GlobalSIP 2017
November 14-16, 2017
Montréal, Canada

Influence of Audio Bandwidth Reduction on Speech Emotion Recognition by Human Subjects

Presenter:
Philippe Gournay

Authors:
Olivier Lahaie
Roch Lefebvre
Philippe Gournay

Department of Electrical and Computer Engineering
Université de Sherbrooke, Québec, Canada

Motivations

- **Bandwidth limitation** is a critical step in the design of a speech coder
- Strong limitation is known to degrade **subjective** speech quality, intelligibility
- It has also been shown to degrade performance of **automatic** speaker identification, speech recognition, emotion recognition
- What effect does it have on **emotion recognition by human subjects?**
- What are the **implications** for the design of future speech coders?

Previous work

- A. Albahri and M. Lech, "Effects of band reduction and coding on speech emotion recognition," ICSPCS 2016
- Effect measured using an **automatic** system (feature extraction followed by a classifier)
- Audio bandwidths **did not correspond to standard telephony bandwidths** and were **limited** to a maximum of 8kHz

Contributions

- A **subjective evaluation** procedure to measure both **accuracy and effort** of **voice emotion recognition** by **human subjects**
- An application to **standard telephony bandwidths**
- An **analysis of the result** to confirm or invalidate the degradation observed using an automatic classifier

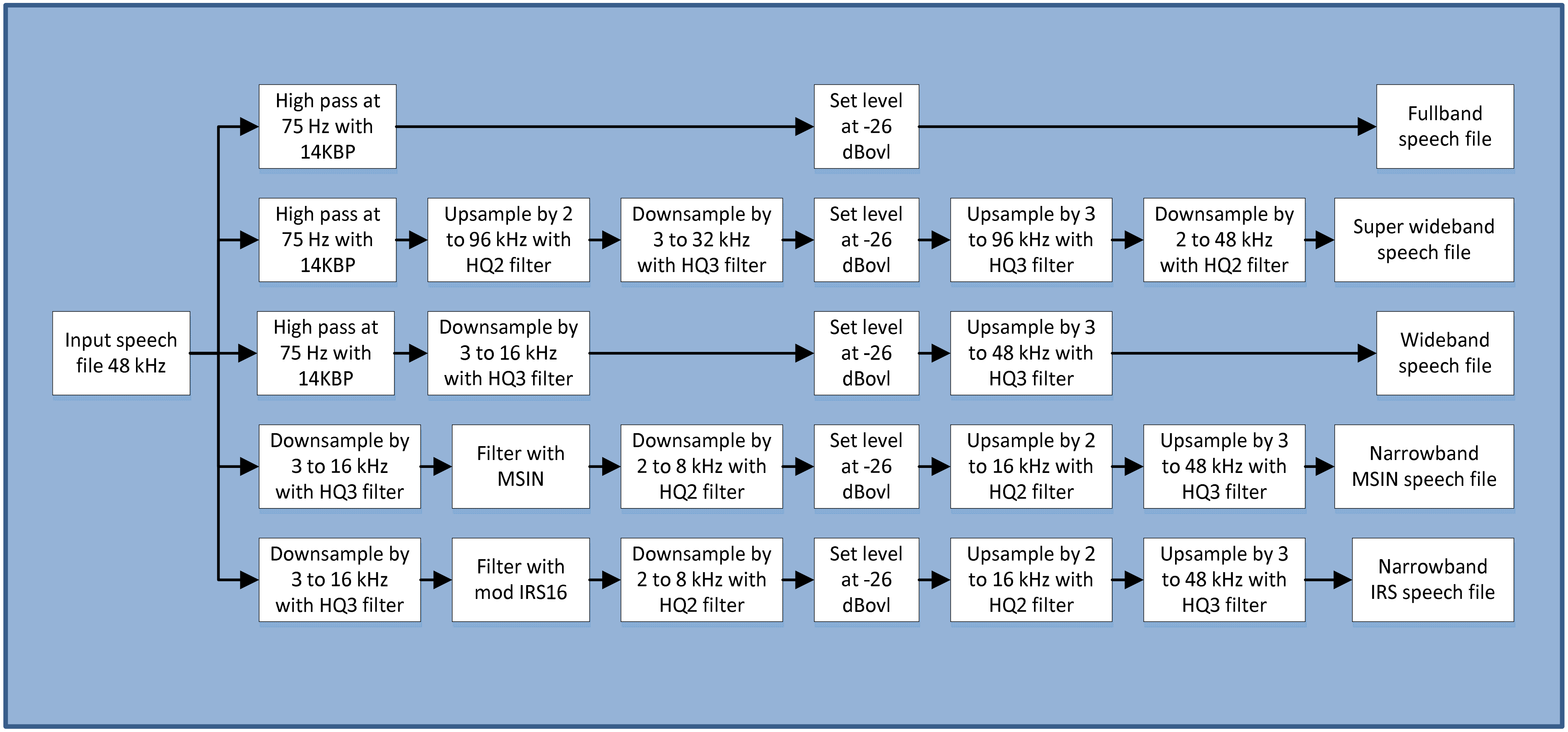
Test material

- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
  - Two semantically neutral sentences
  - 24 actors (12 male, 12 female)
  - Seven different emotions (happiness, sadness, anger, fear disgust, surprise, calm) plus neutral
  - Two emotional intensities (normal and strong)
  - Each combination repeated twice
- Item selection
  - The "calm" emotion was ignored
  - The "normal" intensity was used
  - A **complete and balanced** (across gender, actors, sentences) subset (table in the paper)

Test conditions

- Five **standard telephony bandwidths**
  - Narrowband IRS
  - Narrowband MSIN
  - Wideband
  - Superwideband
  - Fullband
- Obtained using combinations of **standard ITU tools**
- Correspond to **processing plans** used when selecting and characterizing speech coding standards
- The resulting **audio signal conditioning** is close to what can be observed in speech codec implementations

Test conditions



Input speech file 48 kHz

High pass at 75 Hz with 14KBP → Set level at -26 dBovl → Fullband speech file

High pass at 75 Hz with 14KBP → Upsample by 2 to 96 kHz with HQ2 filter → Downsample by 3 to 32 kHz with HQ3 filter → Set level at -26 dBovl → Upsample by 3 to 96 kHz with HQ3 filter → Downsample by 2 to 48 kHz with HQ2 filter → Super wideband speech file

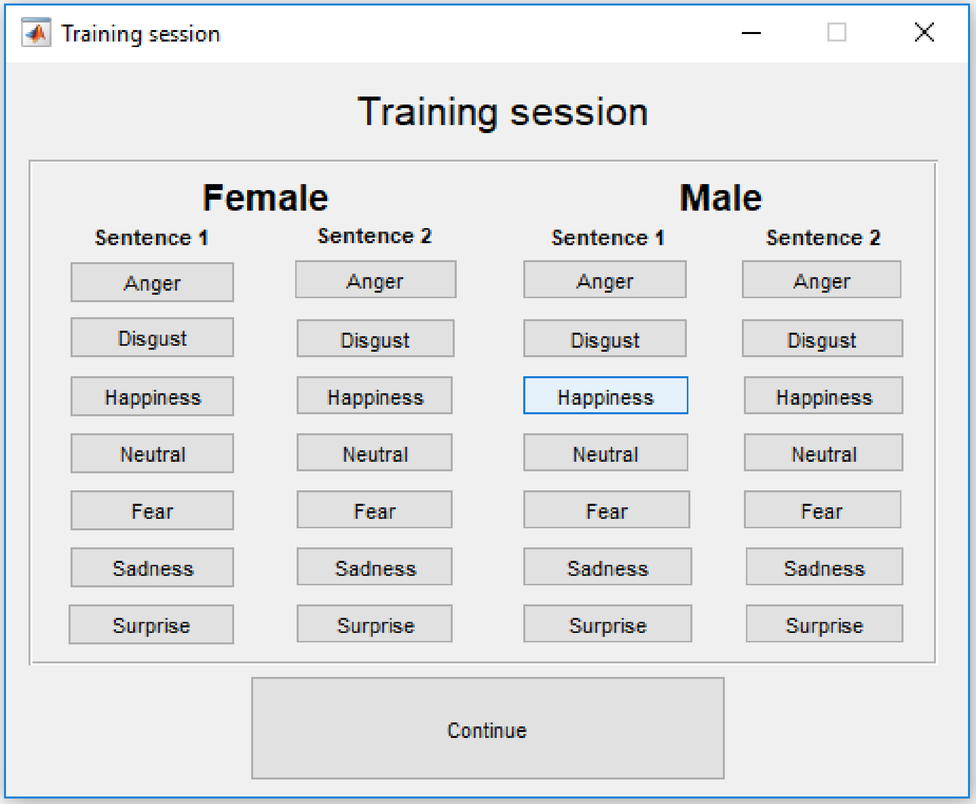
High pass at 75 Hz with 14KBP → Downsample by 3 to 16 kHz with HQ3 filter → Set level at -26 dBovl → Upsample by 3 to 48 kHz with HQ3 filter → Wideband speech file

Downsample by 3 to 16 kHz with HQ3 filter → Filter with MSIN → Downsample by 2 to 8 kHz with HQ2 filter → Set level at -26 dBovl → Upsample by 2 to 16 kHz with HQ2 filter → Upsample by 3 to 48 kHz with HQ3 filter → Narrowband MSIN speech file

Downsample by 3 to 16 kHz with HQ3 filter → Filter with mod IRS16 → Downsample by 2 to 8 kHz with HQ2 filter → Set level at -26 dBovl → Upsample by 2 to 16 kHz with HQ2 filter → Upsample by 3 to 48 kHz with HQ3 filter → Narrowband IRS speech file

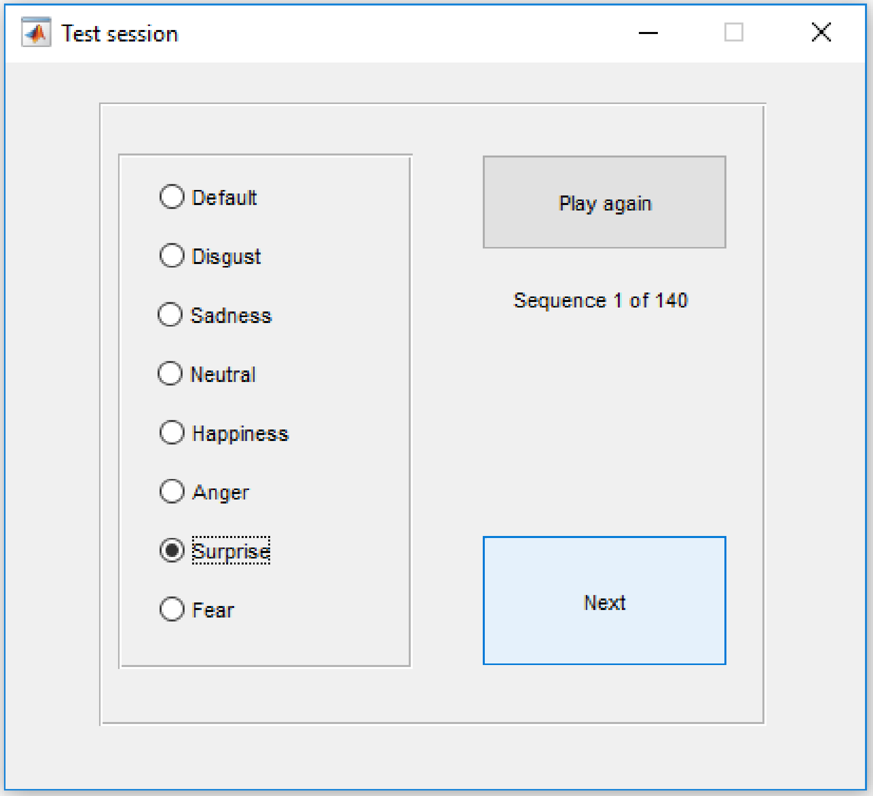
1. Training session

- To familiarize subjects with the **structure and content** of the test material (uses fullband version)



2. Test session

- 7×4×5 = 140 test items covering **all relevant conditions** presented in a **randomized order**



○ Default
○ Disgust
○ Sadness
○ Neutral
○ Happiness
○ Anger
● Surprise
○ Fear

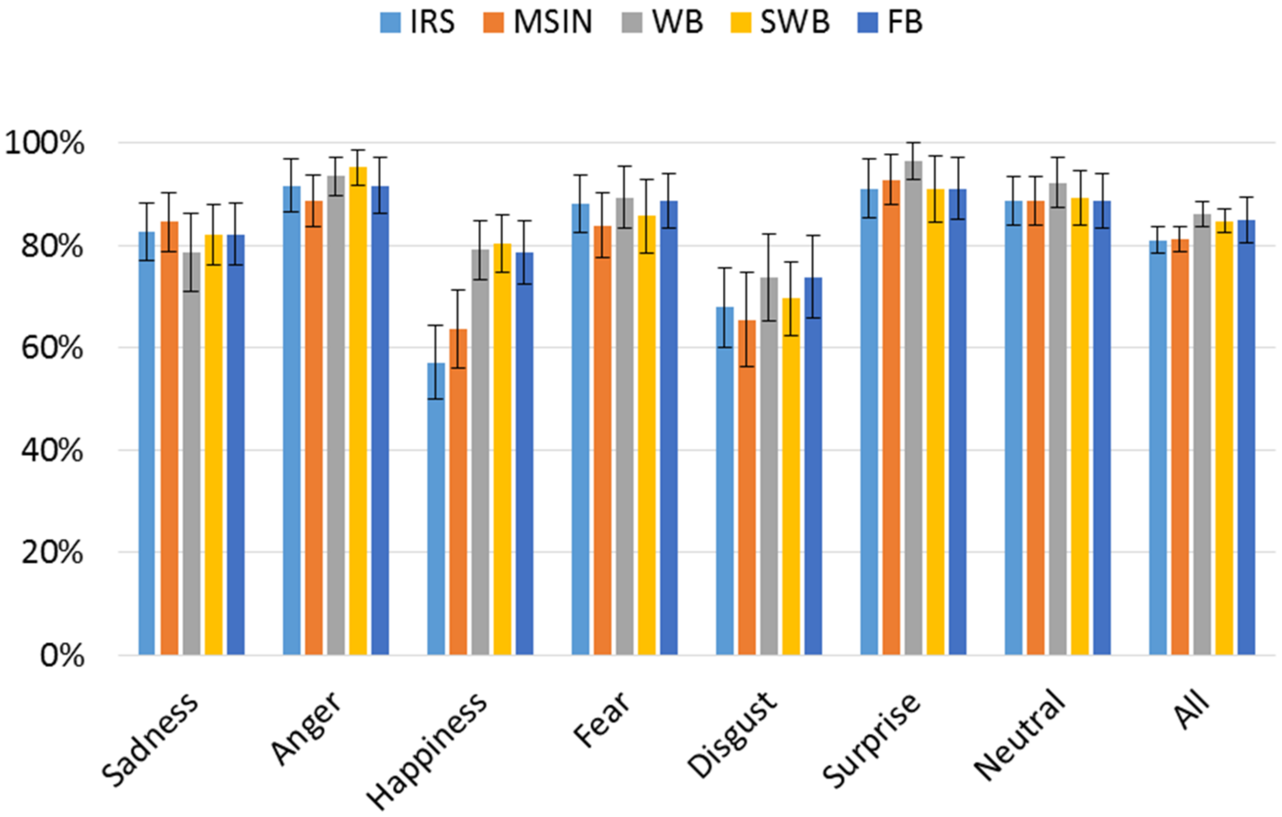
Sequence 1 of 140

Test subject and equipment

- 42 normal-hearing listeners
  - 34 male, 8 female
  - 19 to 48 years old
- Beyerdynamic DT770 headphone
- Rega EAR headphone amplifier
- Performance measured in terms of:
  - Recognition accuracy
  - Number of listenings before taking a decision (listeners were **not aware** of this)

Recognition accuracy

- 42×4 = 168 data points per emotion and condition
- 95% confidence intervals



Recognition accuracy

- No statistically significant difference could be observed between the **fullband, superwideband and wideband** conditions
- No remarkable trend either for any emotion **except for "Happiness"**
- In that case, **accuracy drops** from around 80% for fullband, superwideband and wideband down to 63% for narrowband MSIN and 58% for narrowband IRS
- This degradation is **statistically significant** at the 5% significance level

Detailed results for Happiness

- Same trend can be observed for all four "Happiness" stimuli

	FB	SWB	WB	MSIN	IRS
Female 1	86%	93%	83%	71%	67%
Female 7	55%	52%	57%	36%	33%
Male 2	83%	88%	88%	71%	64%
Male 11	90%	88%	88%	76%	64%

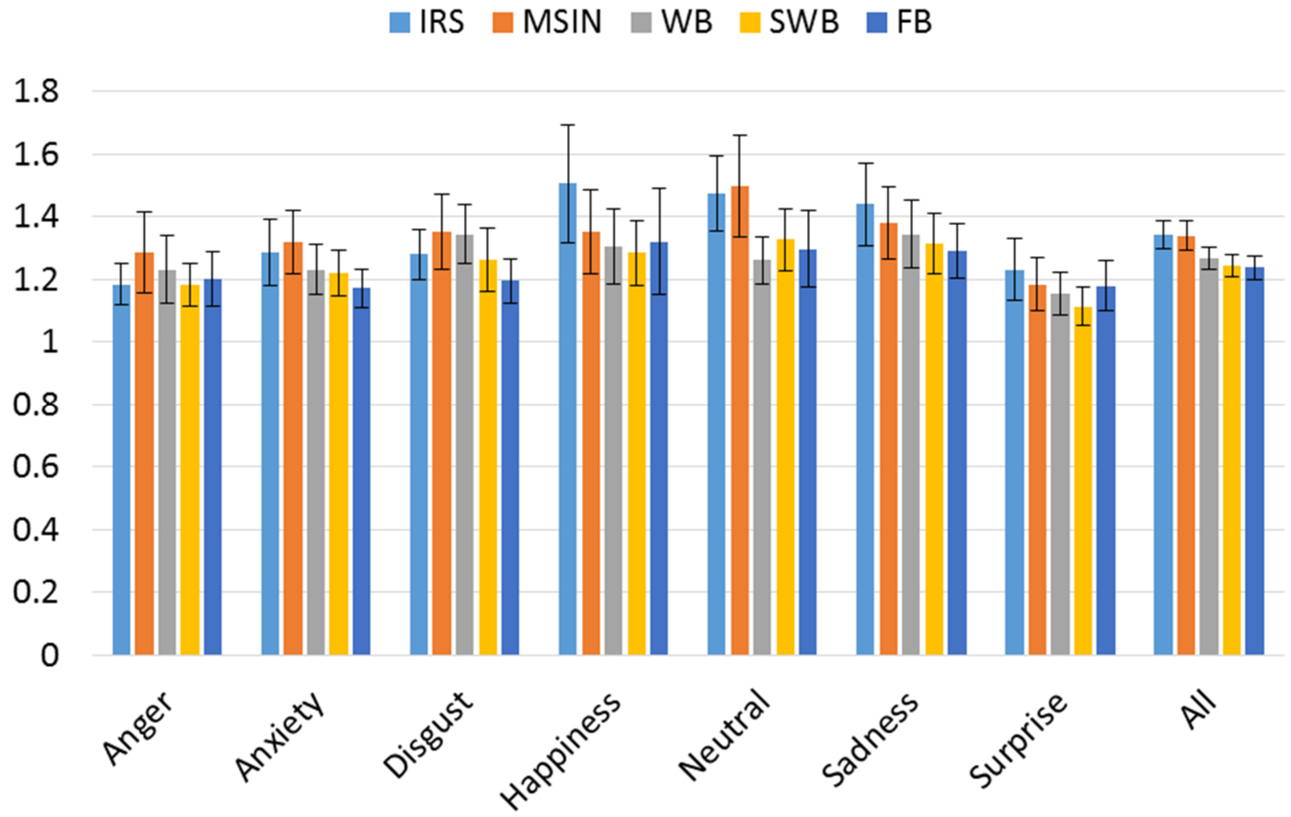
Example of confusion matrix

- Most frequent confusions: "Anger" for "Disgust", "Neutral" or "Fear" for "Happiness"

	Narrowband IRS condition						
	Happiness	Sadness	Anger	Fear	Disgust	Surprise	Neutral
Happiness	96	1	4	20	12	8	27
Sadness	0	139	0	12	6	0	11
Anger	0	0	154	3	7	2	2
Fear	0	11	2	148	0	7	0
Disgust	2	10	31	7	114	2	2
Surprise	1	0	1	5	2	153	6
Neutral	1	15	0	0	3	0	149

Number of listenings

- 168 data points per condition and emotion
- 95% Confidence intervals



Number of listenings

- A clear trend, where the **number of listenings decreases as the audio bandwidth increases**, can be observed for all emotions except "Anger" and "Disgust"
- The observed differences may not be statistically significant when emotions are considered individually…
- … but some are (at the 5% significance level) when all emotions are considered together: the number of listenings is significantly higher for **narrowband IRS and MSIN** than for **superwideband or fullband**

Conclusions

- **Subjective evaluation** of the effect of **bandwidth limitation** on the perception of **speech emotions**
- Several **standard telephony bandwidths** (narrowband, wideband, superwideband and fullband)
- In some cases (specifically, "Happiness") the **recognition accuracy decreases** with the audio bandwidth
- More importantly, **the number of listenings** before subjects made a decision **increases** as bandwidth decreased
- Bandwidth limitation may therefore result in a fatigue for the listener

Perspectives

- Determine why some emotions are more sensitive than others
- Investigate how well **artificial bandwidth extension** techniques **preserve** (or can be used to **restore**) the emotional content in the upper part of the speech spectrum

Thank you!