Vocaine

the Vocoder - Summer School 2015

Yannis Agiomyrgiannakis

Presentation Outline

- A short history of Vocoding
- Vocoders for TTS
  - TTS synthesis
  - TTS Quality
  - Google TTS
- Speech Signal
- Vocaine
  - Overview
  - Speech model
  - Pitch-synchronous framing
  - Spectral sampling
  - Deterministic + stochastic phase model
  - Quadratic phase splines
  - Coherent noise-modulation model
  - Unsafe Super-fast cosines
- Results
- Conclusions

Vocoders - the elder problem in Speech Synthesis

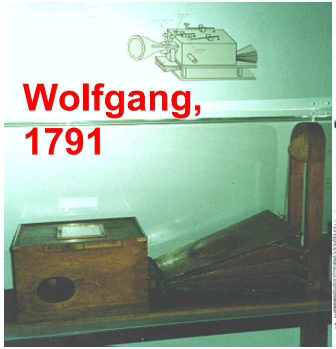
Definition (Wikipedia): A **vocoder** ([/ˈvoʊkoʊdər/](), short for *voice encoder/decoder*) is an analysis/synthesis system, used to reproduce human speech.

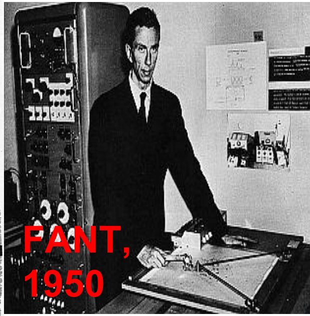
Mechanical era:

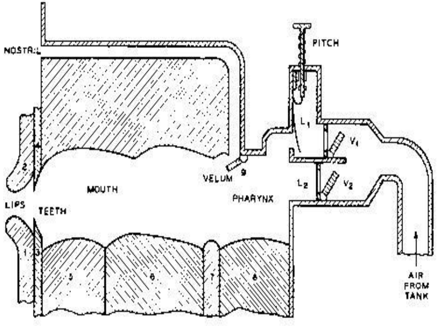
1. Wolfgang von Kempelen, “*Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*”, 1791, Vienna
2. Joseph Faber, "Euphonia", 1846, London
3. R. R. Riesz, “Mechanical Talker”, 1937, USA

Electrical era:

1. Homer Dudley, “VODER”, 1939, New York
2. Gunnar Fant, “Orator Verbis Electris”, 1950s, Sweden







RIESZ,1937

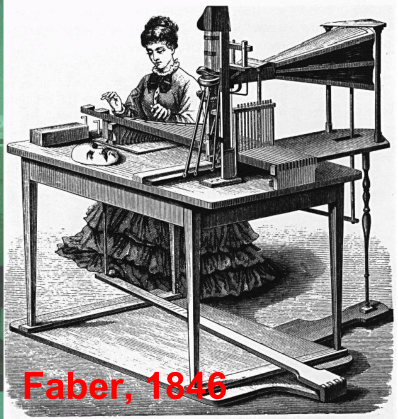
Vocoders - the elder problem in Speech Synthesis

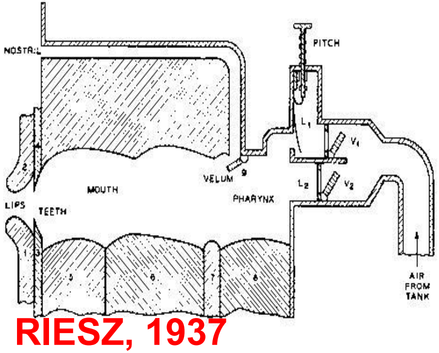
Computer era - Speech Coding:

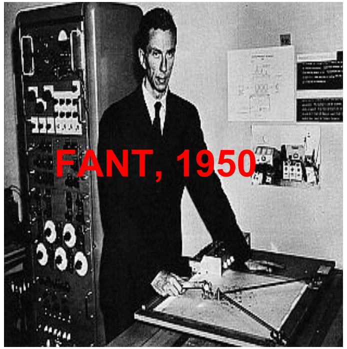
1. **1970s - 1984:** FS1015 2.4 kbps LPC vocoder (LPC-10), MOS 2.20
2. **1993 - 1996:** FS1016 2.4 kbps secure coder, MOS 3.10
3. **1987-2001:** Griffin et al., "Multi-Band Excitation Vocoder" family of vocoders powers most satellite telephony standards (IMBE, …, AMBE+2), via MIT spin-off DVIS Inc.
4. **1995:** McAulay, Quatieri, "Sinusoidal Transform Coding (STC)", MIT Lincoln Labs

Computer era - Speech Synthesis:

1. **2001:** Stylianou et al., "Harmonic + Noise Model", Bell Labs
2. **2008:** Kawahara, "Tandem-Straight" (latest version of STRAIGHT)
3. **2013:** Erro et al., "Harmonic + Noise Model" (STC + HNM hybrid)

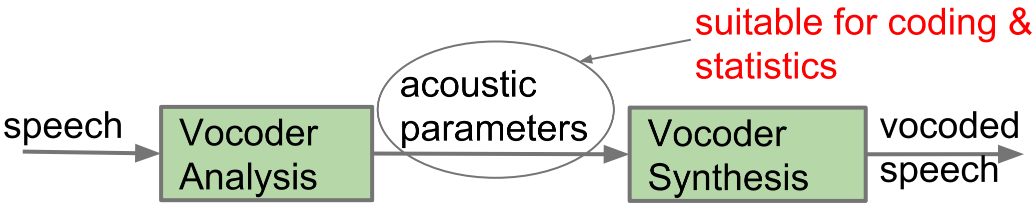




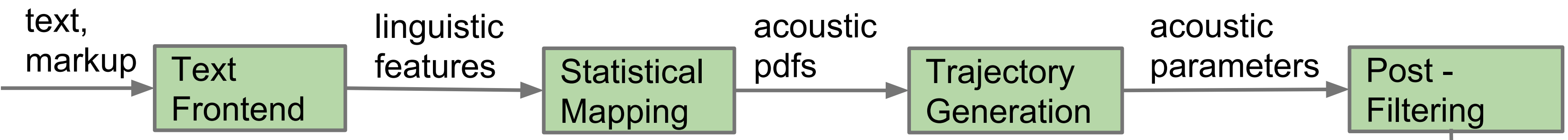


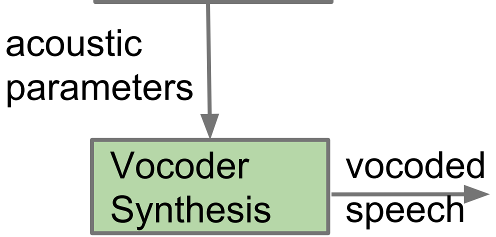
Vocoders - TTS Synthesis

Analysis/Synthesis: vocoders provide a parametric representation of the speech signal suitable for coding & statistics.



Statistical Parametric Speech Synthesis:

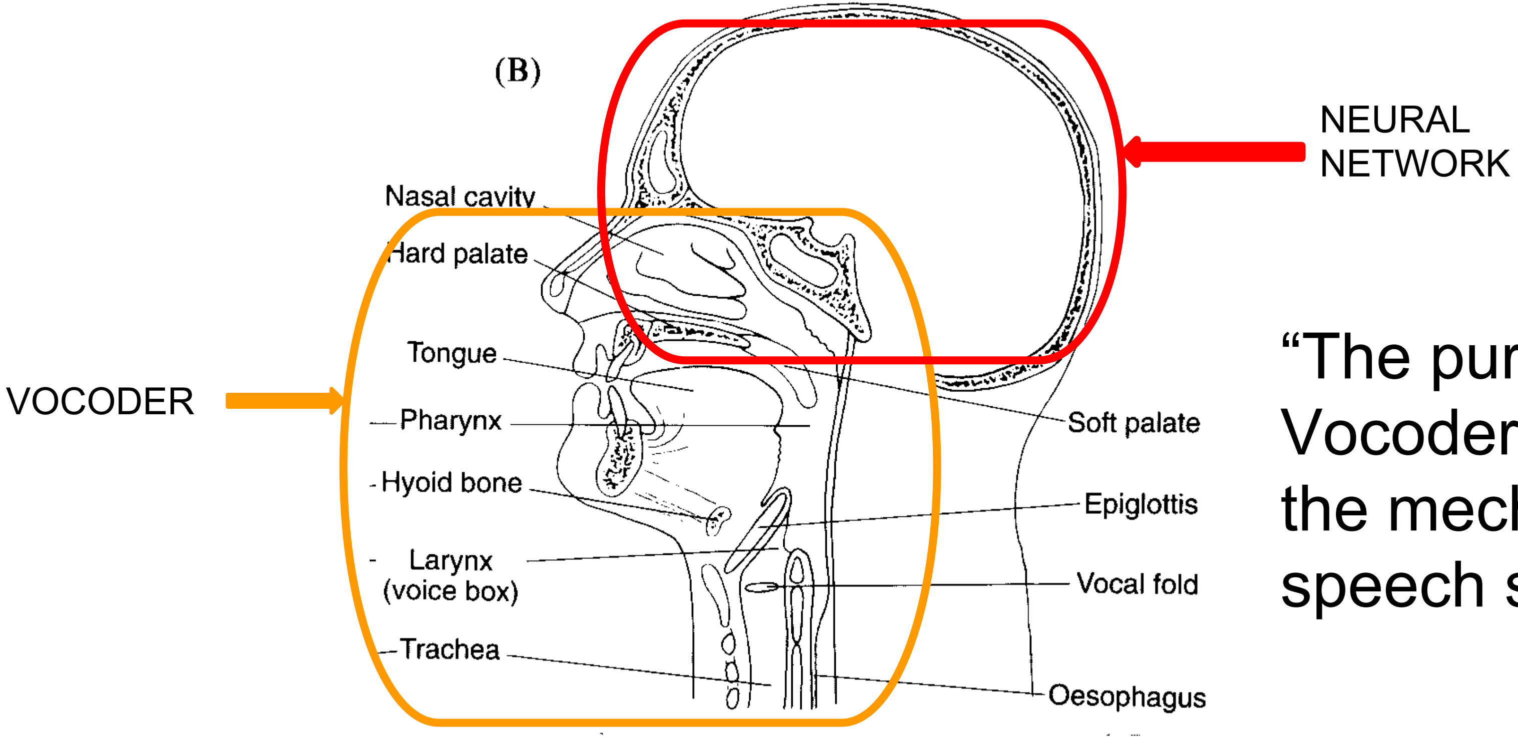




Statistical Mapping:

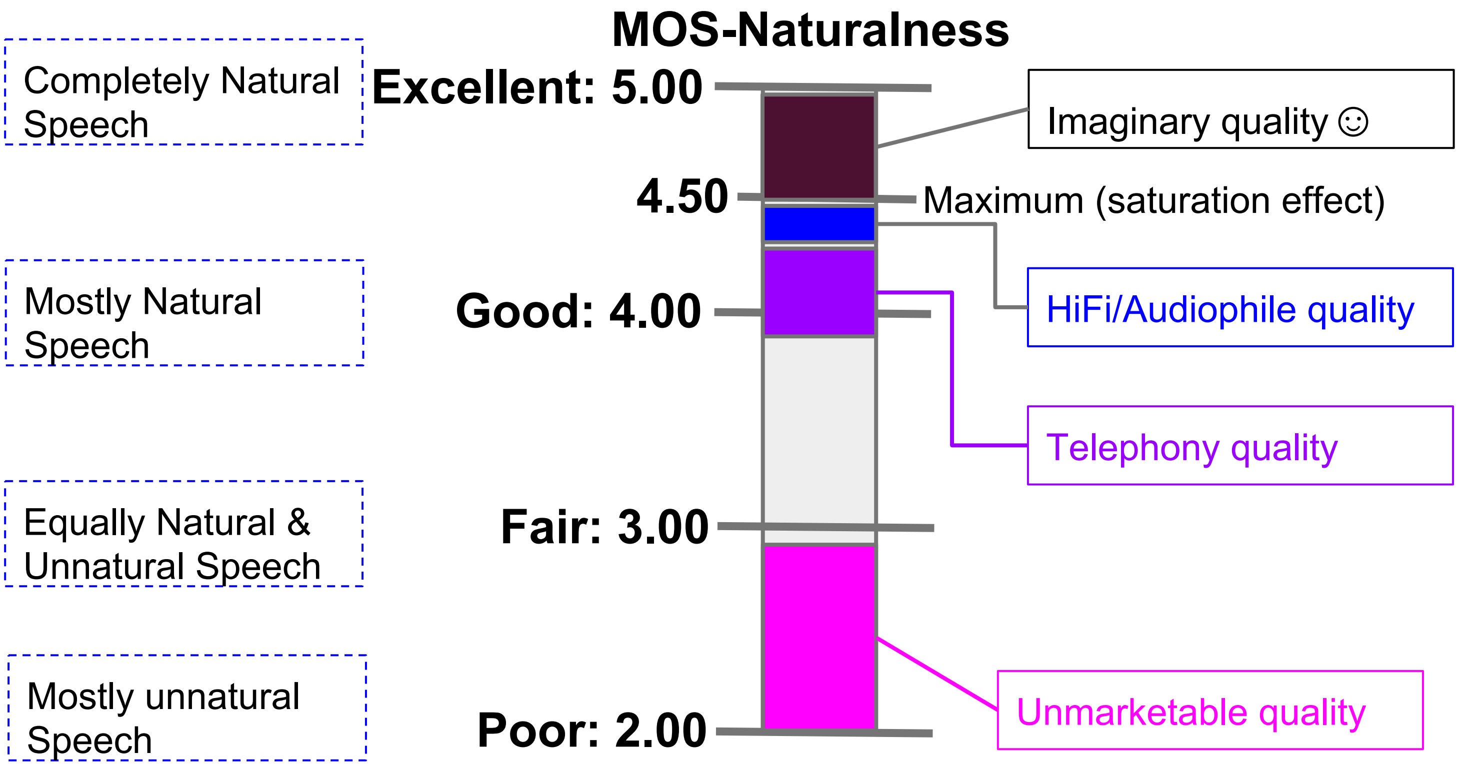
- **Input space:** linguistic features
- **Output space:** acoustic pdfs
- **Methods:** decision trees, DNNs, etc

Vocoders - TTS Synthesis

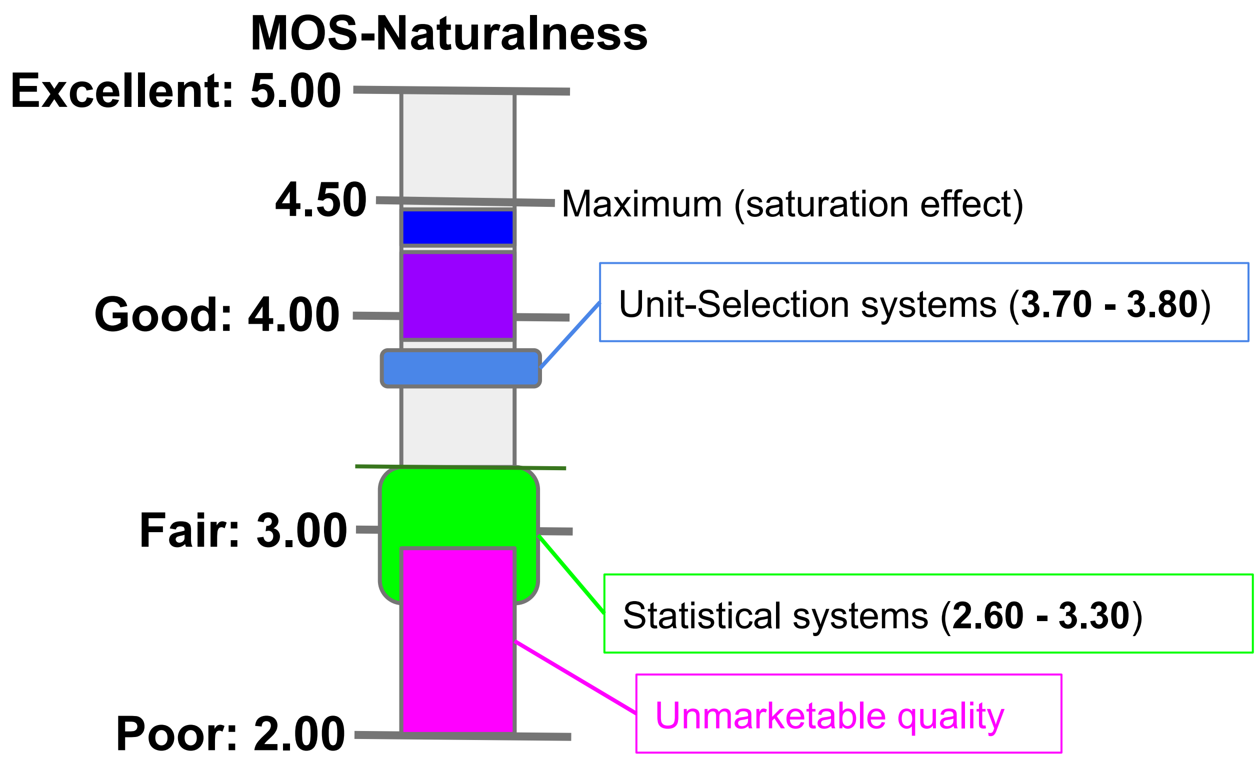


"The purpose of the Vocoder is to replace the mechanics of speech synthesis."

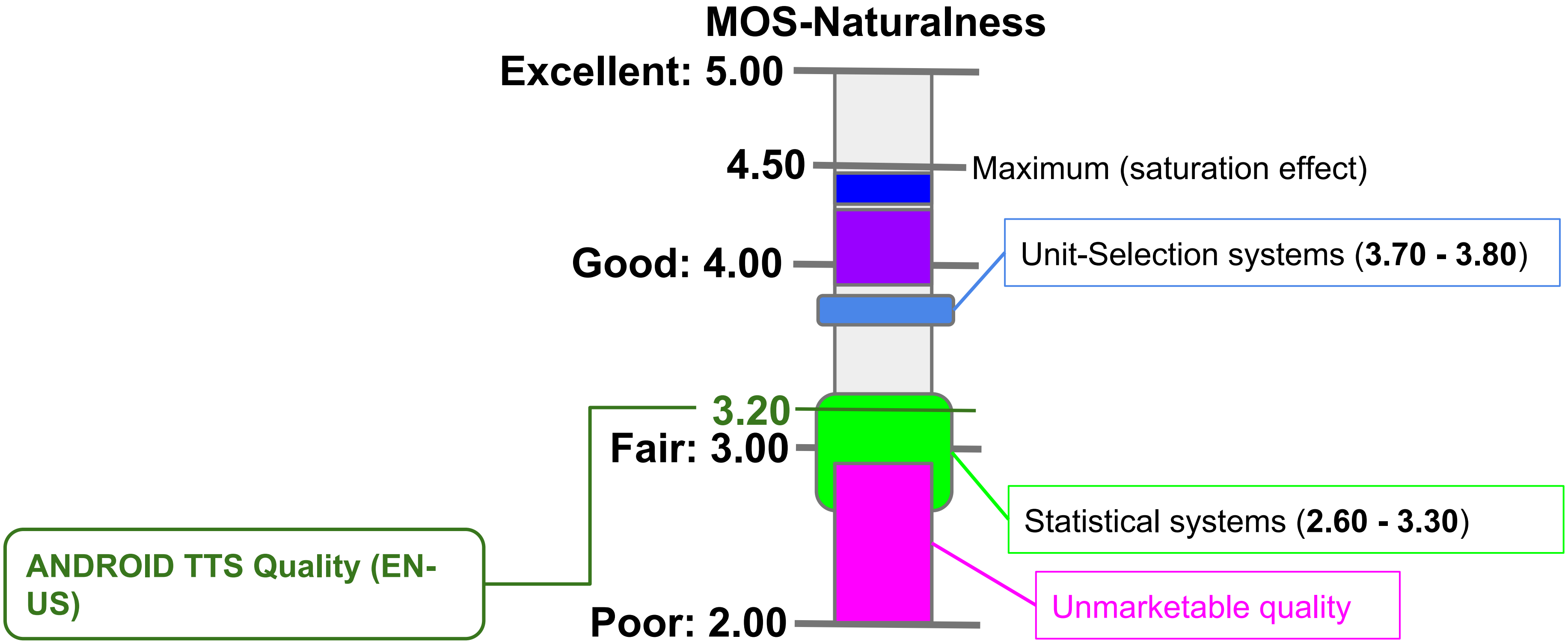
Vocoders - TTS Quality



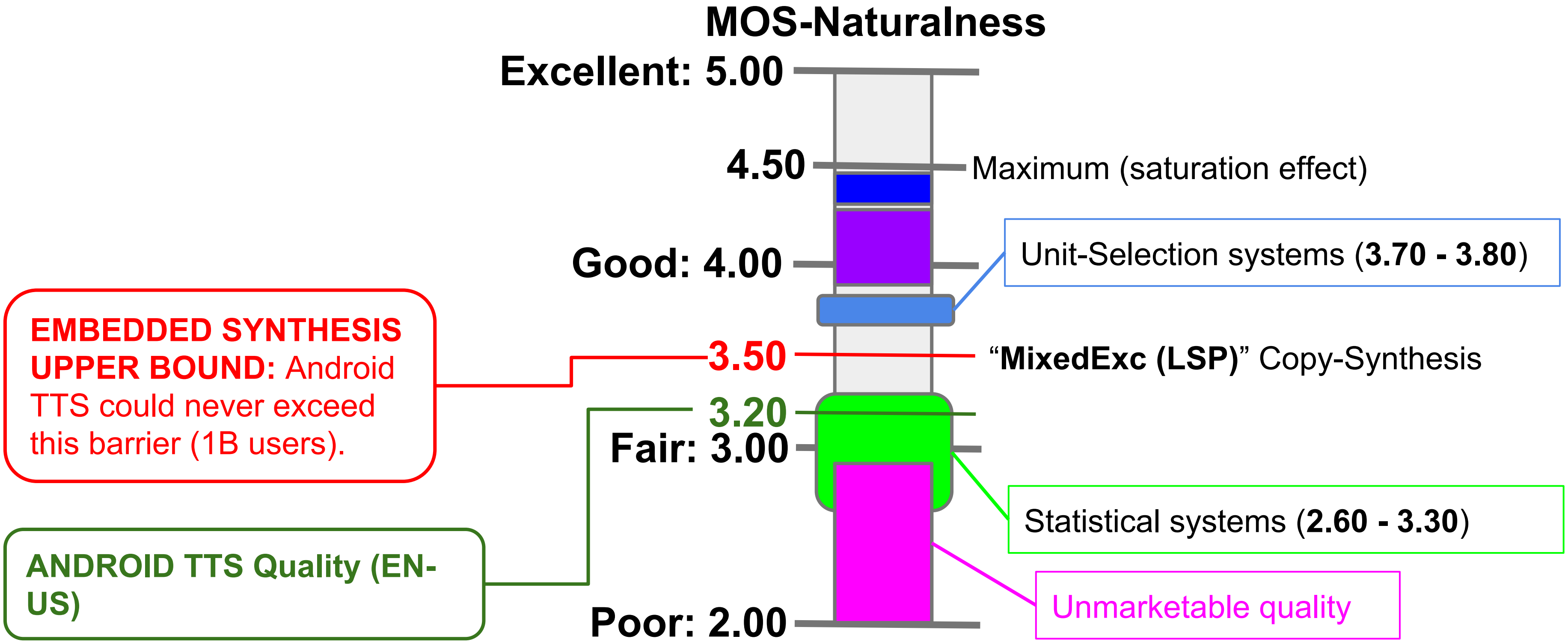
Vocoders - TTS Quality - EN-US/FR Summary (pre-Vocaine).



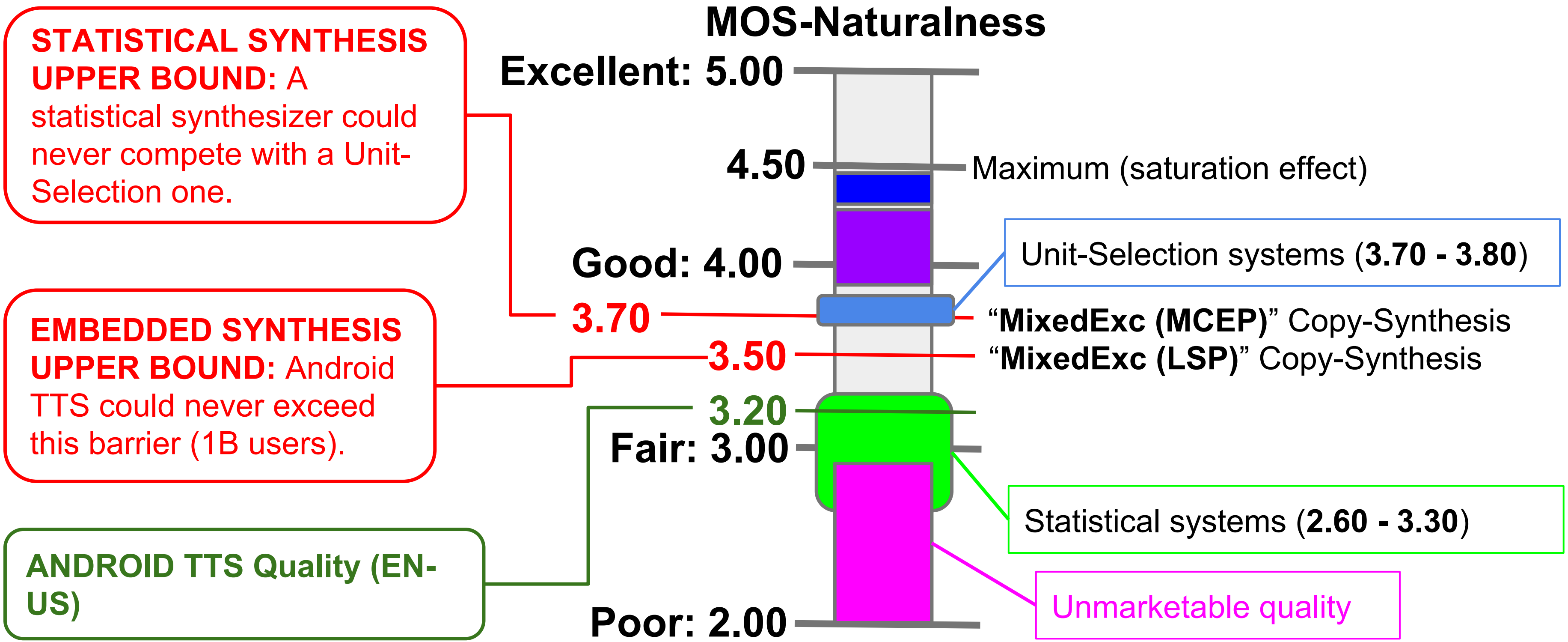
Vocoders - TTS Quality - EN-US/FR Summary (pre-Vocaine).



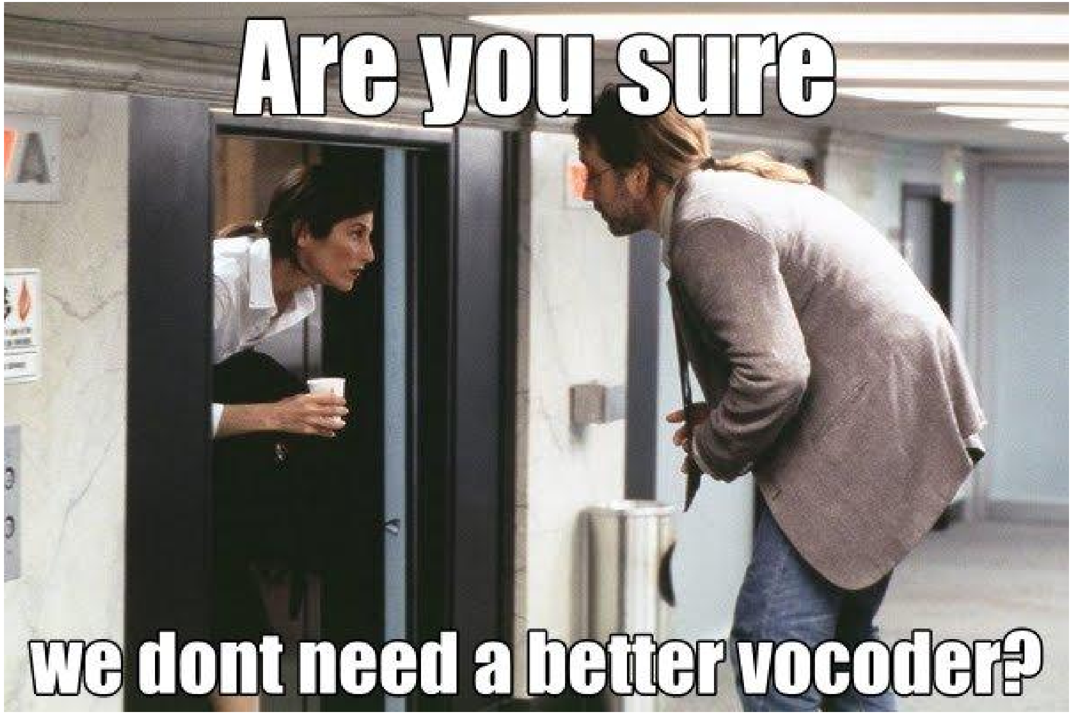
Vocoders - TTS Quality - EN-US/FR Summary (pre-Vocaine).



Vocoders - TTS Quality - EN-US/FR Summary (pre-Vocaine).



Vocoders - TTS synthesis meme



Vocoders - Google TTS - Pre-Vocaine

- **Used in** HMM-based speech synthesizers for **Android, Chrome, Navigation**:
    - **Low-latency** for accessibility & driveabout, etc.
    - **Ultra-low-footprint** versions in Android OS.
    - **Lower quality** than Unit-Selection.
    - **Low-end solution**, suitable for low-spec devices.
    - **Biggest user-base**, the one that most users listen to.
- **Vocoder analysis** based on SWOP-STRAIGHT.
- **Vocoder synthesis** based on:
    - Mixed excitation (embedded excitation, server excitation).
    - Mel-Cepstra (MCEP) using MLSA filter.
    - Mel-Line-Spectrum-Pairs (MLSP).
- **Upper bounds the quality** of a statistical synthesizer:
    - **STRAIGHT:** 4.07 MOS
    - **Server vocoder** (SWOP-MCEP + SERVER-EXC): 3.70 MOS
    - **Embedded vocoder** (SWOP-MCEP + SERVER-EXC): 3.50 MOS
    - Improving upper-bound → improving quality of SPSS.

- **0.50 MOS gap between our current embedded vocoder and the state-of-the-art !!!**

Speech Signal - Waveform Modeling Pillars

Incorporating implicit or explicit assumptions.

Ear

- Auditory models & principles:
  - frequency scaling (mel-scale)
  - Amplitude compression (log)
  - phase coherence

Mouth

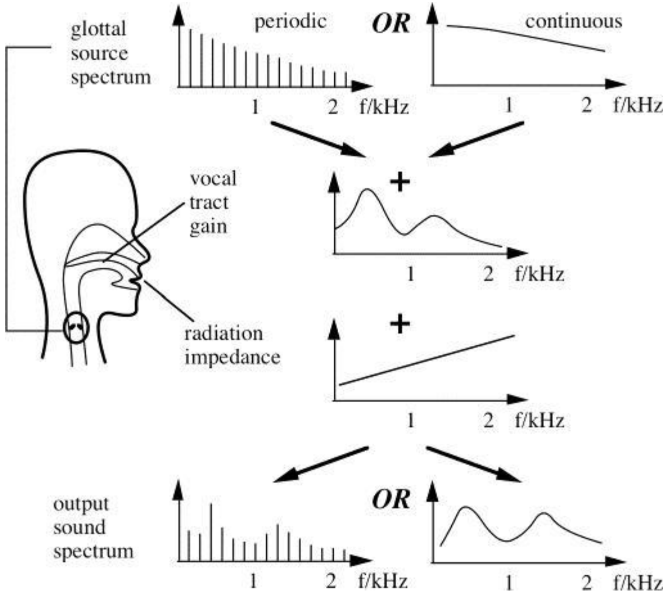
- Speech production models:
  - glottal excitation
  - vocal tract
  - nasal tract
  - aspiration

Speech Signal - The ubiquitous Source / Filter model

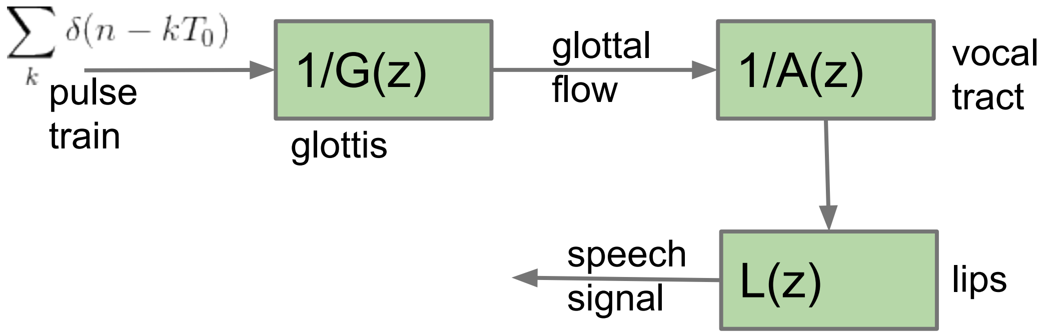
Mechanical models have had a tremendous impact on shaping our perspective about the speech signal.

Dichotomies:

- source / filter
- deterministic / stochastic
- amplitude / phase



Simple Linear Source/Filter model:



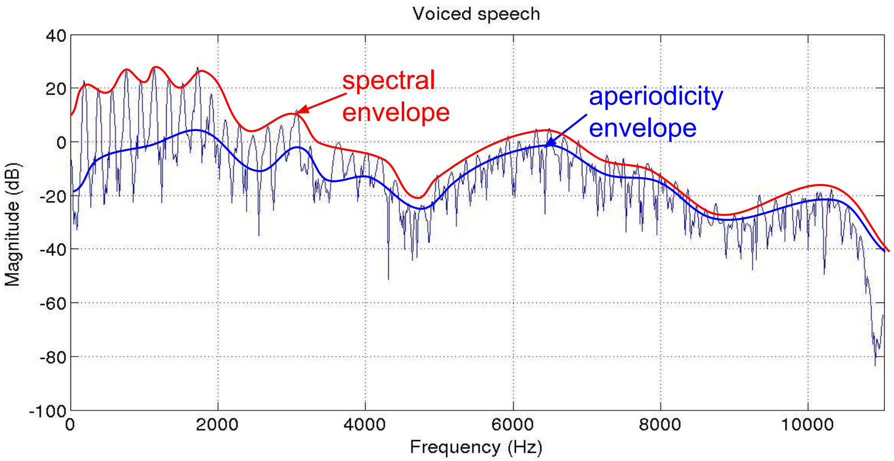
Speech Signal - Deterministic / Stochastic decompositions

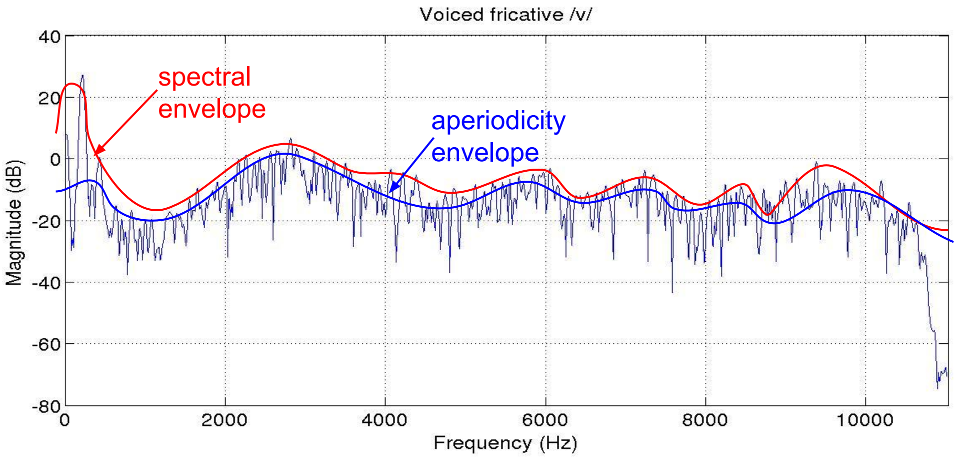
A multitude of phenomena generate **non-deterministic contributions** to the speech signal.

- **aspiration** generated at the glottis introduces aharmonic components.
- **frication** at an vocal tract constriction (i.e. voiced fricatives and plosives).

Dichotomies:

- source / filter
- deterministic / stochastic
- amplitude / phase





Speech Signal - Amplitude / Phase decompositions

Dichotomies:

- source / filter
- deterministic / stochastic
- amplitude / phase

A frequency-domain perspective: The speech signal as a sum of sinusoids.

$$s(n) = \sum_{k=1}^{K} A_k(n) cos(\phi_k(n))$$

Many speech models assume that sinusoidal components are harmonically related:

$$\frac{\partial \phi_k(n)}{\partial n} = k\omega_0$$

- **Amplitude:**
  - measured
  - sampled from a spectral envelope
- **Phase:**
  - measured
  - pulse train with phase model for pulses:
    - minimum-phase (eq. source-filter model)
    - zero-phase (i.e. MBE codecs)
    - fixed random phase envelope (Vocaine)

Vocaine - Overview

- **High spectral resolution:**
  - No inherent restriction in spectral resolution.
  - No complexity penalty.
- **Decouples spectral parameterization from DSP implementation:**
  - Mel-Cepstra, Mel-LSP, band-aperiodicities, MCEP-aperiodicities.
  - easy to extend to arbitrary speech parameterizations.
- **Asynchronous phase model:**
  - **TTS Hybrids** with **Stochastic-Unit**: blending vocoded speech with recorded units.
  - Full signal models - brings **phase information** into the game.
- **Ultra-wideband and beyond:**
  - Supports 8 kHz, 16 kHz, 22kHz, 32 kHz, 48 kHz sampling rates.
- **Universal:**
  - Supports most modern speech models: STRAIGHT, HNM, MBE, STC, AhoCoder, etc.

Vocaine - Overview

- **High quality:**
  - Can we beat STRAIGHT? → YES
  - To the infinite (~4.5 MOS score) and beyond !?
- **Low computational complexity:**
  - Almost as fast as our fastest (embedded) vocoder.
  - Low numerical sensitivity → **fixed-point** implementations are easy.
  - Designed for **SIMD DSP** operations from scratch.
  - Multi-core / streaming friendly.
- **Simplicity:**
  - Keep the math simple.
  - Simple C++ design.

Vocaine - Speech Model

- **Expressing the speech signal in a single equation:**

$$s(n) = A_1(n)\cos(\phi_1(n)) + \sum_{k=2}^{K} A_k(n)\left[\gamma_0 - \gamma_1\alpha_k(n)\cos(\phi_1(n))\right]\cos(\phi_k(n))$$

K : number of harmonics

$n = 1, 2, ..., T_s$: time index (in samples)

T_s : synthesis period

$A_k(n)$: instantaneous amplitude of k-th harmonic

$\phi_k(n)$: instantaneous phase of k-th harmonic

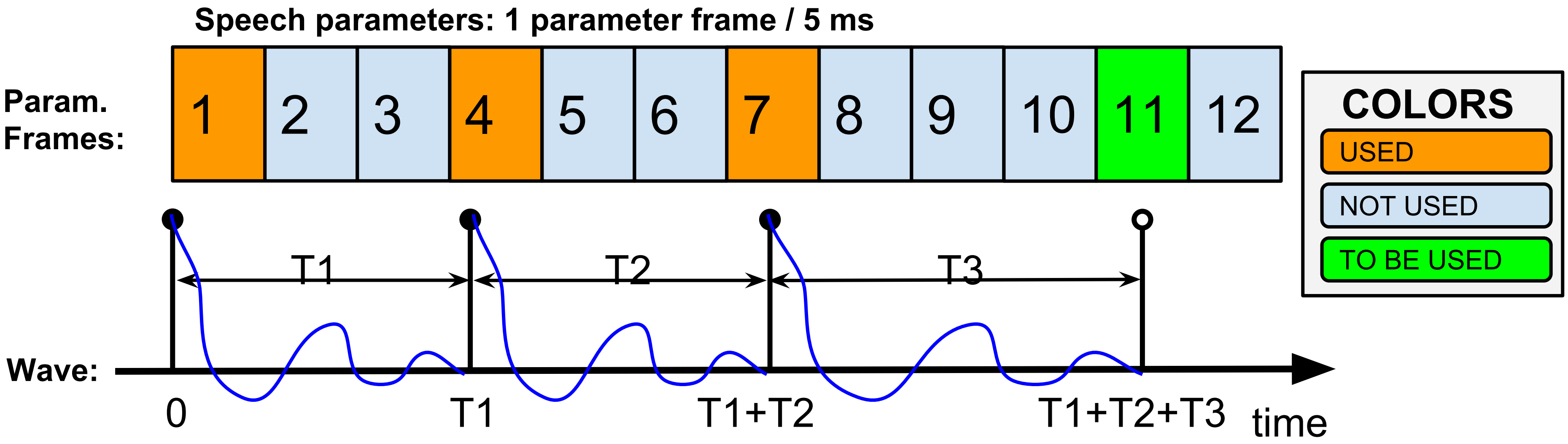
$\alpha_k(n)$: instantaneous aperiodicity of k-th harmonic in $[0, 1]$

γ_0 : modulation bias (i.e. 1.20)

γ_1 : modulation factor (i.e. 0.5)

Vocaine - Pitch-synchronous framing

- **Synthesis is made one period at a time:**



Reference Synthesis Instants (RSI): Glottal Closure Instants (GCI) + unvoiced pitchmarks

Vocaine - Spectral sampling

- **Any speech parameterization can be used.**
  - notice the excessive use of cosines

Spectral Sampling:

Mel-Cepstrum:

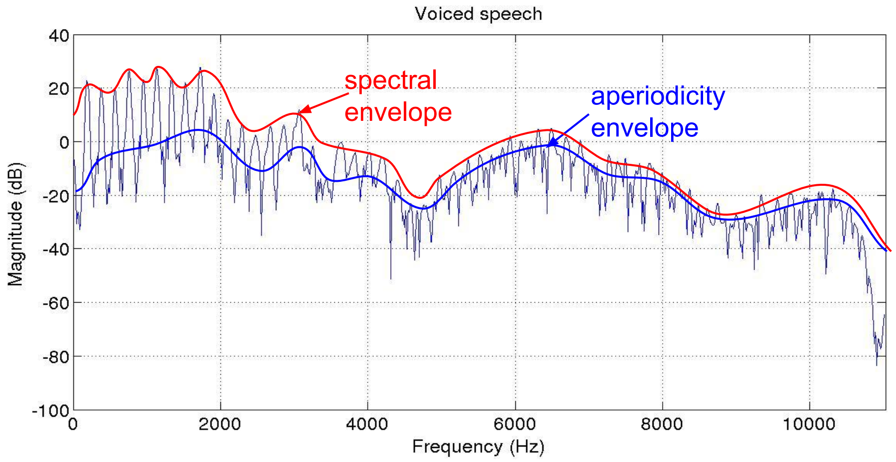
$$H(\omega) = exp\left(\sum_{m=0}^{M} c_m cos(m\omega)\right)$$

Mel-Line-Spectral-Pairs:

$$H(\omega) = \sqrt{\frac{2^{-M}}{\sin(\omega/2)^2 \prod_{m=2,4,\ldots} (\cos(\omega) - \cos(\omega_m))^2 + \cos(\omega/2)^2 \prod_{m=1,3,\ldots} (\cos(\omega) - \cos(\omega_m))^2}}$$

Vocaine - Deterministic + stochastic phase model 1 / 2

- Vocaine accepts **phase values sampled exactly at the RSI** (Glottal Closure Instants for voiced speech).
  - **Enables full-speech models:** can use **explicitly provided phases** from a “phase envelope” → no need to worry about non-stationarity and noise → we can use speech signal models that use both phase and amplitude.
- **Minimal Contamination:** noise is introduced only in phases to reduce the contamination of the (amplitude) spectral envelope.
  - speech sounds more “clear” and “present”.
- Unvoiced phase spectra:
  - phases are uniformly distributed in [0, 2 * pi]
- Voiced phase spectra:
  - Deterministic component: sum-of-sines excitation with some phase dispersion
  - Stochastic component depends on **aperiodicity.**



Vocaine - Deterministic + stochastic phase model 2 / 2

Deterministic phase spectra at RSI:

Sum-of-sines pulse with phase dispersion

$\psi_1 = 0$, always coherent with the RSI

$\psi_k = \frac{\pi}{2} + \text{Uniform}(\pm\frac{\pi}{4})$, for $k = 2, ..., K$

Deterministic + Stochastic phase spectra at RSI

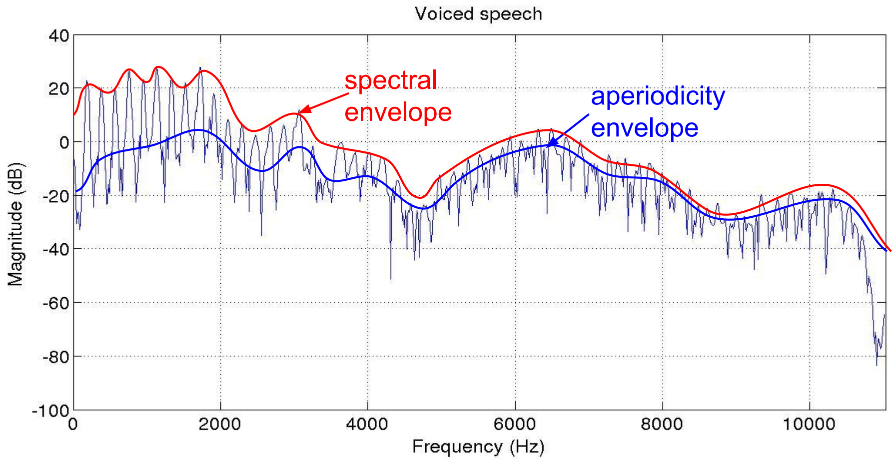
$$\hat{\phi}_k = \psi_k + \text{Uniform}(\pm f(\hat{a}_k)\frac{\pi}{4}), \text{ for } k = 2, ...,$$

Where:

k : index of harmonic

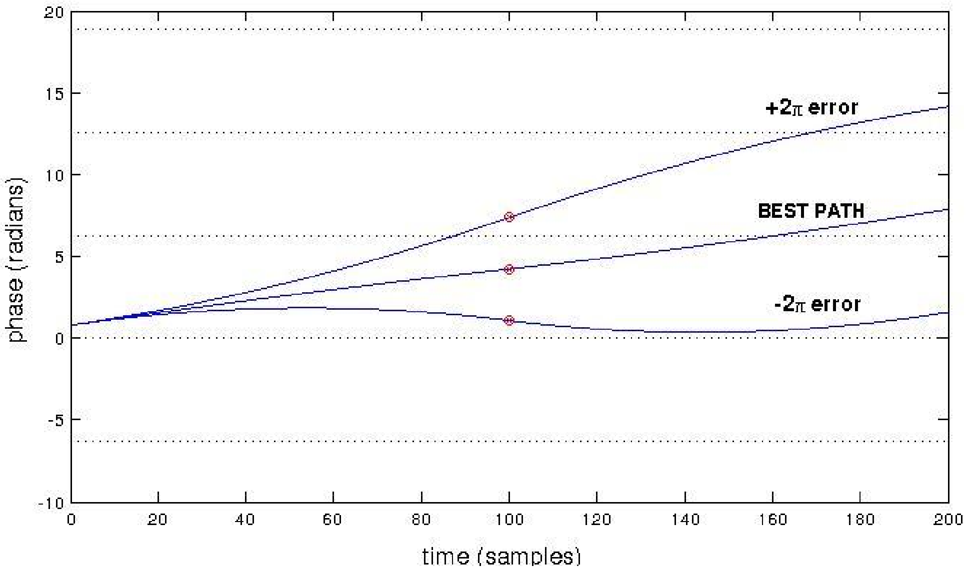
K : number of harmonics

$f(a_k)$: a function of aperiodicity: $[0, 1] \rightarrow [0, 1$



Vocaine - Quadratic phase splines 1 / 5

- **Instantaneous amplitudes & aperiodicities:**
  - Linear interpolation between successive RSIs (piecewise linear spline model).
  - Ignores intermediate frames.
- **Instantaneous phases using a Quadratic Phase Spline Model:**
  - Synthesis period split in two halfs.
  - Uses a quadratic phase model for each half.
  - Corresponds to a piecewise linear frequency model.
  - Mid-period frequency is chosen to maximize smoothness (in the 2-nd derivative sense).
  - very fast: only 2 ADD instructions per harmonic per sample.
  - end-point phases & frequencies are explicitly set.



Vocaine - Quadratic phase splines 2 / 5

Spline Equations:

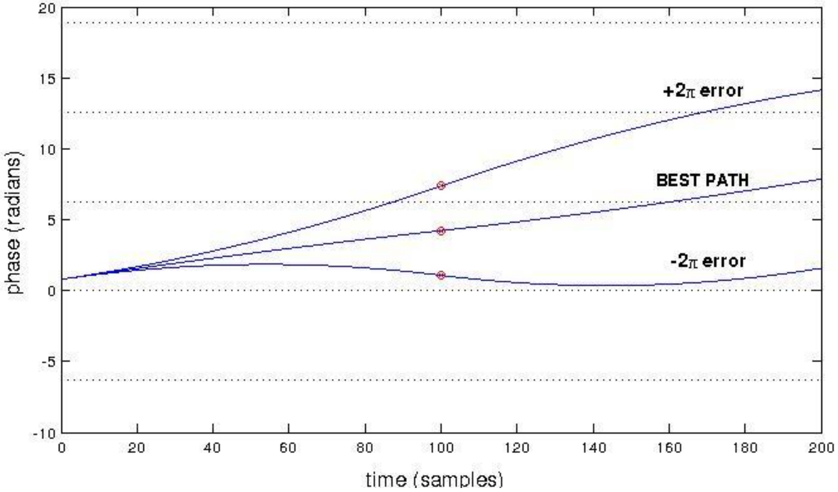
start spline: $\phi_{k,s}(n) = \theta_{k,s} + \omega_{k,s}n + \gamma_{k,s}n^2$

end spline: $\phi_{k,e}(n) = \alpha_{k,s} + \beta_{k,e}(n - n_c) + \gamma_{k,e}(n - n_c)^2 + 2\pi M$

$n \in [0, T]$: time index

$n_c = \lfloor T/2 \rfloor$: break point

M : phase-unwrapping integer

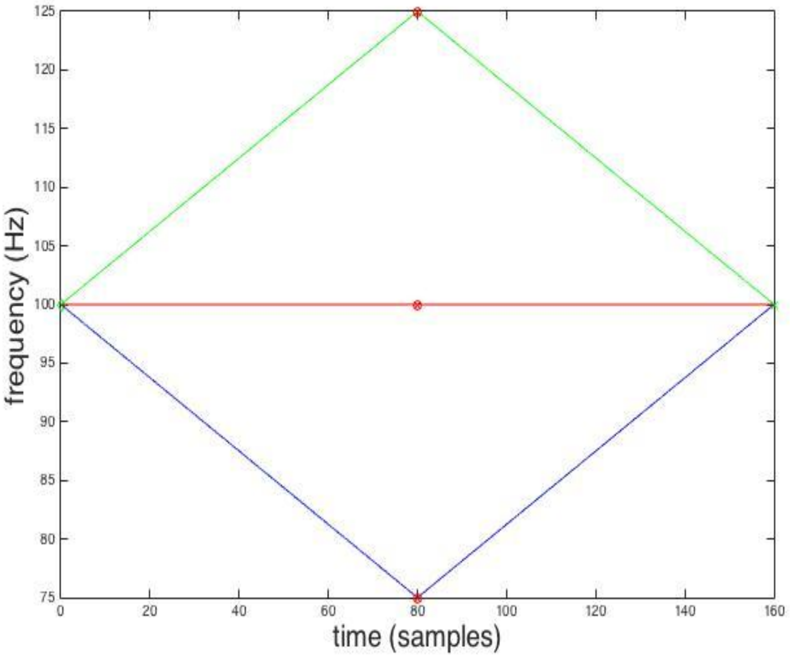


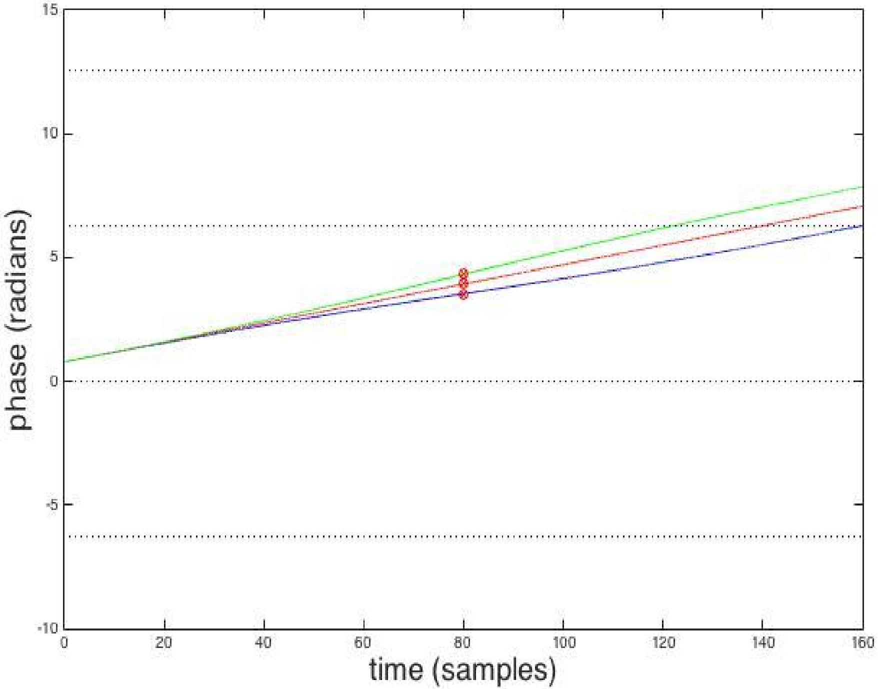
Vocaine - Quadratic phase splines 3 / 5

Constraints:

1. start phase: $\hat{\phi}_{k,s} = \phi_{k,s}(0)$
2. start frequency: $\hat{\omega}_{k,s} = \frac{\partial \phi_{k,s}(n)}{\partial n}$ at $n = 0$
3. end phase: $\hat{\phi}_{k,e} = \phi_{k,e}(T)$
4. end frequency: $\hat{\omega}_{k,e} = \frac{\partial \phi_{k,e}(n)}{\partial n}$ at $n = T$
5. break-point phase continuity: $\phi_{k,s}(n_c) = \phi_{k,e}(n_c)$
6. break-point frequency continuity: $\frac{\partial \phi_{k,s}(n)}{\partial n} = \frac{\partial \phi_{k,e}(n)}{\partial n}$ at $n = n_c$
7. smoothness: $\hat{M} = \left\lfloor \underset{M}{argmin} \left\{ \int_0^{n_c} \left| \frac{\partial^2 \phi_{k,s}(t)}{\partial n^2} \right|^2 dt + \int_{n_c}^{T} \left| \frac{\partial^2 \phi_{k,e}(t)}{\partial n^2} \right|^2 dt \right\} \right\rfloor$

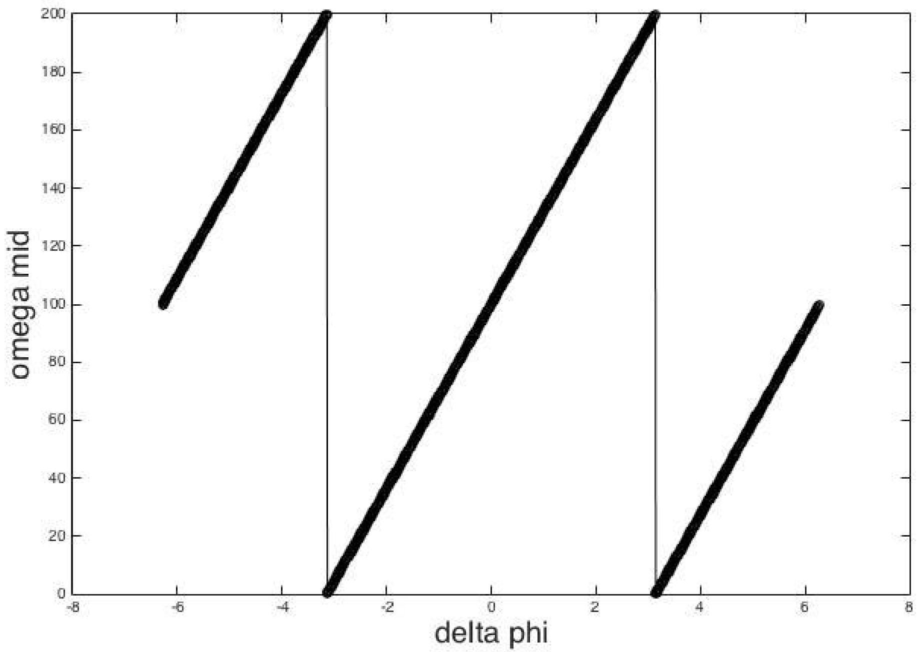
Vocaine - Quadratic phase splines 4 / 5





Vocaine - Quadratic phase splines 5 / 5

- **For aperiodic signals:**
  - sinusoid tracks are not harmonically related.
  - naturally control aperiodicity
- **Quasi-Harmonic model:**
  - Sinusoids are guaranteed to be harmonic only at the pitchmark time-instants.
  - Harmonicity breaks according to noise level (aperiodicity).



Vocaine - Coherent noise modulation model 1 / 3

Vocaine has an explicit frication / aspiration model.

- Aspiration noise in higher frequencies does not sound “incorporated” into the speech signal.
- Vocoders traditionally sound worst in voiced fricatives.
- Some languages like French are very rich in voiced fricatives.
- Voiced fricatives (i.e. /v/, /z/) require a special signal model.
- Same for breathy and laxed speech signals.

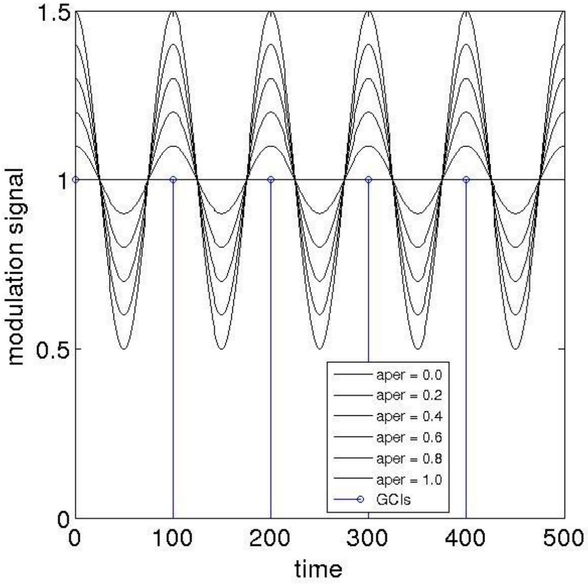
$$s(n) = A_1(n)\cos(\phi_1(n)) + \sum_{k=2}^{K} A_k(n) \underbrace{[\gamma_0 + \gamma_1 \alpha_k(n) \cos(\phi_1(n))]}_{g_k(n)} \cos(\phi_k(n))$$

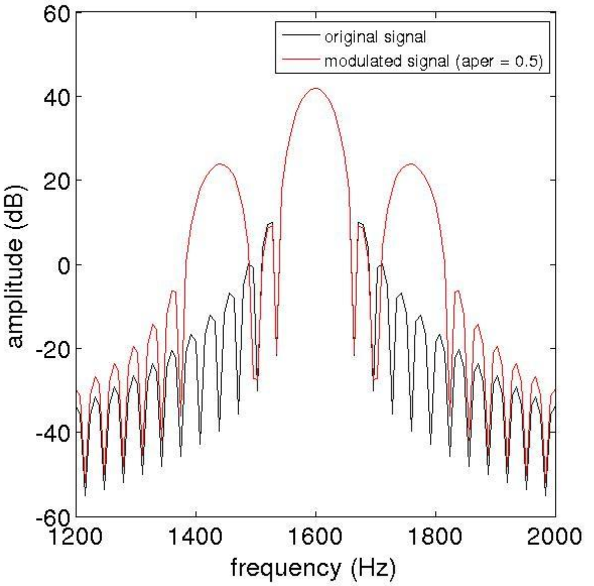
$$\text{coherent noise-modulation: } g_k(n) = \gamma_0 + \gamma_1 \alpha_k(n) \cos(\phi_1(n))$$

Vocaine - Coherent noise modulation model 2 / 3

What does it do?

- **In frequency domain:** convolution spreads the energy of each component.
- **In time domain:** shapes the time-envelope of the noise.
- Frequency-spread and time-modulation becomes stronger with aperiodicity.
- Incorporates noise into the speech signal → **noise is less audible.**
- Simulates aspiration noise patterns of real phonation.





Vocaine - Coherent noise modulation model 3 / 3

- Does it work? → Great improvement in voiced fricatives and breathy phonation. Example: french voice VLF.

References:

- A. McCree, "***A 14 kb/s wideband speech coder with a parametric highband model***", in Proc IEEE Int. Conf. Acoust., Istanbul, 2000, pp. 1153–1156.
- Jan Skoglund and Bastiaan Kleijn, "***On time-frequency masking in voiced speech***," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, July 2000.
- Yannis Agiomyrgiannakis and Yannis Stylianou, "***Combined estimation/coding of highband spectral envelopes for Speech Spectrum Expansion***", ICASSP 2004.
- Pantazis, Yannis, Stylianou, Yannis, "***Improving the modeling of the noise part in the harmonic plus noise model of speech***", ICASSP 2008.
- Yannis Agiomyrgiannakis, and Olivier Rosec. ,"***Towards flexible speech coding for speech synthesis: an LF + modulated noise vocoder.***", ISCA 2008.

Results - Experimental Setup (22050 Hz speech)

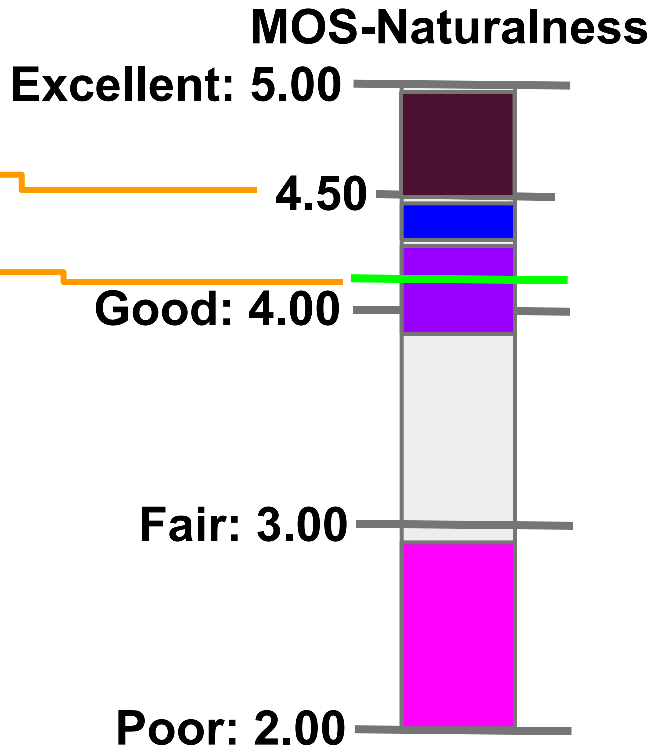
Name	Analysis	Synthesis	#Spectral params	#Aper. params
Embedded+MixedExc (LSP)	MixedExc	Embedded	24	7
STRAIGHT	STRAIGHT	STRAIGHT	1025	513
Vocaine+STRAIGHT	STRAIGHT	Vocaine	1025	513
Vocaine+MixedExc (MCEP)	MixedExc	Vocaine	40	7

Results - Speed - Copy Synthesis

Synthesizer	Median execution time (ms)
Embedded+MixedExc (MCEP)	10150 (100%) ← previous
Vocaine+MixedExc (MCEP)	10264 (101%) ← new

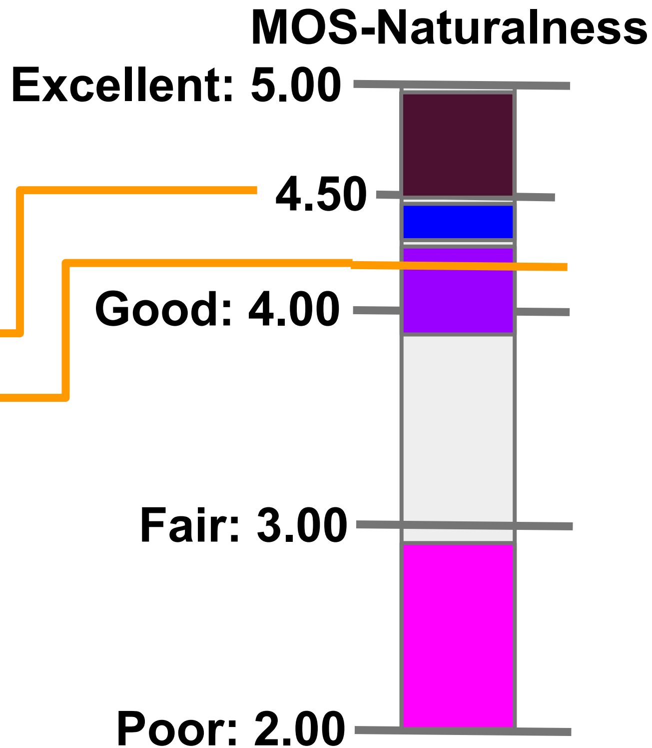
Results - Copy-Synthesis - Quality - English

Recorded Speech	**4.493** ± 0.101
Vocaine+STRAIGHT	**4.144** ± 0.132
Vocaine+MixedExc (MCEP)	**4.079** ± 0.116
STRAIGHT	**4.074** ± 0.126
Embedded+MixedExc (LSP)	**3.699** ± 0.140



Results - Copy-Synthesis - Quality - French

Recorded Speech	**4.568** ± 0.058
Vocaine+STRAIGHT	**4.265** ± 0.073
Vocaine+MixedExc (MCEP)	**4.031** ± 0.076
STRAIGHT	**4.016** ± 0.080
Embedded+MixedExc (LSP)	**3.307** ± 0.106



Results - Quality - Copy Synthesis - Summary

Experiment: Copy-Synthesis

2 MOS tests, 5 English voices (2 males, 3 females), 1 French voice (female):

Summary:

- Original speech MOS: **~4.50**
- STRAIGHT + Vocaine: **~4.20**
- STRAIGHT: **~4.05**
- CODER + Vocaine: **~4.05**
- CODER with SERVER excitation: **~3.710**
- CODER with EMBEDDED excitation: **~3.503**

Improvement in French:

1. Server synthesizer: **0.50 MOS - 0.75 MOS**
2. Embedded synthesizer: **0.7 - 1.0 MOS**

Improvement in English:

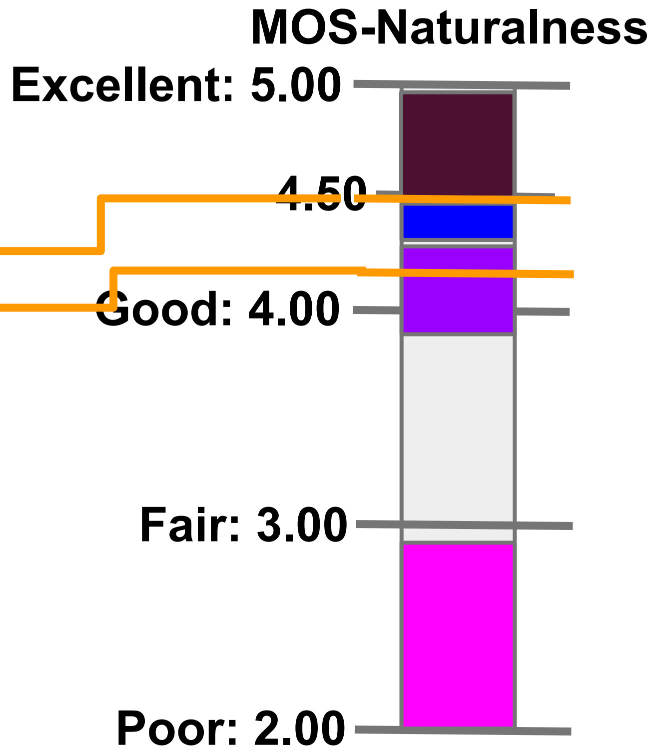
1. Server synthesizer: **0.20 - 0.26 MOS**
2. Embedded synthesizer: **0.38 - 0.45 MOS**

Results - TTS - English

Recorded Speech	**4.529** ± 0.086
Vocaine+STRAIGHT Copy-Synthesis	**4.337** ± 0.094
Vocaine+MixedExc (MCEP) Copy-Synthesis	**4.176** ± 0.114
STRAIGHT Copy-Synthesis	**4.090** ± 0.111
Barracuda Unit-Selection	**3.788** ± 0.128
Manhattan Unit-Selection	**3.773** ± 0.128
Vocaine+MixedExc+LSTM synthesizer	**3.738** ± 0.095
Vocaine+MixedExc+HMM synthesizer	**3.472** ± 0.103
Embedded+MixedExc+HMM synthesizer (LSP)	**3.218** ± 0.112

Results - TTS - French

Recorded Speech	**4.477** ± 0.054
Vocaine+STRAIGHT Copy-Synthesis	**4.209** ± 0.077
Vocaine+MixedExc (MCEP) Copy-Synthesis	**3.958** ± 0.080
STRAIGHT Copy-Synthesis	**3.613** ± 0.087
Vocaine+MixedExc+HMM synthesizer (MCEP)	**3.373** ± 0.154
Embedded+MixedExc+HMM (LSP)	**2.749** ± 0.173



Results - TTS - Summary & Discussion

- Vocaine is significantly better than the state-of-the-art vocoder (STRAIGHT) in copy-synthesis experiment by **~0.2 MOS** for French (richer in voiced fricatives) and **~0.1 MOS** for English. Vocaine shows that it is possible to parameterize the speech signal to a quality level of **~4.20 MOS** without any phase information. The result is both significant and surprising as **4.20 MOS** values were previously only reported when phase information is used.
- Vocaine+HMM synthesizer yields an **~0.350 MOS** improvement over our current baseline for English, significantly narrowing the GAP between HMM-based and unit-selection TTS systems.
- Languages rich in voiced-fricatives which are well modelled by Vocaine benefit significantly more (**+0.625 MOS** points for French).
- **The combination of Vocaine and LSTM statistical mapping with extended input features has matched the performance of a mature unit selection synthesizer.**