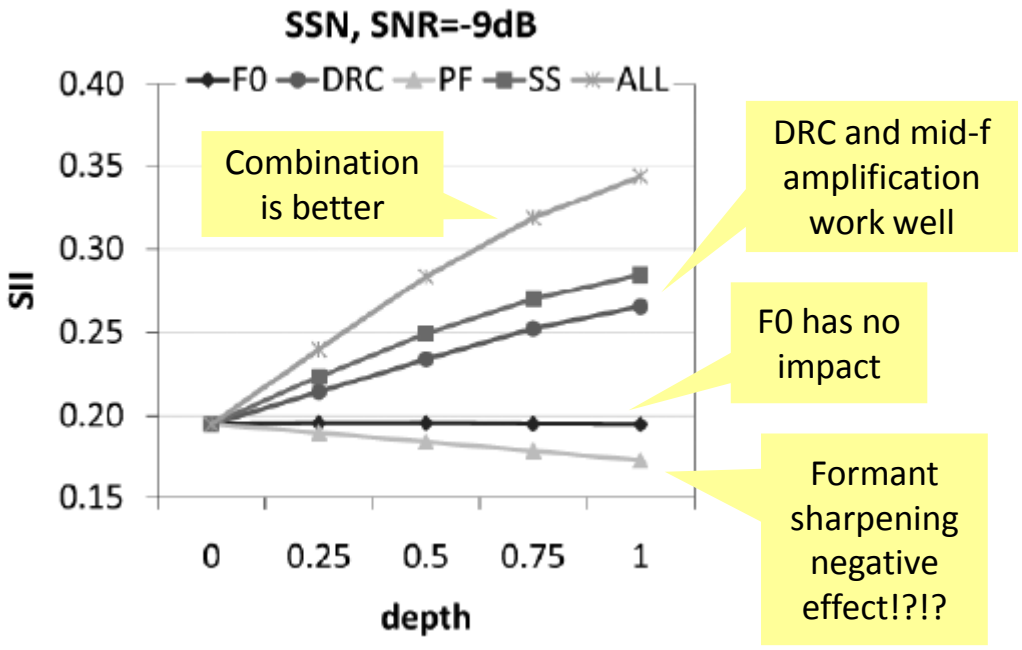
- ~~Introduction~~
- ~~Modifications~~
- Experiments
- Conclusions

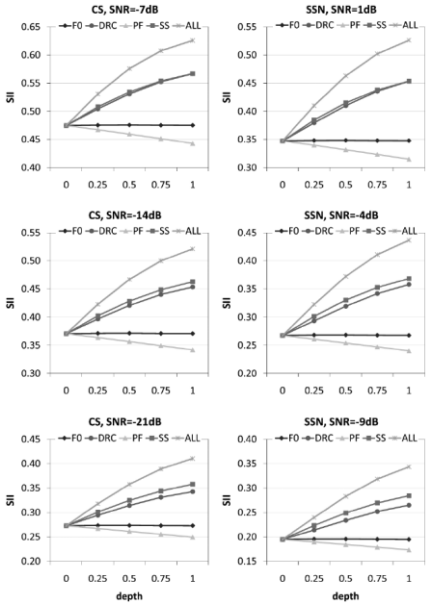
Experiments

- Calculate eSII, ANSI S3.5-1997, 0.8 correlation with subjective scores (Rhebergen & Versfeld, 2005)
  - Competing speaker, SNR = -7, -14, -21dB
  - Speech-shaped noise, SNR = 1, -4, -9dB

Experiments

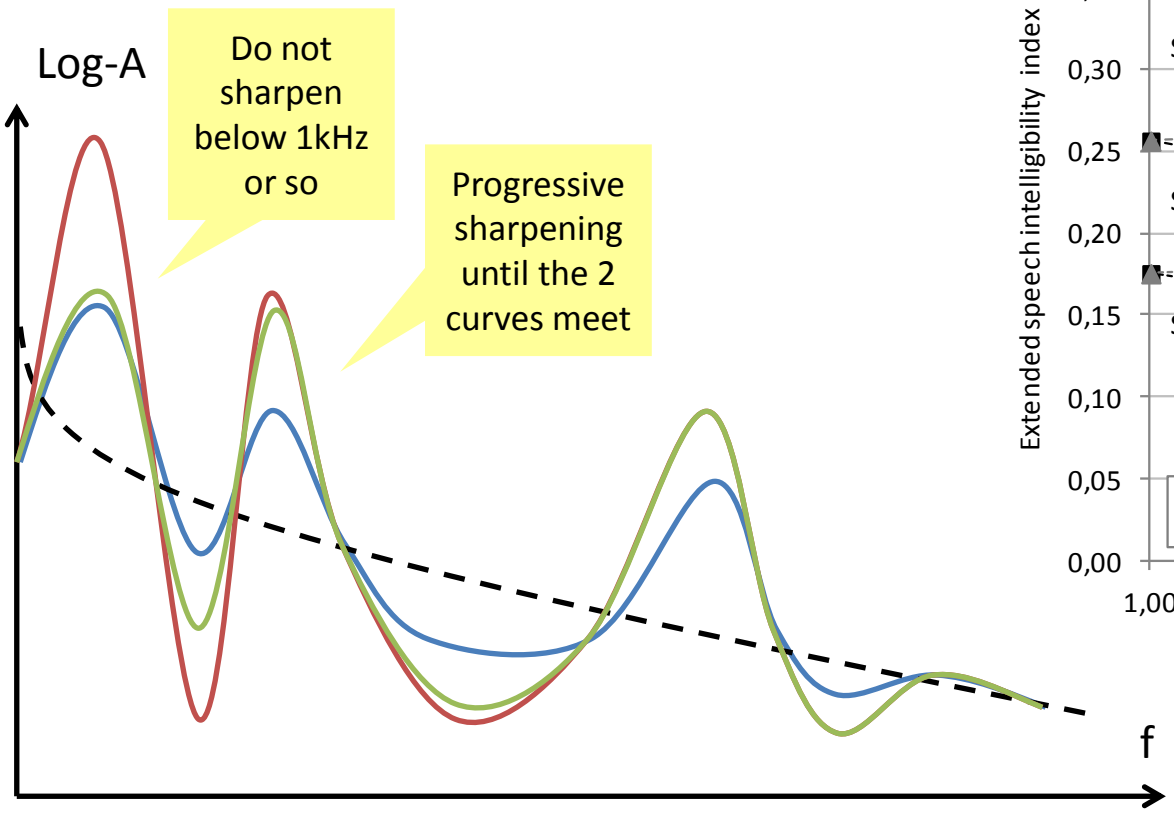
- Calculate eSII, ANSI S3.5-1997, 0.8 correlation with subjective scores (Rhebergen & Versfeld, 2005)

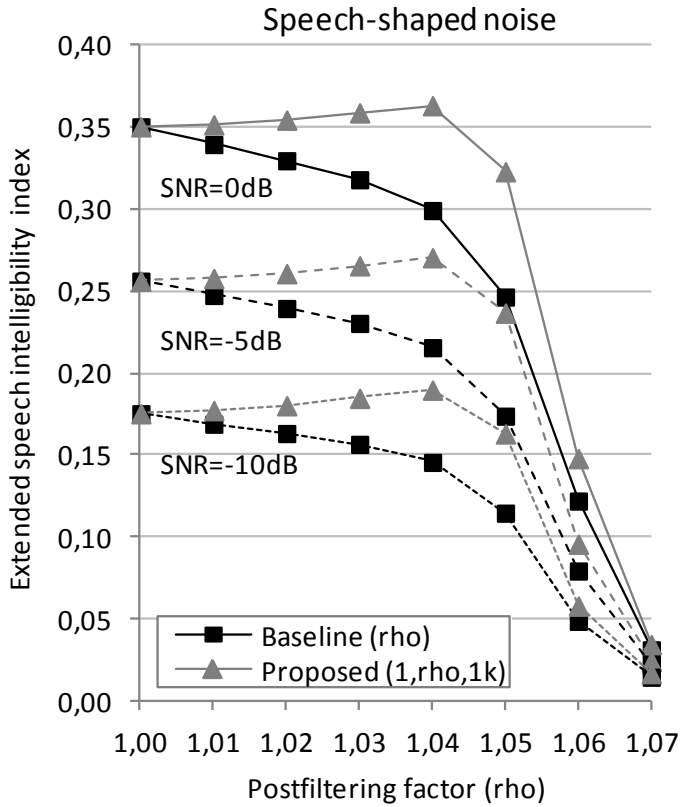




Experiments

- Parenthesis



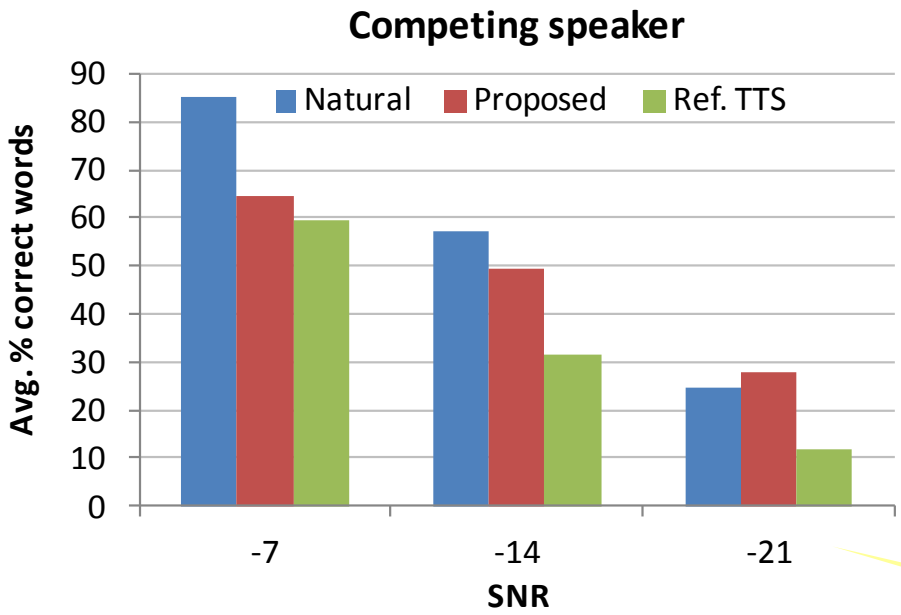


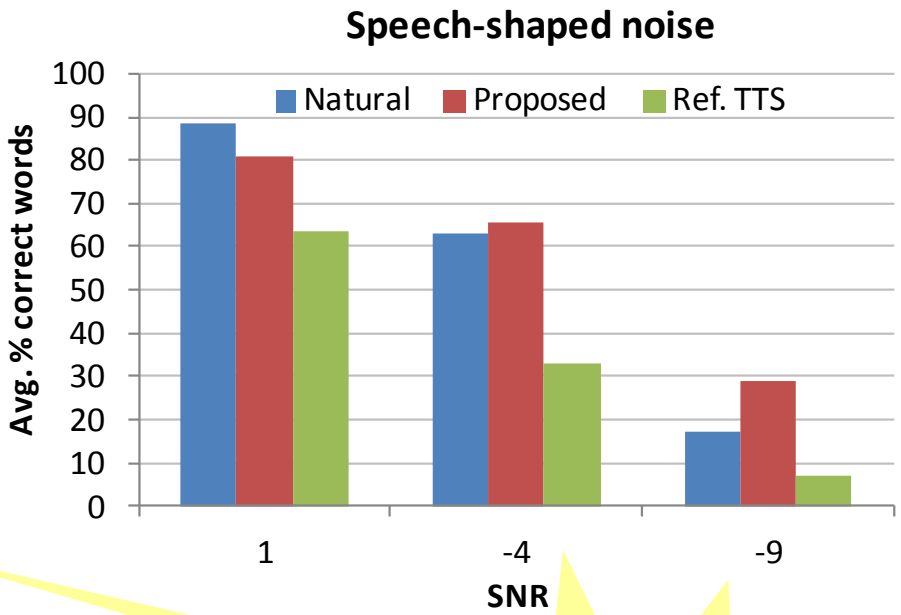
Experiments

- Hurricane Challenge (Cooke et al., 2013)
  - 175 native listeners, Univ. Edinburgh
  - "Listen once and type what you hear"
  - Avg % correct words excluding very short ones (a, the, in, to, on, is, and, of, for, at)
  - 15 natural speech enhancement systems, 5 TTS
  - 2 types of noise: speech-shaped, competing spkr
  - TTS: 2863 short sentences for training, 180 for test

Experiments

- Results of Hurricane Challenge



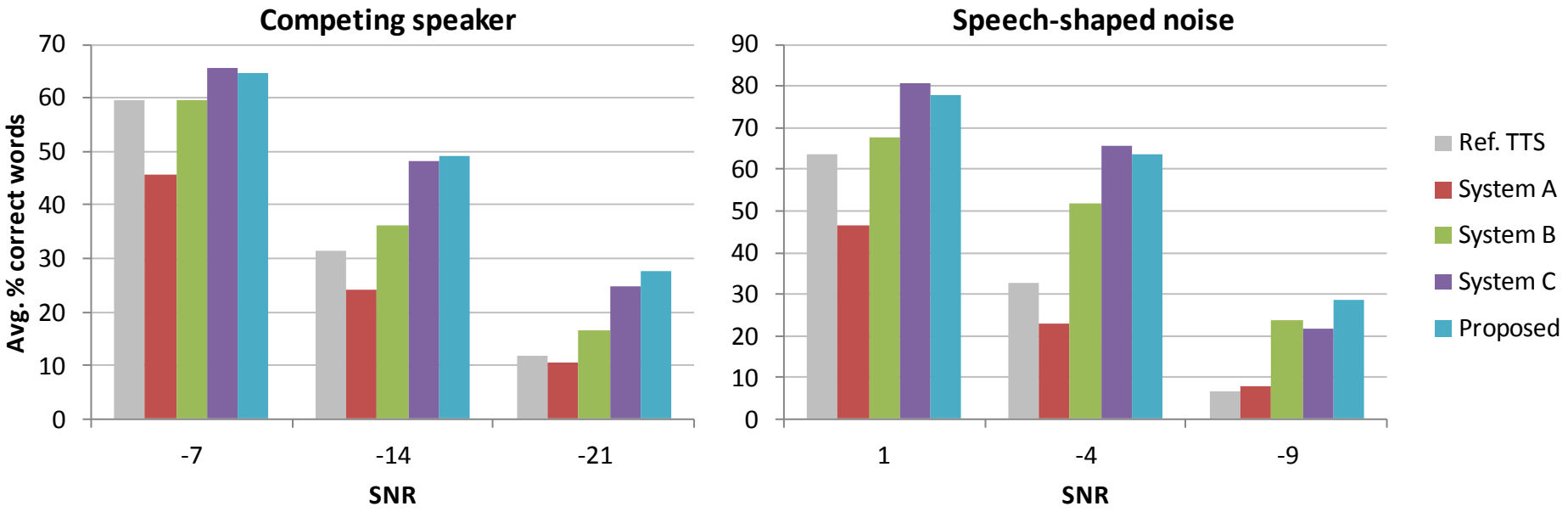


Always outperforms the ref TTS

Outperforms natural speech in mid-low SNR

Experiments

- Results of Hurricane Challenge

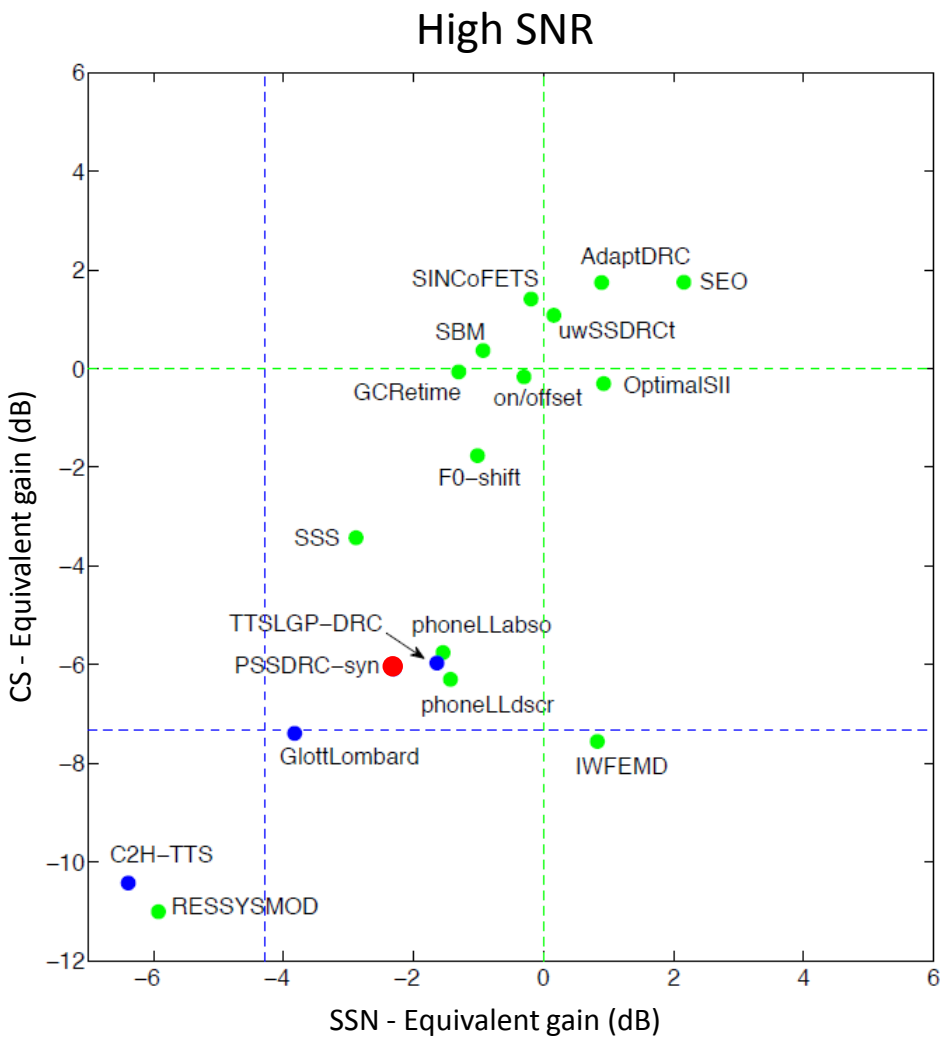


Better than A and B, comparable to C

C based on a speaker-adaptive framework and noise-adaptive modifications

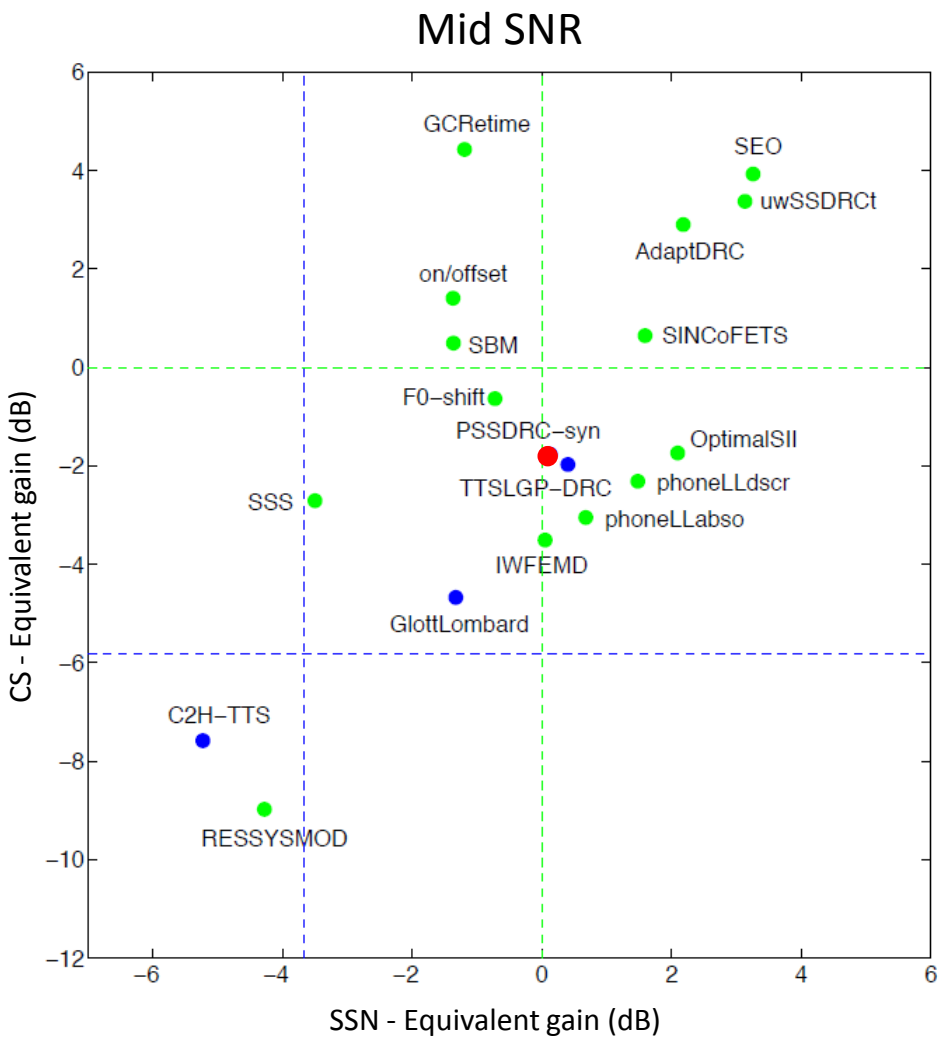
Experiments

- Results



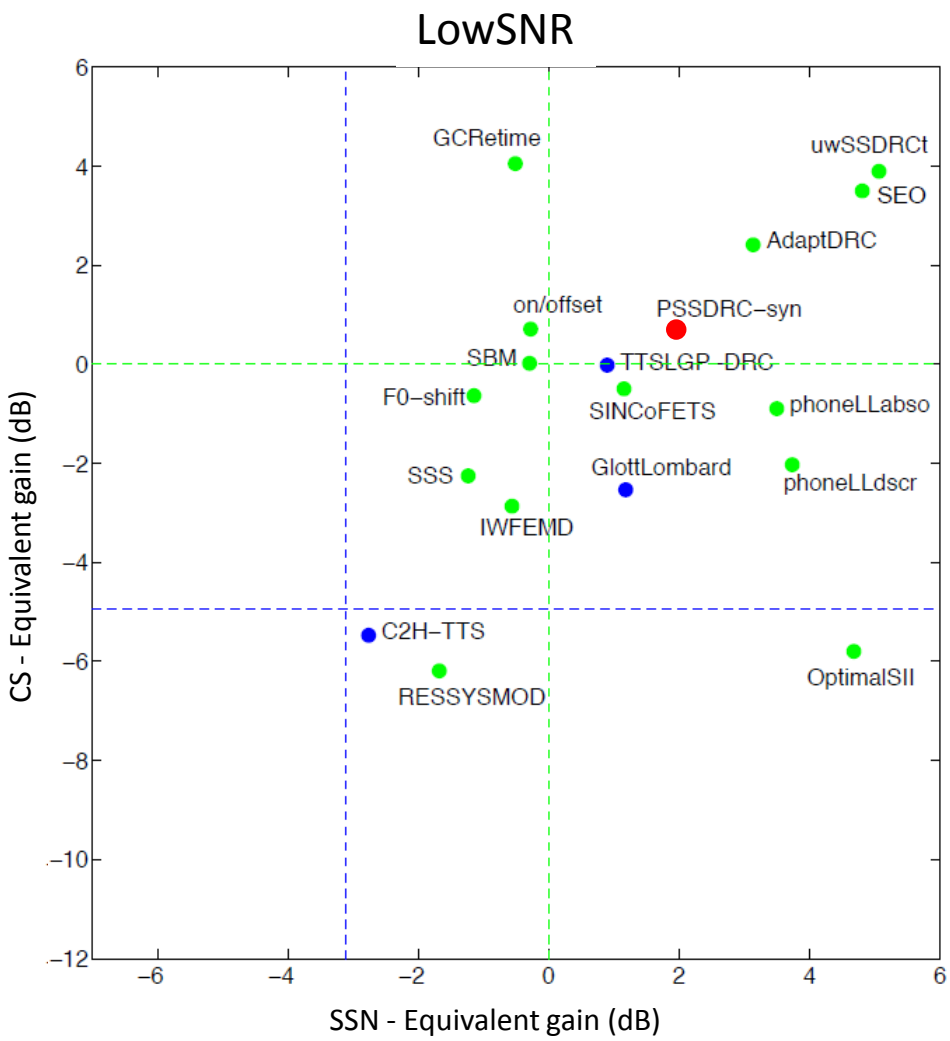
Experiments

- Results



Experiments

- Results



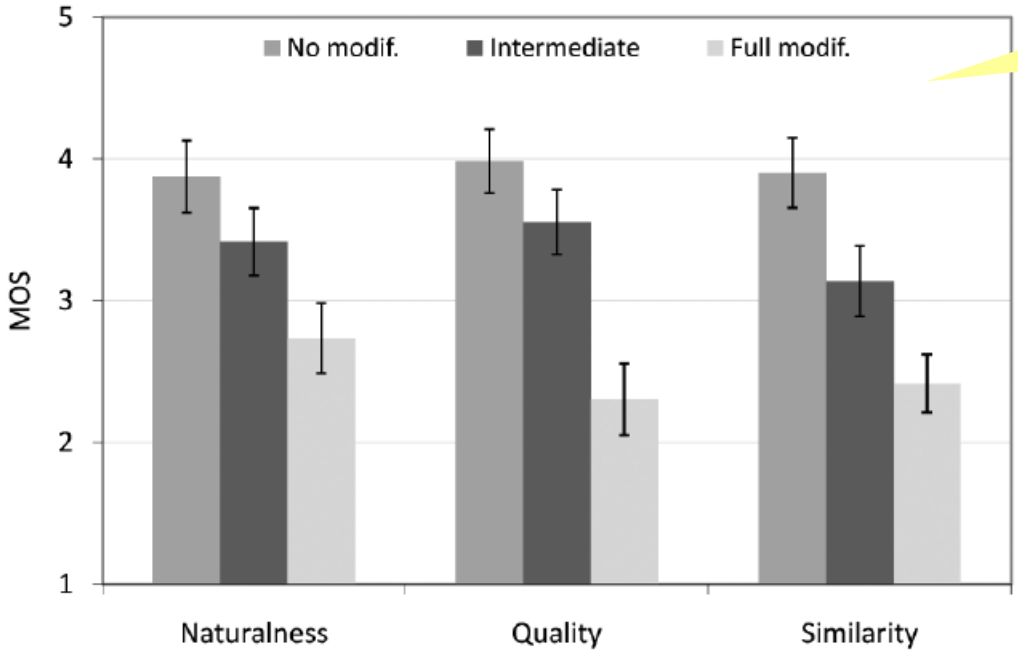
Experiments

http://listening-talker.org/showcase

(PSSDRC-syn)

Experiments

- Impact of modifications on other possibly important perceptual aspects



At some point the system should be made adaptive

Outline

- ~~Introduction~~
- ~~Modifications~~
- ~~Experiments~~
- Conclusions

Conclusions

- Noise-independent modifications based on hand-crafted rules: simple & cheap, no need to retrain!
- Easy to implement through a harmonic vocoder that takes MCEP + logF0 as input
- Probably speaker-independent
- Duration, F0 mean and range, DRC, formant sharpening (!), mid-f enhancement
- Very good results in an international campaign, even without any external data
- Speech is perceived as less natural → roughly-noise-adaptive version

Outline

- ~~Introduction~~
- ~~Modifications~~
- ~~Experiments~~
- ~~Conclusions~~

Acknowledgements

Tudor-Catalin Zorila

Yannis Stylianou

Cassia & Hurricane Challenge organizers

References

- D. Erro et al., “Enhancing the Intelligibility of Statistically Generated Synthetic Speech by means of Noise-Independent Modifications”, IEEE T ASLP, 2014
- C. Valentini-Botinhao, “Intelligibility enhancement of synthetic speech in noise” PhD thesis, Univ. Edinburgh, 2013
- M. Cooke et al., “The listening talker: a review of human and algorithmic context-induced modifications of speech”, CSL, 2014
- R. McAulay, T. Quatieri, “Sinusoidal Coding”, chapter in “Speech Coding and Synthesis”, Elsevier, 1995
- K. Rhebergen, N. Versfeld, “A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal hearing listeners”, JASA, 2005.
- M. Cooke et al., “Intelligibility-enhancing speech modifications: the Hurricane Challenge”, Proc. Interspeech, 2013

Sinusoidal Models for Highly Intelligible Text-to-Speech Synthesis

… or …

Ingelligibility enhancement using a harmonic vocoder

Daniel Erro - derro@aholab.ehu.es

ikerbasque
Basque Foundation for Science

Bilbao, Spain

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea