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

N. Garcia¹, J.C Vásquez-Correa¹, J.D Arias-Londoño¹, J.F Vargas-Bonilla¹ and J.R Orozco-Arroyave¹,²
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.

Methodology

- Berlin:
  - enterface05:

   - 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

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

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.