Misperceptions of the emotional content of natural and vocoded speech in a car

Jaime Lorenzo-Trueba[1], Cassia Valentini-Botinhao[2], Gustav Eje Henter[1], Junichi Yamagishi[1,2]

[1] National Institute of Informatics, Tokyo, Japan,
[2] The University of Edinburgh, Edinburgh, United Kingdom

1 - Motivation

- Speech synthesis has advanced remarkably recently
  - With naturalness almost comparable to human speech!
- But **generation of speech in adverse environments** still requires fundamental research
- Noise is known to degrade speech intelligibility, and is compensated automatically by humans with **Lombard speech**

Hypotheses

- **Emotional communication capabilities** also degrade with noise

Objective

- Validate by means of a perceptual evaluation in an accurate adverse environment simulation if the hypotheses holds

2 - Emotional speech corpus

- Female professional Japanese voice actress recorded in a high quality studio environment
- Three emotion pairs (happy - sad, calm - insecure, excited - angry) and neutral speech, 1200 sentences per emotion
- The utterances' text presents no emotional context, and the speaker was controlled to produce stable emotional strength

---

- Recordings took place in a noiseless environment
  - The adverse environmental conditions had to be simulated

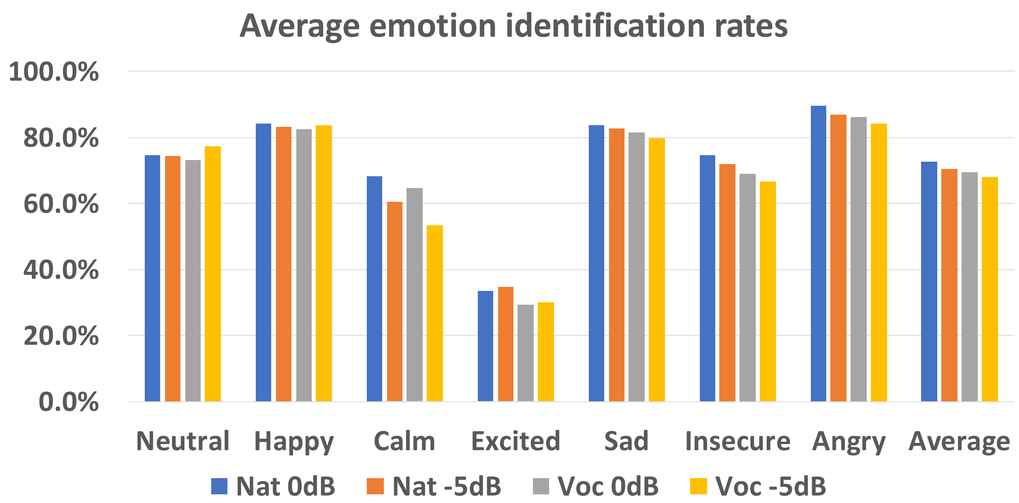
3- Generating emotional speech under adverse conditions

- We wanted to reproduce an environment as realistic as possible. So we recorded **real in-car noise**
- Recordings were carried out using a **binaural microphone in a head-and-torso mannequin** in the front passenger seat of a hybrid car
- Both a **city route** and a **highway route** were considered, both in different environmental conditions (open/closed windows, competing male speaker)
- We also recorded the **room impulse response** of the car so that we could **properly simulate the reverberation** of the environment

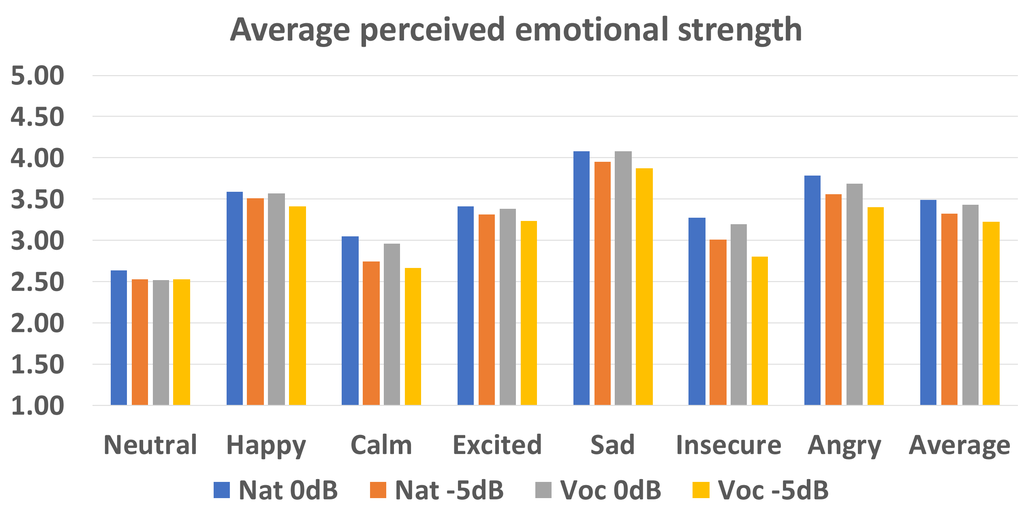
4 - Experimental evaluation

- To validate the hypotheses we carried out a crowdsourced perceptual evaluation:
  - 414 native listeners with driving license
  - 63% female, 37% male
  - Ages ranging between 18 and 74 years old
- All seven emotions, for a total 8400 natural speech samples and 700 additional high quality vocoded speech samples randomly selected from the database
- Two simulated SNR conditions: -5 dB and 0 dB and five environmental conditions. **91,000 data points** were evaluated
- Both emotion identification rates (EIR) and perceived emotional strength (ES, Likert scale) were evaluated

5 - Evaluation Results



- As expected, there is a **significant decrease** in recognition rates between the 0 dB and -5 dB SNR conditions
- The effect is stronger when talking about vocoded speech



- There is a similarly **significant degradation in emotional strength perception** between the different noise conditions.
- Neutral speech is not perceived as more (or less) neutral, though!

Results in terms of listener's factors

- **Listener age** plays the most significant role in **both recognition rates and perceived emotional strength**:
  - Young people show 4% better average EIR
  - **Elder people** perceive emotions as **less expressive** (-0.3 average)
- Listener gender shows a lesser but also significant impact:
  - **Females show better recognition capabilities** (1% higher EIR)
  - And also tend to perceive them as **more intense** (+0.1 ES)
- A combined analysis of **gender and age** show that their **interaction is also significant:**
  - Young females show 12% better average EIR than elder males!

6 - Conclusions and future work

- The evaluation has proven that **noise plays a significant role in emotional communication degradation**, as in intelligibility
- This affects **low-arousal emotions (sad, calm, insecure)**, with high arousal emotions being more robust in this environment
- Age and gender have also shown to be relevant for emotion recognition, with **male elder people showing significantly worse recognition rates than young females** (12% in average)
- **Vocoded speech** not only reproduces this effect **but makes it more significant**

ACKNOWLEDGEMENTS: The work presented in this poster was partially supported by Toyota Motor Corporation.