LANGUAGE DEPENDENCY OF OBJECTIVE SPEECH QUALITY CRITERIA IN MOBILE NETWORKS

N. Méchergui, F. Ben Ali, S. Larbi and M. Jaïdane
Unité de recherche Signaux et Systèmes
Ecole Nationale d'Ingénieurs de Tunis
Université Tunis-El Manar, Tunisia

E-mail: nader.mechergui@gmail.com, ben-ali-faten@yahoo.fr, sonia.larbi@enit.rnu.tn, meriem.jaidane@planet.tn

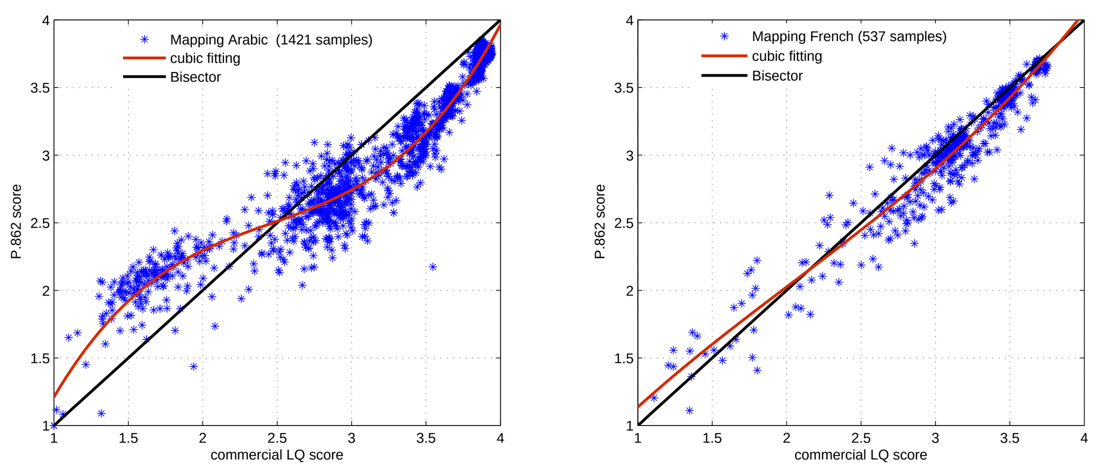
Introduction

- Main motivation
  - Non adequacy of quality assessment algorithms to Arabic speech
  - Which language feature is responsible for this dependency?
- T/F non stationarity measure of languages:
  - The stationarity index (SI)
  - Language dependency of the local T/F characteristics
- Effect of arbitrary frame by frame analysis on the T/F speech content: Case of PESQ

- Commercial objective quality assessment criteria has shown language dependency when used in actual mobile communications network. To analyze this dependency, we focus on time/frequency analysis of speech and we show that different languages have different "non stationary" behavior.

Main motivation

Commercial LQ scores vs. PESQ P.862: mapping for different languages



Arabic samples mapping — French samples mapping

=> Language dependent behavior

Test procedure: measurements in the mobile network of Tunisiana, one ref. speech sample per language (male/female speakers, 6s)

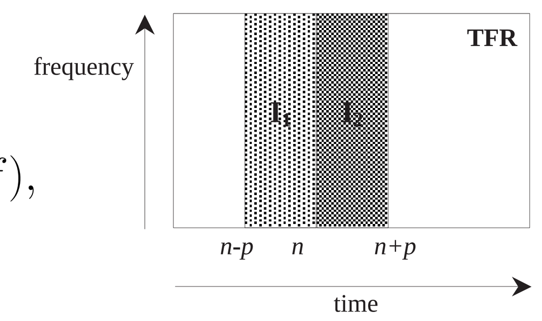
Which language feature is responsible for language dependency?

A feature which discriminates between languages:

- Linguistic approach [Grabe et al., 2002]: rhythmic language classification (stressed, syllable and mora-timed), based on *isochronously repeated rhythmic units*
- Statistical-linguistic approach [Ramus et al., 1999]: Statistics of vocalic-consonantal intervals duration(%V, Δ C)
- Signal processing approach: Voiced-unvoiced transitions detected and measured in the time-frequency domain.

T/F stationarity measure of languages: stationarity indices SI

- Sliding sub-images I_1 and I_2:



$$I_1(n;\tau,f) = TFR(n-p+\tau,f),$$
$$I_2(n;\tau,f) = TFR(n+\tau,f).$$

Sliding step $\tau \in [0,p]$ and p is a sensitivity parameter.

- Normalized sub-images NI_1 and NI_2

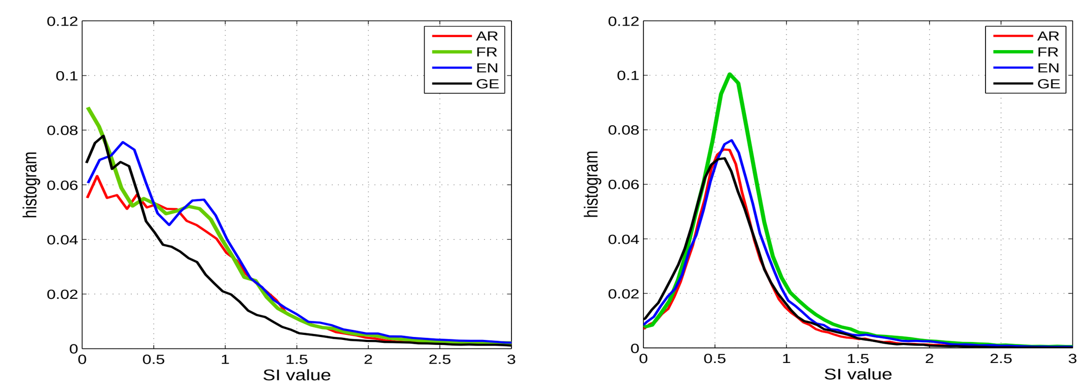
$$NI_k(n;\tau,f) = \frac{|I_k(n;\tau,f)|}{\int_{\tau=0}^{p}\int_{-\infty}^{+\infty}|I_k(n;\tau,f)|df d\tau} \quad k=1,2$$

- Küllback Distance between sub-images:

$$SI_{ku}(n) = \int_{\tau=0}^{p}\int_{-\infty}^{+\infty}(NI_1(n;\tau,f)-NI_2(n;\tau,f))\log\left(\frac{NI_1(n;\tau,f)}{NI_2(n;\tau,f)}\right)df d\tau$$

Variability of the local T/F speech content: Histograms of stationarity indices

- Languages have different SI histograms: bimodality for English, flatness for Arabic, rather unimodal for French



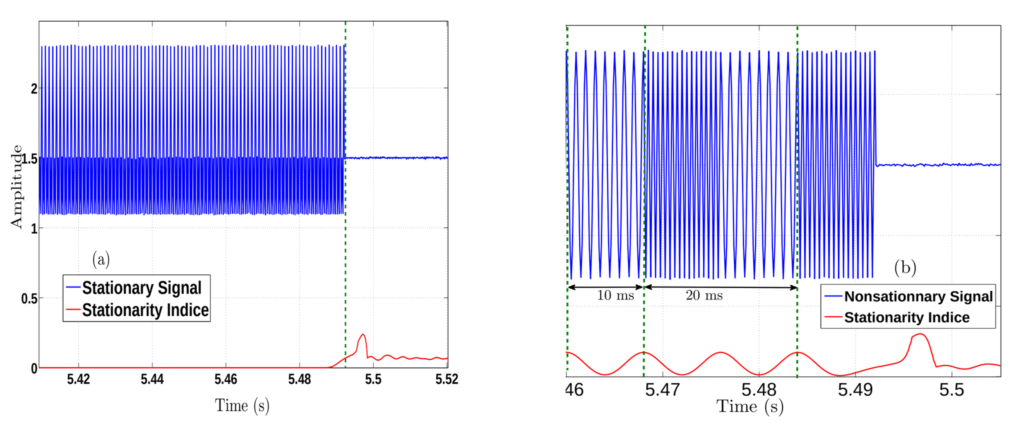
UIT-T P.50 original speech samples — stationarized(LPC modeled)samples

⇒ Languages have different non stationary behavior

- Histograms of stationarized languages show all the same unimodal behavior: Differences in the non stationarity characteristics between languages are reduced

Compromise between frame size and stationarity of signals

SI of test signals

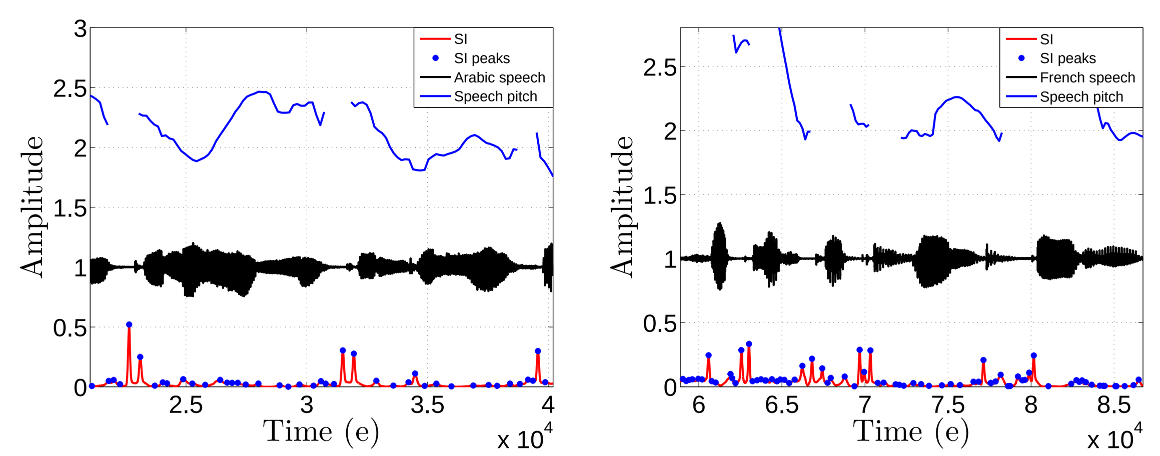


(a) Sum of 2 sinusoids (stationary) — **(b)** 2 concatenated sinusoids (non stationary)

- Signal (a) is stationary: use of a large analysis frame
- Signal (b) is non stationary (SI peaks): frame size depends on stationary segment duration

Optimal frame size = Distance between SI peaks

Speech test material for frame size optimization



Pitch and Stationarity Indices of an Arabic and a French sentence

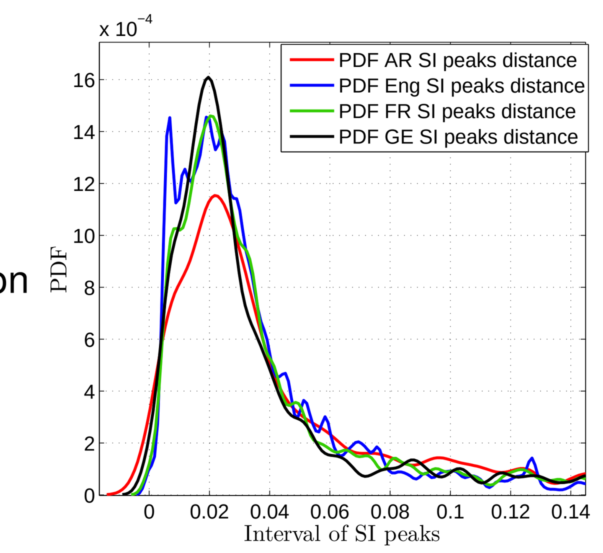
We compute the distance between SI peaks:

- 16 sentences (8s) in Arabic, French, German and English
- Speech Database: ITU-T P.50 (F_s=16 kHz, 16 bits)

SI threshold $= 0.02$ => Voiced/Unvoiced transition

Optimal analysis frame size for different languages

- 20 ms frame size suitable for AR, GE, FR
- ENG case: 2 frame duration - 10ms and 20ms - seems to be suitable



⇒ Optimal analysis frame size is 20ms as usually stated

⇒ For some languages, a variable frame size should be used

Conclusions

- Many speech processing systems are based on signal stationarity over 20 ms analysis frames
- This work confirms the analysis frame size of 20 ms (usually stated) for FR, GE and AR
- Some languages, like English, have a different period of stationarity: 10ms and 20ms
- A variable analysis frame size would enhance speech processing and reduce the effect of language dependency (as the example of AAC-coder for Music coding)

2èmes Journées Perception Sonore (JPS), December 10-12, 2012 • Laboratoire de Mécanique et d'Acoustique (LMA-CNRS), Marseille, France