UR Signals and Systems

Perceptual Long-Term Harmonic plus Noise Modeling for Speech Data Compression

Faten Ben Ali, Sonia Djaziri Larbi

Signals and Systems Lab (U2S)
National Engineering School of Tunis (ENIT)
Tunis El Manar University, Tunisia

Outline

- ❑ Long-Term Harmonic plus Noise Model (LT-HNM)
- ❑ Perceptual LT-HNM for Data reduction
- ❑ Experimental Results for Data Compression

Outline

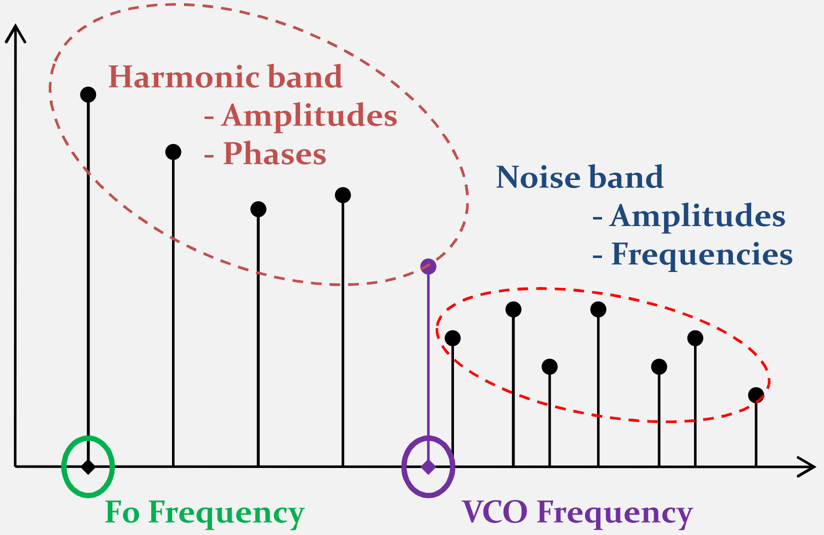
❑ Long-Term Harmonic plus Noise Model (LT-HNM)

❑ Perceptual LT-HNM for Data reduction

❑ Experimental Results for Data Compression

Long-Term Harmonic plus Noise Model (LT-HNM)

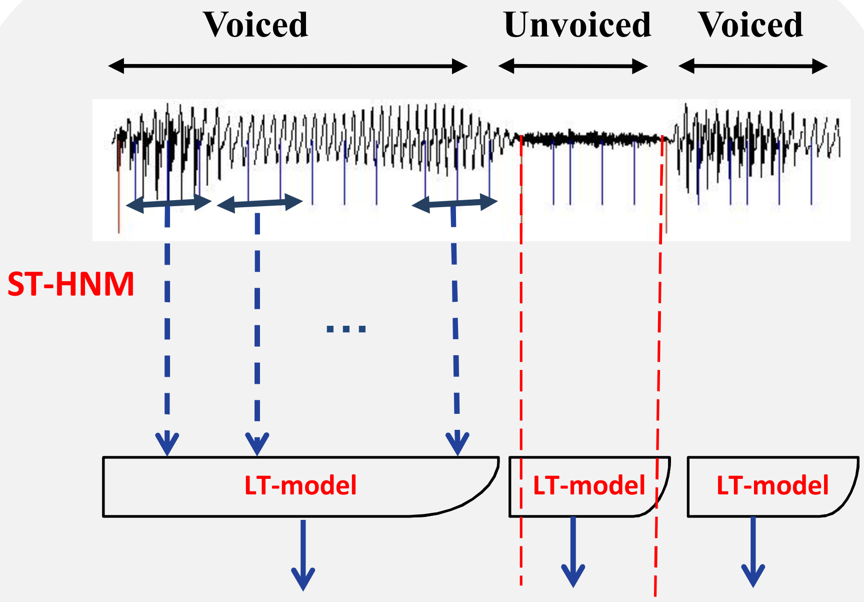
ST-model: Harmonic plus Noise Model (HNM)



❑Harmonic plus Noise Models:

- harmonic band: multiples of the fundamental frequency (F0)
- noise band: peak picking frequencies

❑The two bands are separated by a voicing cut-off frequency (VCO)



❑Speech signal is segmented into voiced ($F_0 \neq 0$) and unvoiced ($F_0 = 0$) LT-sections

❑ A Long-Term model is applied to the ST-HNM parameters (F_0, VCO frequencies and spectral amplitudes) for each LT-section.

Long-Term Harmonic plus Noise Model (LT-HNM)

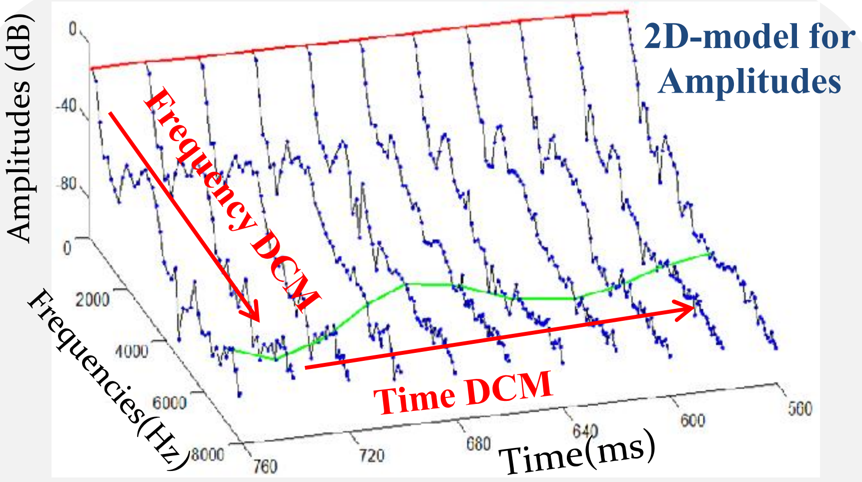
LT-model: Discret Cosine Model (DCM)

$$\hat{X}(n) = \sum_{p=0}^{P} c_p \cos\left(\frac{p\pi n}{N}\right)$$

❑ Applying a DCM to the time trajectory of the ST-HNM parameters (F0, FV and amplitudes) in a LT time section.

❑ Exploits the correlation between successive ST-parameters

❑ Optimization of the model order P



For amplitudes, we apply a DCM twice:

❑ first along the frequency axis to model the spectral amplitudes in a ST-frame (1D-DCM)

❑ second along the time axis to model the time trajectory of the 1D-DCM coefficients in a LT-section (2D-DCM).

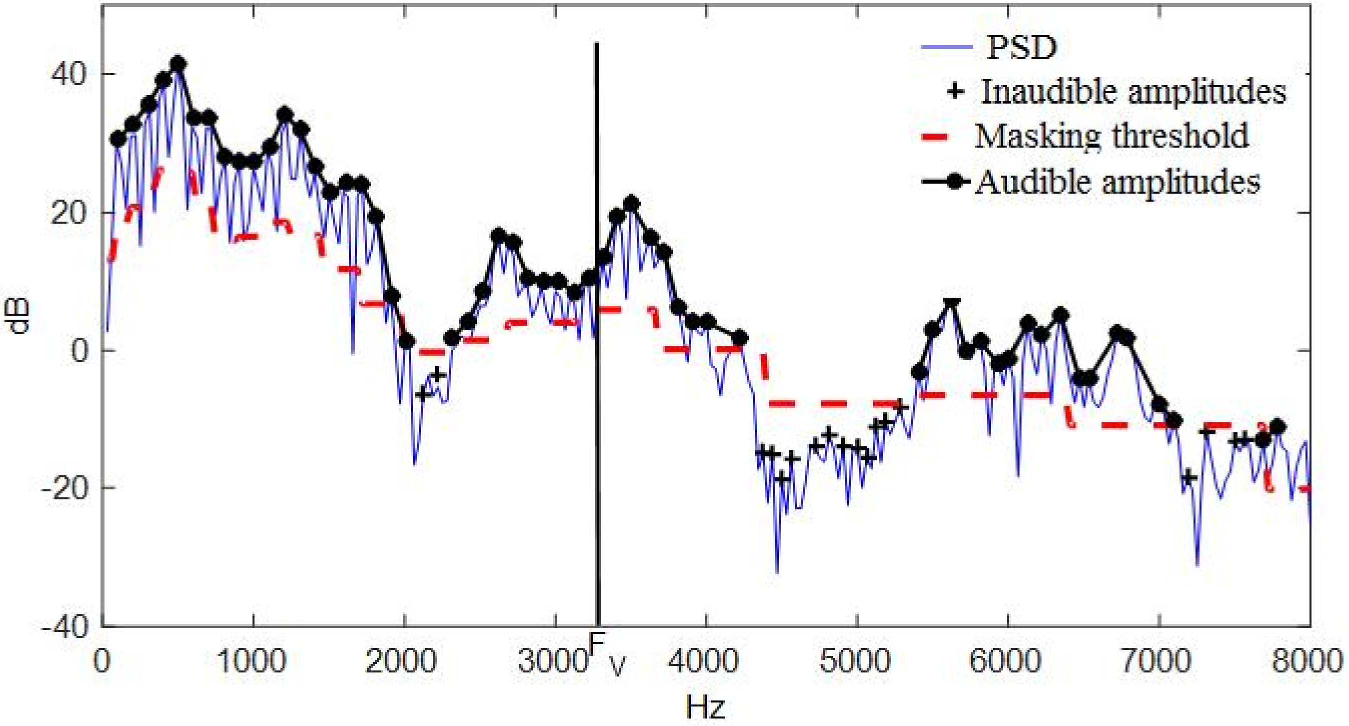
Outline

□ Long-Term Harmonic plus Noise Model (LT-HNM)

□ Perceptual LT-HNM for Data reduction

□ Experimental Results for Data Compression

Auditory Masking



❑ A masking threshold is computed for each ST-frame

❑ The spectrum amplitudes (harmonic + noise) are compared to the masking threshold

→ **Only amplitudes above the mask are selected as audible**

ST-HNM Data Reduction

p-ST-HNM

Only audible amplitudes are considered in the ST-HNM, inaudible ones are discarded from the model

→ The data size of the model parameters is considerably reduced: up to 50% in a ST-frame

→ Reduction of the data-rate with the equivalent perceptual quality

⇩

p-LT-HNM

The LT-modeling is applied to the parameters of the p-ST-HNM

⇩

Double data-compression: auditory masking + LT-modeling

Outline

❑ Long-Term Harmonic plus Noise Model (LT-HNM)

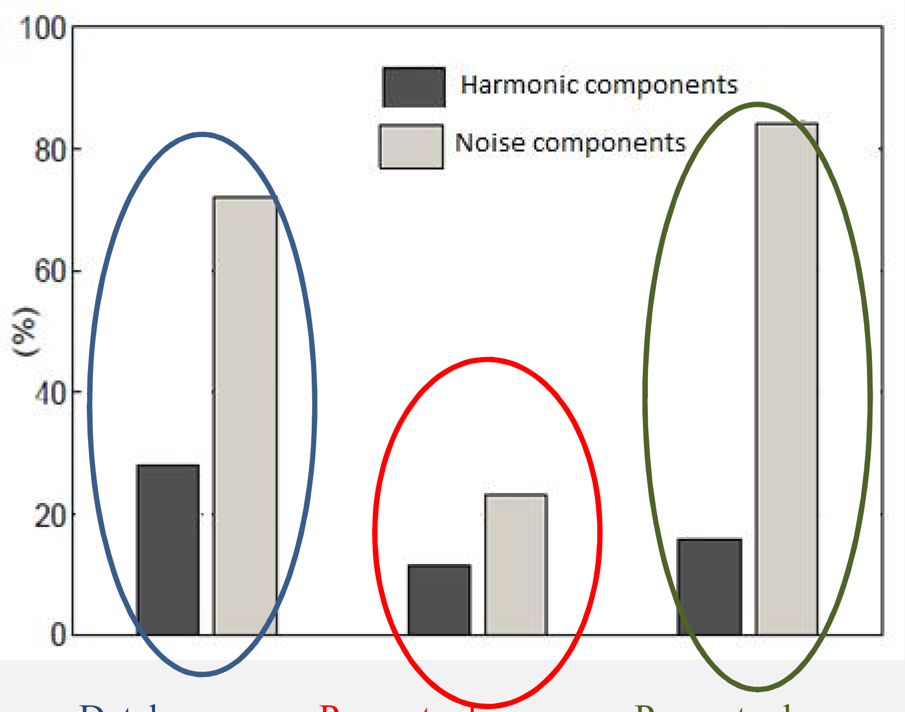
❑ Perceptual LT-HNM for Data reduction

❑ Experimental Results for Data Compression

Overall Data Reduction

Database

- 600 TIMIT speech samples at 16kHz
- Total duration ~ 17mn
- ST-HNM analysis with 30ms ST-frames and a hop-size of 20ms (81539 ST-frames)
- The masking threshold is attenuated with -5dB



Database statistics

Perceptual compression rate

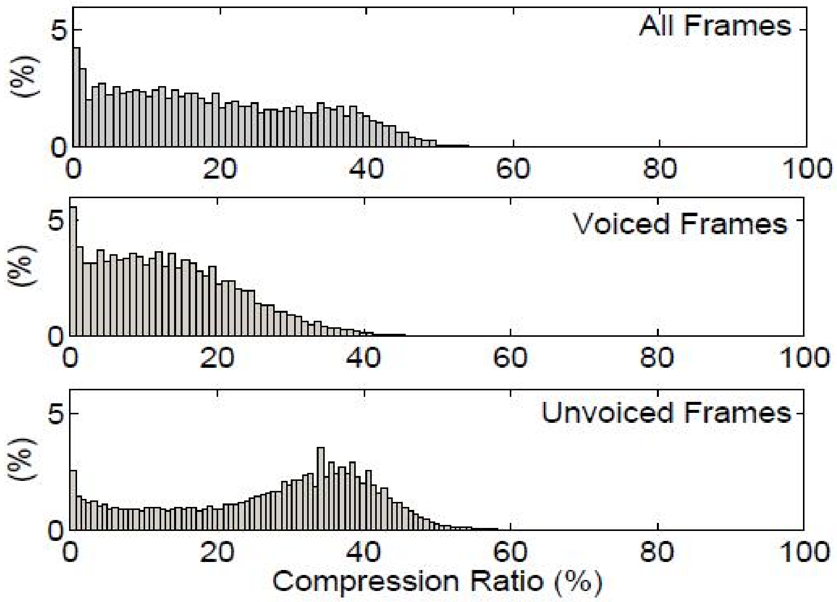
Perceptual compression statistics

❑ 72% of database frequencies are noise frequencies, while 27.8% are harmonics

❑ Total frequency components compression ≈ 20%: noise band compression ≈ 23.3%, harmonic band compression ≈ 11.3%

❑ 84.1% of achieved compression is due to the noise band, while the harmonic frequencies contribute only by 15.% to the total compression

Data Reduction in V/UV ST-frames

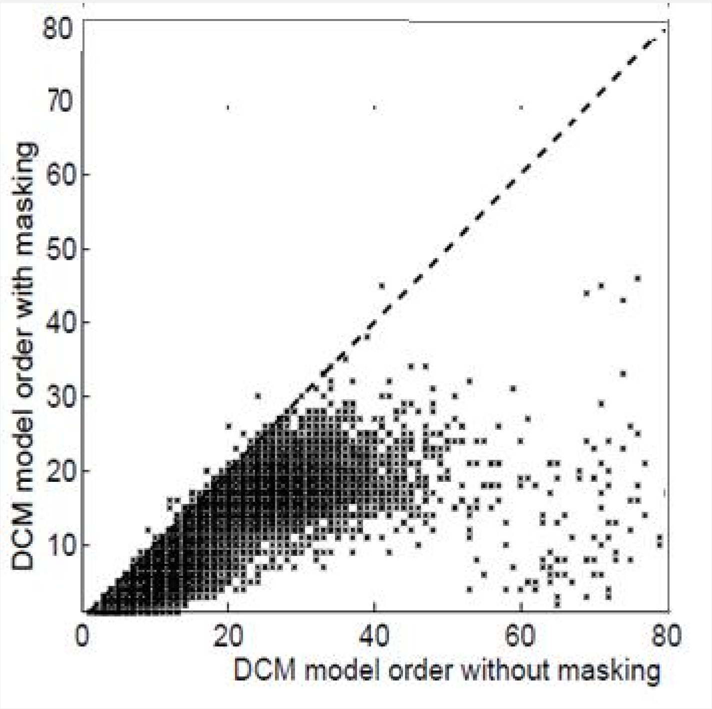


❑ Data compression rate is up to 50% in a ST-frame

❑ Compression rate is higher for unvoiced ST-frame (entirely composed of noise frequencies)

❑ Higher contribution of noise band to the total compression

Reduction of the LT-HNM coefficients rate



The 1D-DCM order is considerably reduced when applying the auditory masking

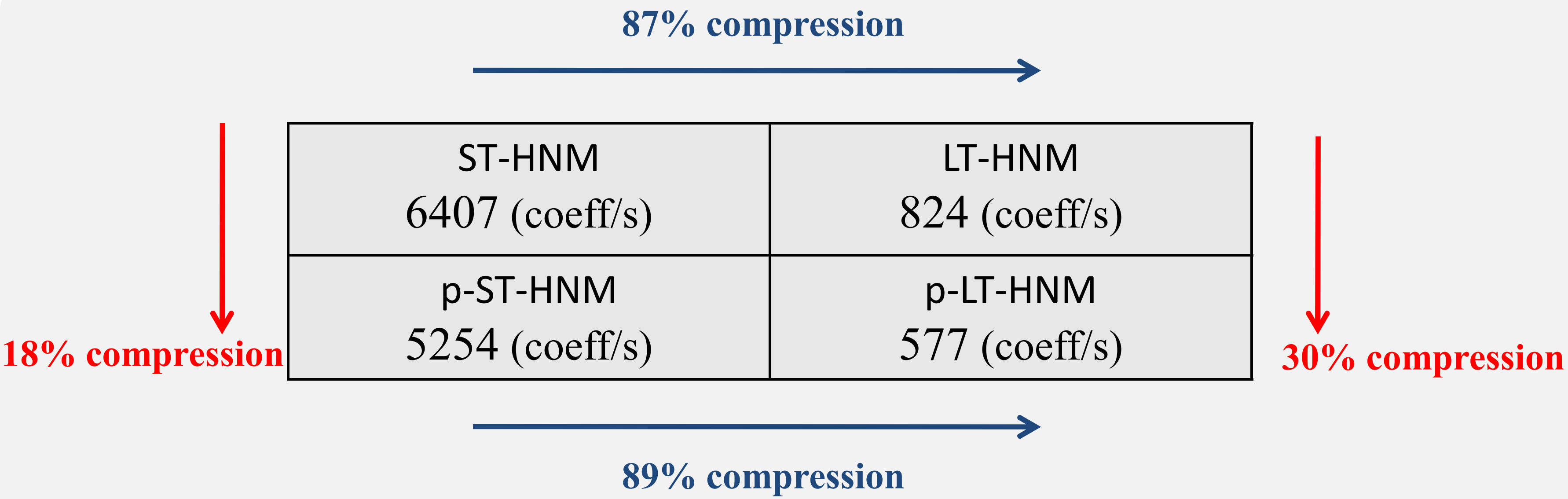
⇩

The data size to be LT-modeled is reduced

⇩

Reduction of the LT-HNM data size

Parameters-Rate Gain



87% compression

ST-HNM 6407 (coeff/s)	LT-HNM 824 (coeff/s)
p-ST-HNM 5254 (coeff/s)	p-LT-HNM 577 (coeff/s)

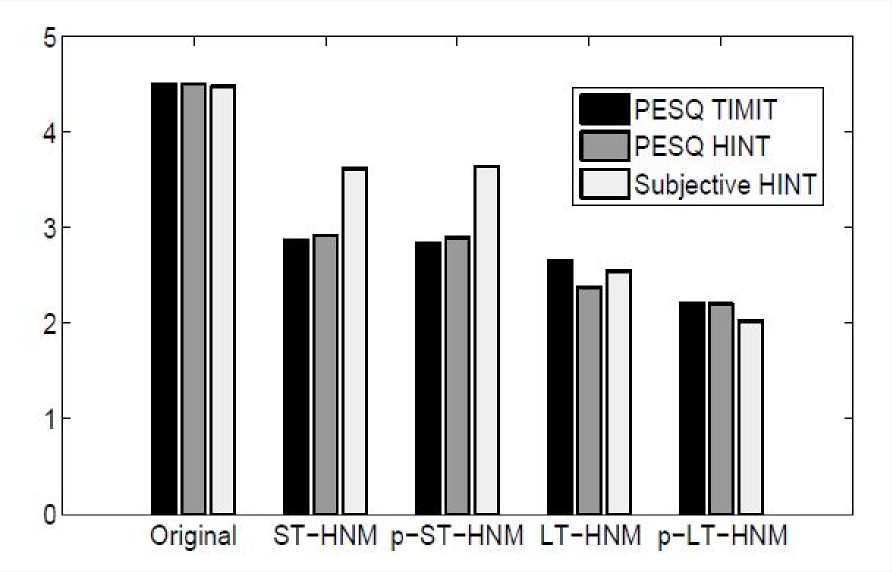
18% compression

30% compression

89% compression

Data rate gain of 18% for p-ST-HNM and 30% for p-LT-HNM

Listening Quality (PESQ and MOS Scores)



❑ Mean PESQ scores: 40 speech samples from TIMIT (English) and 20 samples from HINT (French)

❑ MOS Subjective listening test applied to HINT samples (12 French speakers participants)

→ No auditory distortion when applying the p-ST-HNM (PESQ and MOS)

→ No significant auditory distortion when applying the p-LT-HNM (PESQ and MOS)

Conclusion and Perspectives

Conclusion:

Two stages of compression:

❑ Perceptual based compression: 18% (ST-HNM → p-ST-HNM)
❑ Perceptual LT modeling: 89% (p-ST-HNM → p-LT-HNM)

→ **Total compression: 90% (ST-HNM → p-LT-HNM)**

Perceptual Quality

❑The perceptual HNM and generic HNM provide equivalent quality scores

Perspectives:

❑ A two stage vector quantization is currently being applied to the perceptual LT-HNM parameters and will to design a low bit-rate speech codec.

Faten Ben Ali, Sonia Djaziri Larbi

Signals and Systems Lab (U2S), National Engineering School of Tunis (ENIT),
Tunis El Manar University, Tunisia

UR Signals and Systems

U2S ENIT <u2s@enit.rnu.tn>