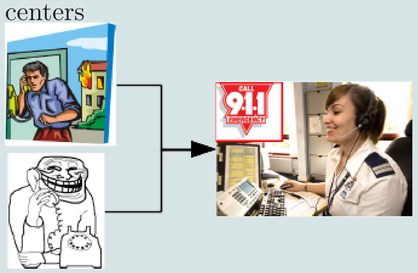
Automatic Emotion Recognition in Compressed Speech Using Acoustic and Non-Linear Features

N. Garcia[1], J.C Vásquez-Correa[1], J.D Arias-Londoño[1], J.F Várgas-Bonilla[1] and J.R Orozco-Arroyave[1,2]
nicanor.garcia@udea.edu.co,
[1] Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia
[2] Friedrich-Alexander-Universität, Erlangen, Germany.

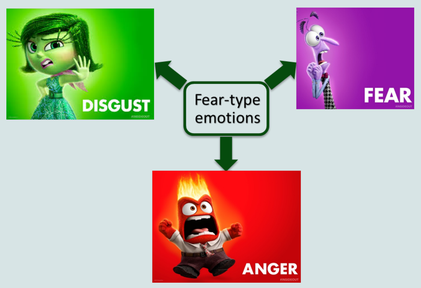
Outline

- Automatic emotion recognition can be applied in call center scenarios to:
  - Evaluate the quality of the service provided in commercial call centers.

  - Discriminate real emergency calls from pranks or diversions in emergency call centers



- The most relevant emotions to recognize in these applications are Fear-type emotions.

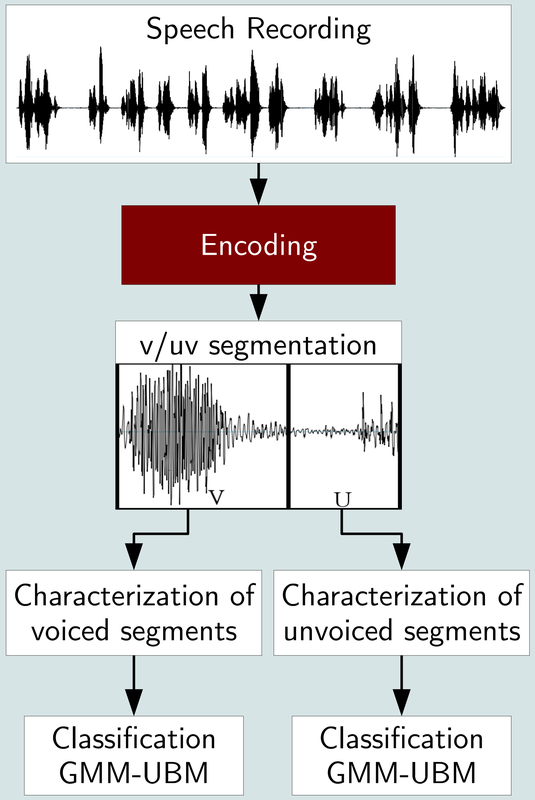


- The audio of incoming calls is compressed by different encoding schemes.

Aim of this study

- Perform automatic emotion recognition on speech compressed by different encoding techiniques.

Methodology



Databases

- Berlin:
- enterface05:

Encoding

Evaluated codecs:

- G.722, used in LAN VoIP
- G.726, used in International Calls
- AMR-NB, used in mobile phone networks
- GSM-EFR, used in mobile phone networks
- AMR-WB, used in modern mobile phone networks
- SILK, used by Skype
- Opus, used by WebRTC (Google, Facebook).

Characterization

Voiced frames features:

- 12 MFCC
- Noise measures: HNR, GNE and NNE.
- Non-Linear Dynamics: correlation dimension, the largest Lyapunov exponent, the Hurst exponent, and the Lempel-Ziv Complexity

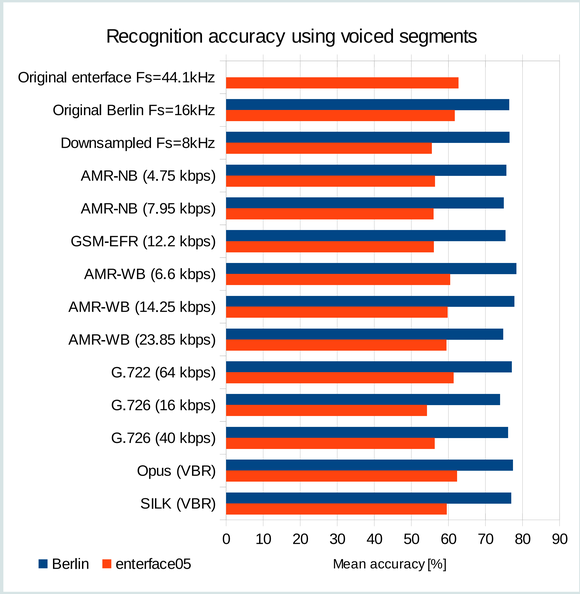
Unvoiced segments features:

- 12 MFCC
- Energy content over 25 bands in Bark scale

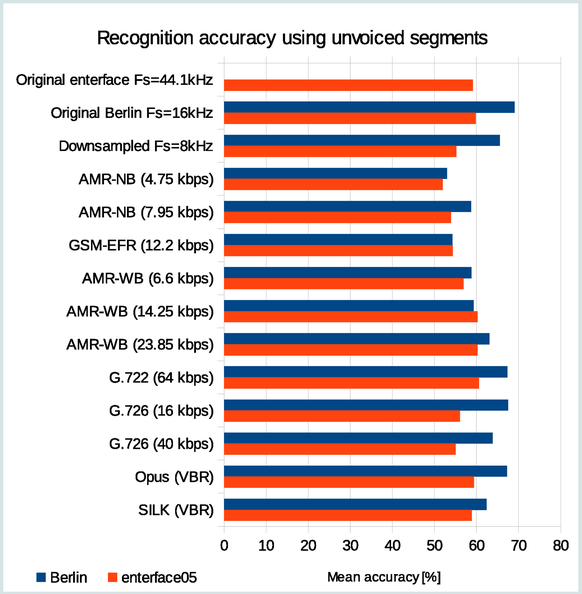
Classification

- Features extracted from voiced and unvoiced segments are classified separately.
- The classifier used is a Gaussian Mixture model adapted from a Universal backgroud model (GMM-UBM), using a Maximum A Posteriori (MAP) rule to derive a speaker-specific GMM from the UBM.

Results voiced



Results unvoiced



Conclusions

- The compression by encoding produces different effects according to the type of segments used:
  - Voiced frames: Little to no degradation of accuracy.
  - Unvoiced frames: Considerable degradation of accuracy.
- The bandwidth of the signal is a significant factor in the recognition accuracy.
- The bit-rate also affects the recognition accuracy.
- Future work should address other distortions generated by the communications channel.