Speech Intelligibilty

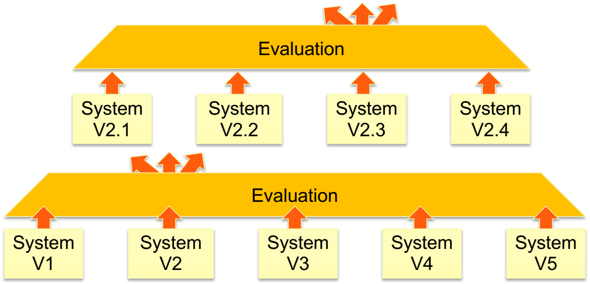
TTS evaluations

1

Evaluation : why, how

- **Why** we need to evaluate
  - diagnostic test to guide future development
  - comparative test against another system, or a baseline
  - pass/fail test for a product release
- **When** to evaluate
  - individual components (during development) *-or-* the finished system ?
- **Which** aspects to evaluate
  - intelligibility, naturalness, speaker similarity, ...
- **How** to evaluate
  - listener task, test design, materials used, objective measures, ...
- **What** to do with the outcome

2

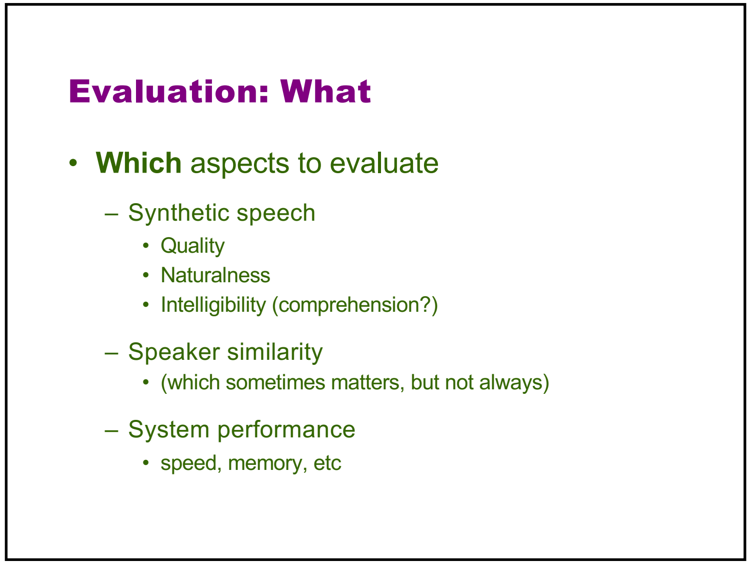


3

Evaluation : When

- During development
  - isolated components
  - components working within a complete system
- After building a complete system
  - **pass/fail** - does commercial product meet user or market requirements
  - **cross-system comparisons**
    - control certain components, such as
      - a common database
      - fixed annotation and label alignments
      - common front end

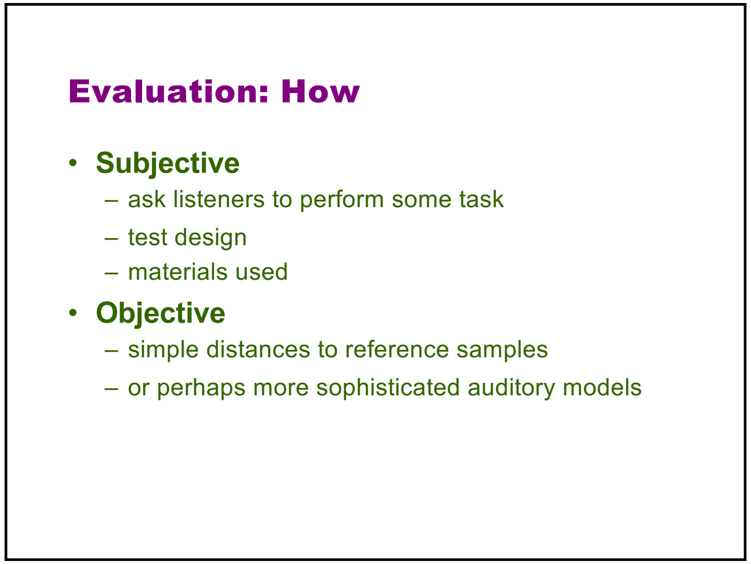
4



Evaluation: What

- **Which** aspects to evaluate
  - Synthetic speech
    - Quality
    - Naturalness
    - Intelligibility (comprehension?)
  - Speaker similarity
    - (which sometimes matters, but not always)
  - System performance
    - speed, memory, etc

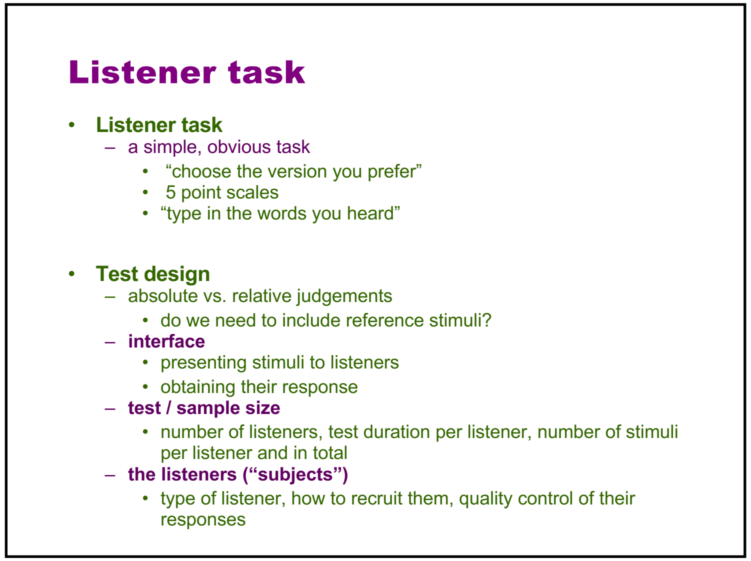
5



Evaluation: How

- **Subjective**
  - ask listeners to perform some task
  - test design
  - materials used
- **Objective**
  - simple distances to reference samples
  - or perhaps more sophisticated auditory models

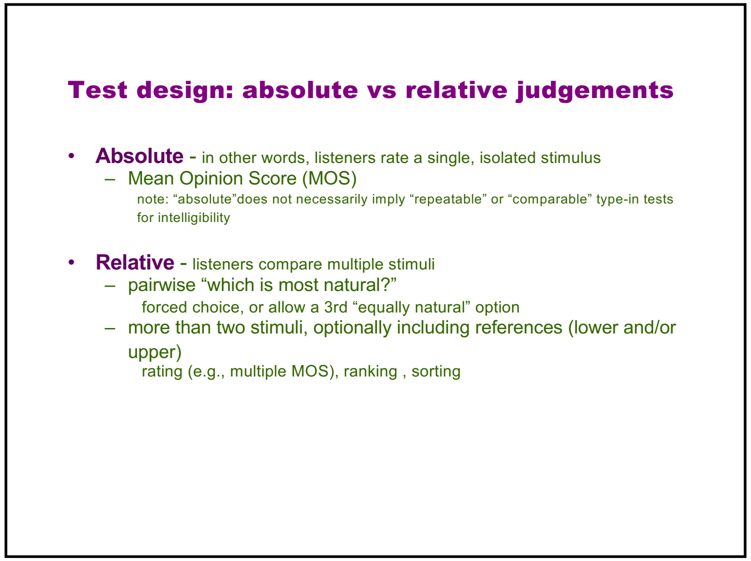
6



Listener task

- **Listener task**
  - a simple, obvious task
    - “choose the version you prefer”
    - 5 point scales
    - “type in the words you heard”
- **Test design**
  - absolute vs. relative judgements
    - do we need to include reference stimuli?
  - **interface**
    - presenting stimuli to listeners
    - obtaining their response
  - **test / sample size**
    - number of listeners, test duration per listener, number of stimuli per listener and in total
  - **the listeners (“subjects”)**
    - type of listener, how to recruit them, quality control of their responses

7



Test design: absolute vs relative judgements

- **Absolute** - in other words, listeners rate a single, isolated stimulus
  - Mean Opinion Score (MOS)

    note: “absolute”does not necessarily imply “repeatable” or “comparable” type-in tests for intelligibility
- **Relative** - listeners compare multiple stimuli
  - pairwise “which is most natural?”

    forced choice, or allow a 3rd “equally natural” option
  - more than two stimuli, optionally including references (lower and/or upper)

    rating (e.g., multiple MOS), ranking , sorting

8

Test design: size

- Sample size determines statistical power
  - Find significant differences
  - *see a course of statistics !*

Main idea (own experience)

- maximum test duration 30-45 minutes, for paid or *motivated* listeners in a controlled environment
  - much less for online (remote) listeners
- at least 10-20 listeners, and preferably more
- as many different sentences as possible

9

Participants (listeners)

- All listeners should hear exactly the same stimuli randomly
- Each stimulus should be presented once (repetition should be handled carefully)
- Cannot participate in the same evaluation twice.

10

Presentation: intelligibility

- **Normal sentences/word**
  - tend to get a ceiling effect, due to interference from semantics (predictability)
- **Semantically Unpredictable Sentences**
  - e.g.,“The unsure steaks overcame the zippy rudder”
  - e.g., « je n'entends point insulter la raison. »

  ➔ not representative of actual system usage, but avoids ceiling effect
- **Minimal pairs**
  - specific to individual phonemes
  - very time consuming and therefore rarely used
- **Add noise**

11

Objective evaluation

- **Simple objective measures**
- Compare acoustic properties to a natural reference sample
  - assumes that natural version is the ‘gold standard’
- Time-align natural and synthetic
  - perform frame-by-frame comparison, sum up local differences
- Does not account for natural variation (could use multiple natural examples)

Objective evaluation

- Based only on properties of the signal
  - spectral envelope: Mel-Cepstral Distortion (MCD)
  - F0 contour: Root Mean Square Error of F0 (RMSE F0) and/or correlation

which do not correlate perfectly with human perception

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred,i} - y_i)^2}{n}}$$

12

Audiovisual Speech Intelligibility

13

Audiovisual speech intelligibility: Why?

- To assess the quality of audiovisual speech synthesis
- To understand the mechanism of audiovisual intelligibility
- What makes an audiovisual message intelligible?
- How does the visual component contribute to the audiovisual perception?
- How to assess the audiovisual intelligibility?
- How to compare the individual differences in speech intelligibility of different talkers?

14

Motivations

- Measuring audiovisual intelligibility
- The importance of the audiovisual speech in language learning
- The benefit of viewing internal articulators (i.e., tongue, palate, etc.)
- The influence of faces and voices on audiovisual intelligibility

Measuring AV Intelligibility

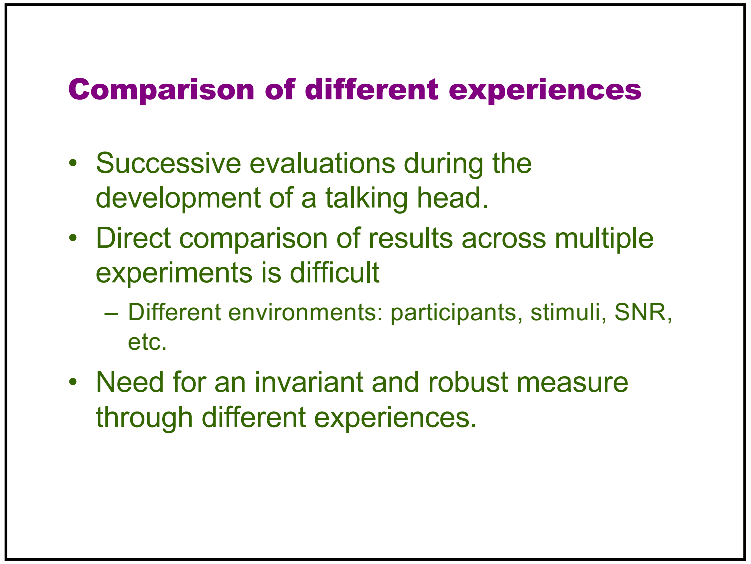
- Comparing the intelligibility of a talking head relatively to a reference
- Direct comparisons across different sessions of experiments
- Measures the benefit of a synthetic face in comparison with the natural face.

15

General outline of evaluating talking heads

- At least two conditions:
  - Unimodal Audio
  - Bimodal audiovisual
- Presentation of a set of sentences (words, syllables, etc.) in a noisy environment.
- The same acoustic signal is used in both conditions.
- Choice of noise level:
  - Strong enough, to make the acoustic signal difficult to understand, but audible.
  - Pre A pre-test may be necessary.
- Recognition task by participants

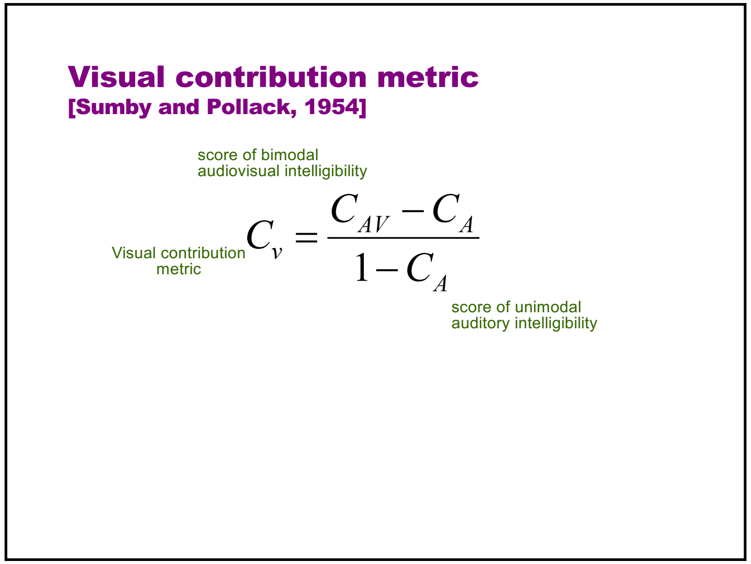
16



Comparison of different experiences

- Successive evaluations during the development of a talking head.
- Direct comparison of results across multiple experiments is difficult
  - Different environments: participants, stimuli, SNR, etc.
- Need for an invariant and robust measure through different experiences.

17



Visual contribution metric
[Sumby and Pollack, 1954]

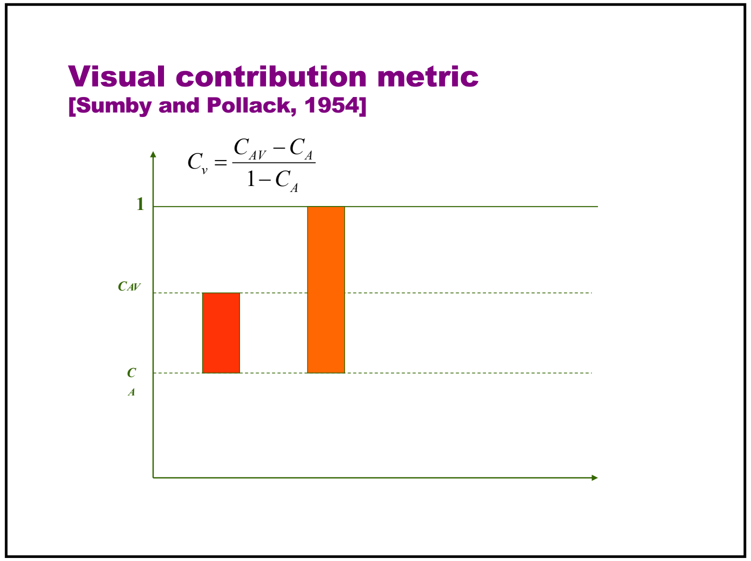
score of bimodal
audiovisual intelligibility

Visual contribution
metric

$$C_v = \frac{C_{AV} - C_A}{1 - C_A}$$

score of unimodal
auditory intelligibility

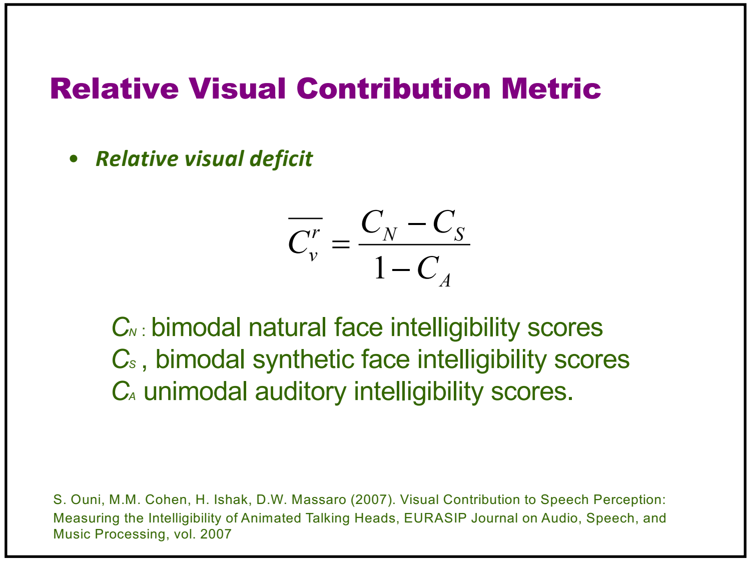
18



Visual contribution metric
[Sumby and Pollack, 1954]

$$C_v = \frac{C_{AV} - C_A}{1 - C_A}$$

19



Relative Visual Contribution Metric

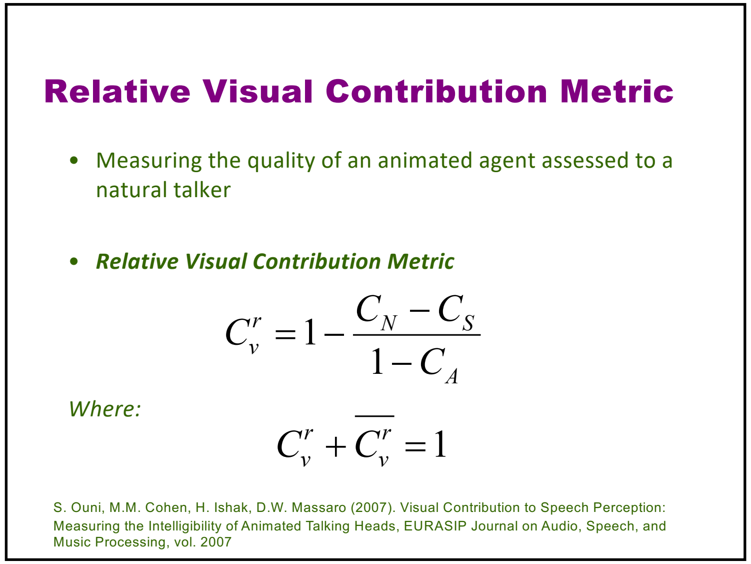
- ***Relative visual deficit***

$$\overline{C_v^r} = \frac{C_N - C_S}{1 - C_A}$$

C_N : bimodal natural face intelligibility scores
C_S , bimodal synthetic face intelligibility scores
C_A unimodal auditory intelligibility scores.

S. Ouni, M.M. Cohen, H. Ishak, D.W. Massaro (2007). Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2007

20



Relative Visual Contribution Metric

- Measuring the quality of an animated agent assessed to a natural talker
- ***Relative Visual Contribution Metric***

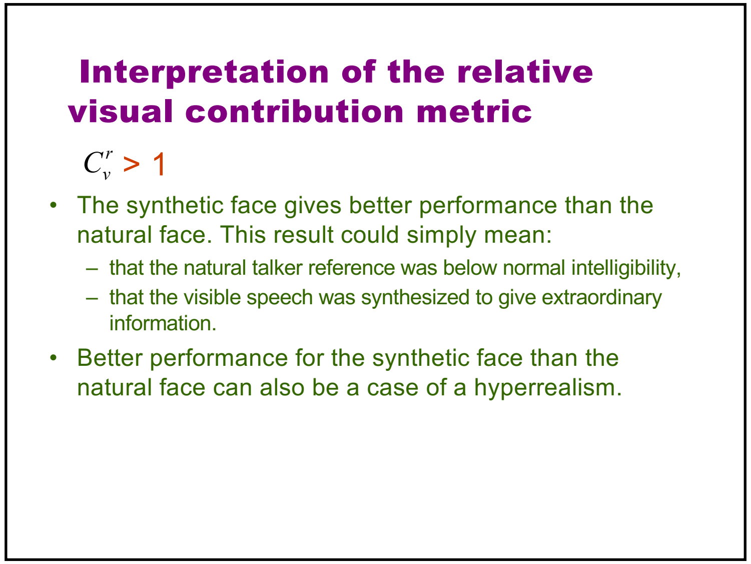
$$C_v^r = 1 - \frac{C_N - C_S}{1 - C_A}$$

Where:

$$C_v^r + \overline{C_v^r} = 1$$

S. Ouni, M.M. Cohen, H. Ishak, D.W. Massaro (2007). Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2007

21

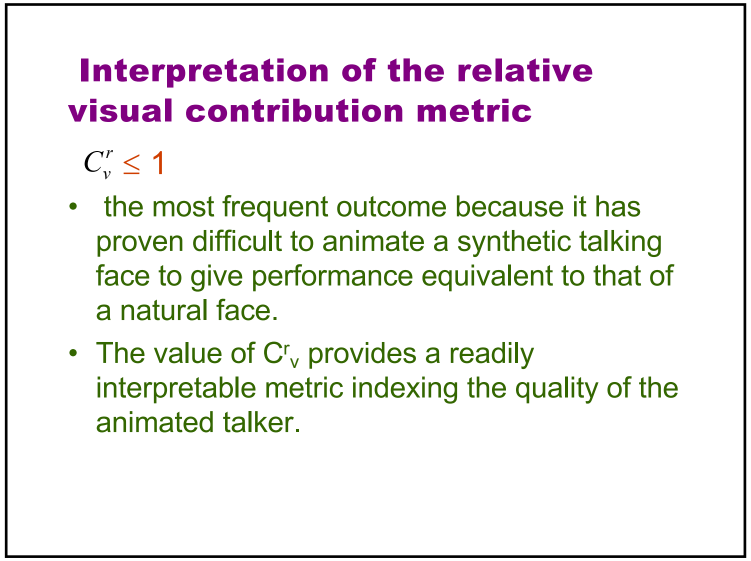


Interpretation of the relative visual contribution metric

$C_v^r > 1$

- The synthetic face gives better performance than the natural face. This result could simply mean:
  - that the natural talker reference was below normal intelligibility,
  - that the visible speech was synthesized to give extraordinary information.
- Better performance for the synthetic face than the natural face can also be a case of a hyperrealism.

22

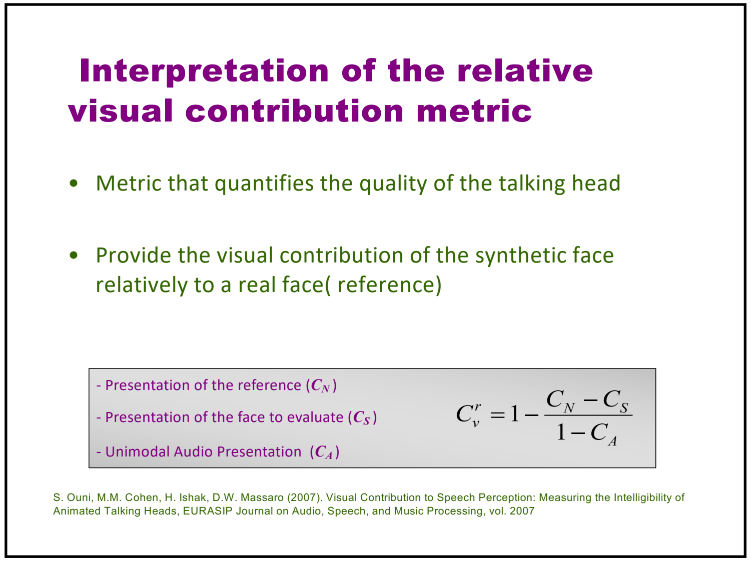


Interpretation of the relative visual contribution metric

$C_v^r \leq 1$

- the most frequent outcome because it has proven difficult to animate a synthetic talking face to give performance equivalent to that of a natural face.
- The value of C_v^r provides a readily interpretable metric indexing the quality of the animated talker.

23



Interpretation of the relative visual contribution metric

- Metric that quantifies the quality of the talking head
- Provide the visual contribution of the synthetic face relatively to a real face(reference)

- Presentation of the reference (C_N)
- Presentation of the face to evaluate (C_S)
- Unimodal Audio Presentation (C_A)

$$C_v^r = 1 - \frac{C_N - C_S}{1 - C_A}$$

S. Ouni, M.M. Cohen, H. Ishak, D.W. Massaro (2007). Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2007

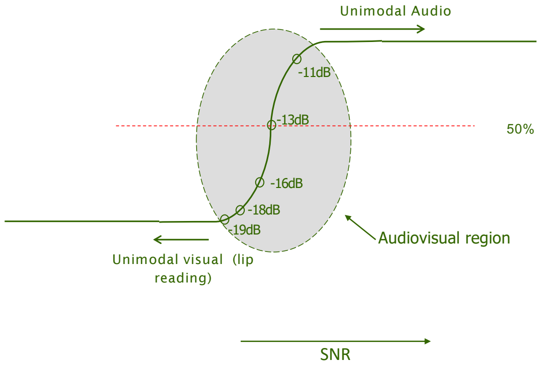
24

Using the relative visual contribution metric in perceptive experiments

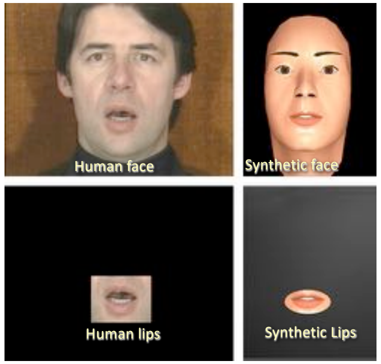
- 3 perceptual experiments in noisy environment.
- Differents modalites
- Audio modality :
  - Same acoustic signal
- 5 noise leves (SNR: -11dB, -13dB, -16dB, -18dB et –19 dB)
- Stimuli : CV (consonant-vowel)

  C = {/f/, /p/, /l/, /s/, /ʃ/, /t/, /θ/, /r/, /w/}
  V = {/a/, /i/, /u/}
  random presentation

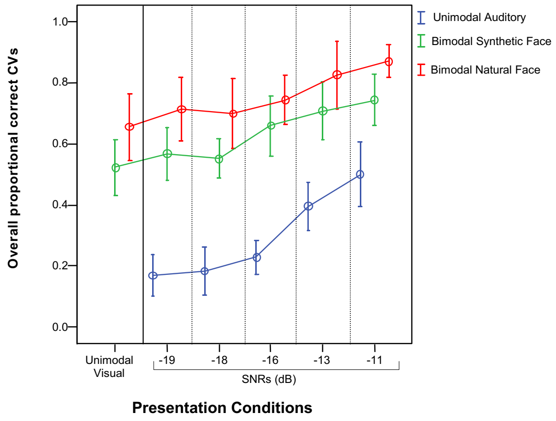
Noise Level



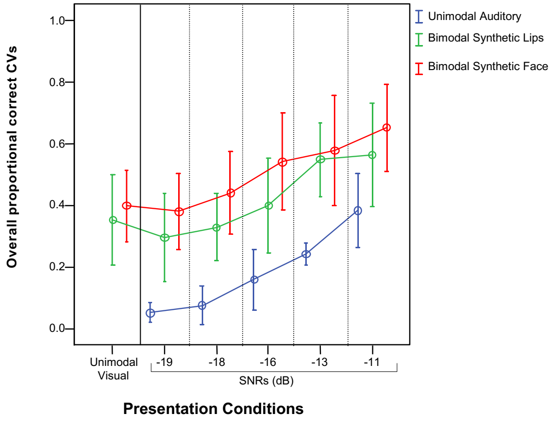
Visual Modalities



Experience 1

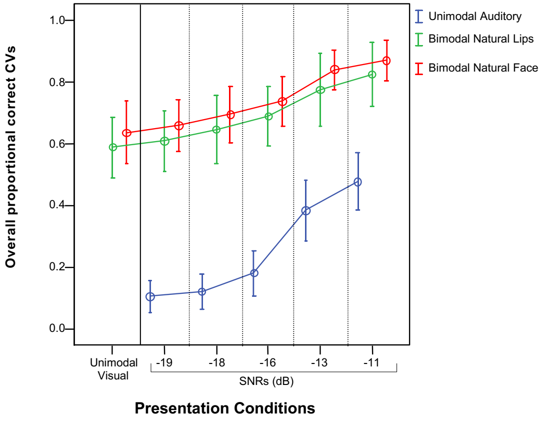


Experience 2



29

Experience 3



30

Contribution visuelle (CV)
Contribution visuelle relative (CVR)

SNR	-19dB	-18dB	-16dB	-13dB	-11dB
S&P CV – Visage synthétique	0.53	0.49	0.58	0.57	0.56
S&P CV - Visage humain	0.66	0.64	0.67	0.71	0.76
CVR	**0.87**	**0.85**	**0.90**	**0.86**	**0.80**

SNR	-19dB	-18dB	-16dB	-13dB	-11dB
S&P CV - Lèvres du visage humain	0.26	0.27	0.29	0.40	0.29
S&P CV - Visage humain	0.35	0.39	0.46	0.44	0.44
CVR	**0.74**	**0.66**	**0.64**	**0.68**	**0.57**

SNR	-19dB	-18dB	-16dB	-13dB	-11dB
S&P CV - Lèvres du visage synthétique	0.56	0.60	0.62	0.64	0.66
S&P CV - Visage synthétique	0.62	0.65	0.68	0.74	0.75
CVR	**0.92**	**0.92**	**0.93**	**0.87**	**0.87**

31

Evaluation:
Visual Contribution (CV)
Relative Visual Contribution (CVR)

- **Analysis of variance**
  - (Participant scores, experiences and noise level)
- **S & P Visual Contribution (SP-CV)**
  - SP-CV varies significantly depending on the noise level.
  - ➔ SP-CV is not independent of noise.
- **Relative Visual Contribution (CVr)**
  - CVr did not vary with noise, and there was no interaction between noise level and experiments.
  - ➔ CVR measurement is independent of noise.

32

Evaluation of an bimodal unit-selection base Audiovisual TTS

Perceptual evaluation

- Audiovisual presentation
- 50 french words (e.g. anneau, bien, chance, cuisine)
  - In-corpus words (original : 39 words)
  - Out-of-corpus words (synthesis : 11 words)
- + 2 white noises (6 dB, 10 dB)
- Human recognition task
- 34 Participants: 15 females and 24 males, aged 19 to 65 (average of 30.5 years, SD = 10.97)

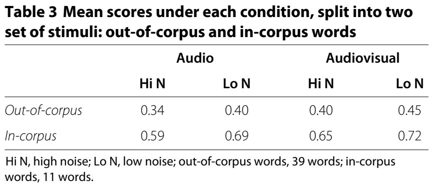


Table 3 Mean scores under each condition, split into two set of stimuli: out-of-corpus and in-corpus words

	Audio		Audiovisual	
	Hi N	**Lo N**	**Hi N**	**Lo N**
Out-of-corpus	0.34	0.40	0.40	0.45
In-corpus	0.59	0.69	0.65	0.72

Hi N, high noise; Lo N, low noise; out-of-corpus words, 39 words; in-corpus words, 11 words.

Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M. O., & Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 16.

33

Evaluation of an bimodal unit-selection base Audiovisual TTS

Subjective evaluation (MOS) *mean opinion score*

- 20 synthesized audiovisual sentences (without any added noise)
  - 7 *in-corpus*
  - 13 out-of-corpus
- 34 participants: rate each sentence by answering 5 questions.

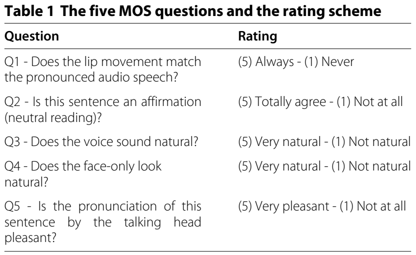


Table 1 The five MOS questions and the rating scheme

Question	Rating
Q1 - Does the lip movement match the pronounced audio speech?	(5) Always - (1) Never
Q2 - Is this sentence an affirmation (neutral reading)?	(5) Totally agree - (1) Not at all
Q3 - Does the voice sound natural?	(5) Very natural - (1) Not natural
Q4 - Does the face-only look natural?	(5) Very natural - (1) Not natural
Q5 - Is the pronunciation of this sentence by the talking head pleasant?	(5) Very pleasant - (1) Not at all

Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M. O., & Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 16.

34

Evaluation of an bimodal unit-selection base Audiovisual TTS

- **Subjective evaluation (MOS)**

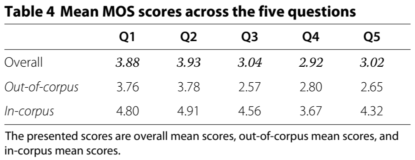


Table 4 Mean MOS scores across the five questions

	Q1	Q2	Q3	Q4	Q5
Overall	*3.88*	*3.93*	*3.04*	*2.92*	*3.02*
Out-of-corpus	3.76	3.78	2.57	2.80	2.65
In-corpus	4.80	4.91	4.56	3.67	4.32

The presented scores are overall mean scores, out-of-corpus mean scores, and in-corpus mean scores.

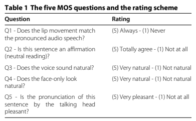


Table 1 The five MOS questions and the rating scheme

Question	Rating
Q1 - Does the lip movement match the pronounced audio speech?	(5) Always - (1) Never
Q2 - Is this sentence an affirmation (neutral reading)?	(5) Totally agree - (1) Not at all
Q3 - Does the voice sound natural?	(5) Very natural - (1) Not natural
Q4 - Does the face-only look natural?	(5) Very natural - (1) Not natural
Q5 - Is the pronunciation of this sentence by the talking head pleasant?	(5) Very pleasant - (1) Not at all

Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M. O., & Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 16.

35

Expressivity

Example : Perceptual evaluation of the quality of the expressivity of an audiovisual corpus.

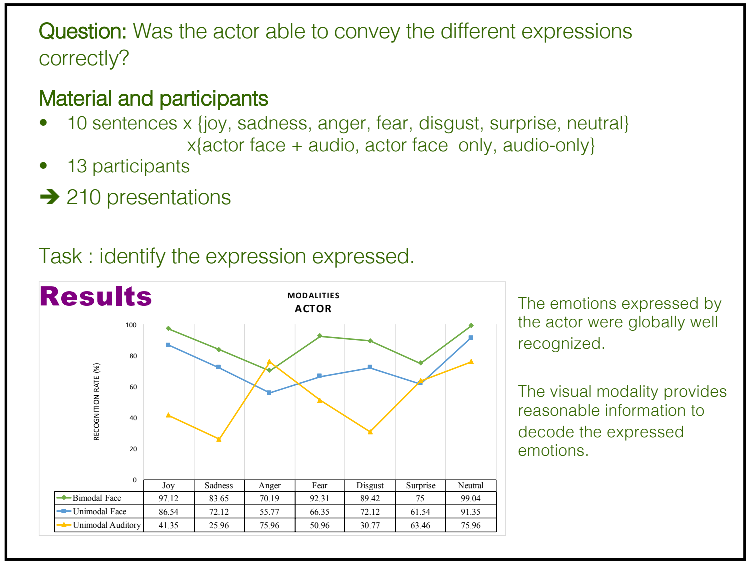
- A case study of a semi-professional actor
- A set of sentences in 6 different emotions (joy, sadness, anger, fear, disgust, surprise) + neutral.
- Quality of acting

Technique exercise in style

- The same set of sentences uttered in different emotions (joy, sadness, anger, fear, disgust, surprise) + neutral.
  ➔The linguistic content of the sentences does not help to identify the expressed emotion (dissociation between semantics and syntax).

Slim Ouni, Sara Dahmani, Vincent Colotte (2017). On the quality of an expressive audiovisual corpus: a case study of acted speech. AVSP 2017: 53-57

36



Question: Was the actor able to convey the different expressions correctly?

Material and participants

- 10 sentences x {joy, sadness, anger, fear, disgust, surprise, neutral} x{actor face + audio, actor face only, audio-only}
- 13 participants

➔ 210 presentations

Task : identify the expression expressed.

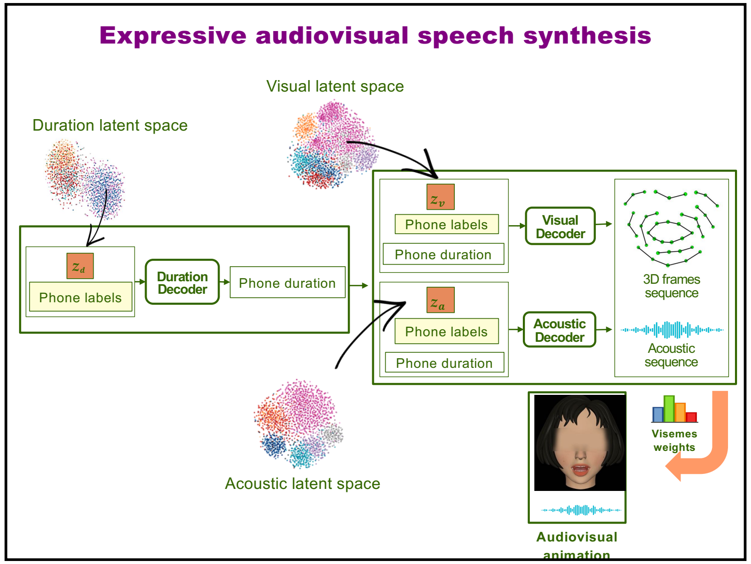
Results

	Joy	Sadness	Anger	Fear	Disgust	Surprise	Neutral
Bimodal Face	97.12	83.65	70.19	92.31	89.42	75	99.04
Unimodal Face	86.54	72.12	55.77	66.35	72.12	61.54	91.35
Unimodal Auditory	41.35	25.96	75.96	50.96	30.77	63.46	75.96

The emotions expressed by the actor were globally well recognized.

The visual modality provides reasonable information to decode the expressed emotions.

37



Expressive audiovisual speech synthesis

Visual latent space

Duration latent space

z_v

Phone labels

Phone duration

Visual Decoder

3D frames sequence

z_d

Phone labels

Duration Decoder

Phone duration

z_a

Phone labels

Phone duration

Acoustic Decoder

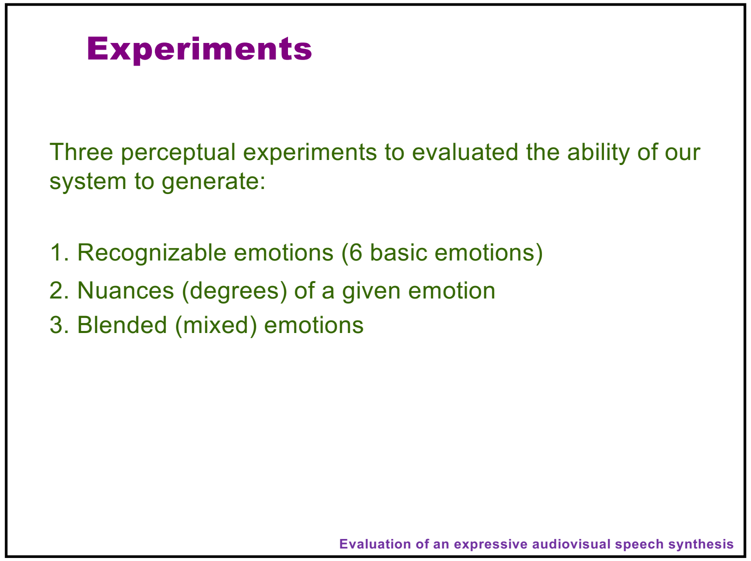
Acoustic sequence

Acoustic latent space

Visemes weights

Audiovisual animation

38



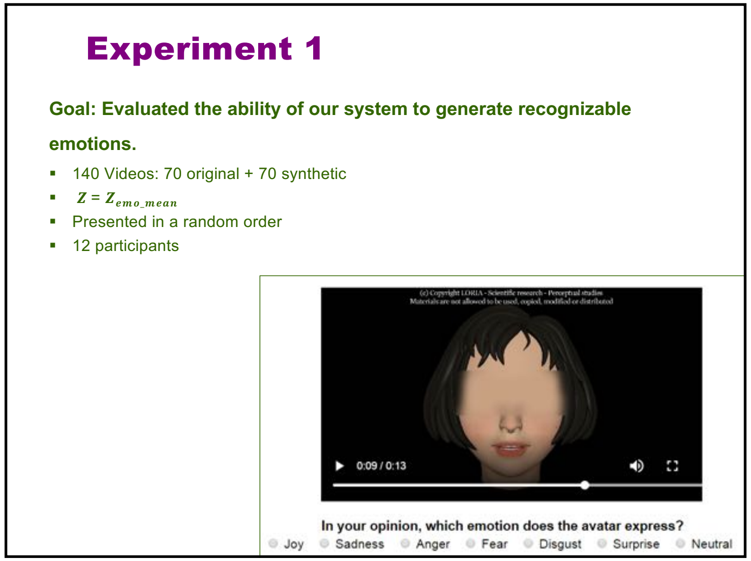
Experiments

Three perceptual experiments to evaluated the ability of our system to generate:

1. Recognizable emotions (6 basic emotions)
2. Nuances (degrees) of a given emotion
3. Blended (mixed) emotions

Evaluation of an expressive audiovisual speech synthesis

39



Experiment 1

Goal: Evaluated the ability of our system to generate recognizable emotions.

- 140 Videos: 70 original + 70 synthetic
- $Z = Z_{emo\_mean}$
- Presented in a random order
- 12 participants

(c) Copyright LORIA - Scientific research - Perceptual studies
Materials are not allowed to be used, copied, modified or distributed

In your opinion, which emotion does the avatar express?

○ Joy ○ Sadness ○ Anger ○ Fear ○ Disgust ○ Surprise ○ Neutral

40

Experiments – Experiment I Results

The recognition rates of the original and the synthetic emotions.

	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Original	97%	67%	42%	69%	77%	57%	72%
Synthetic	71%	83%	**11%**	71%	92%	**26%**	73%

- Emotions were globally well perceived
- The upper part of the face is important in conveying fear and sadness (Costantin et al. 2005)
- Some breathing and fine voice variations were smoothed and lost in synthetic samples

Experiments – Experiment II

Goal: Evaluate the ability of our system to generate nuances of emotions.

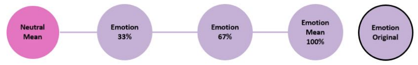
Linear combination of latent vectors:

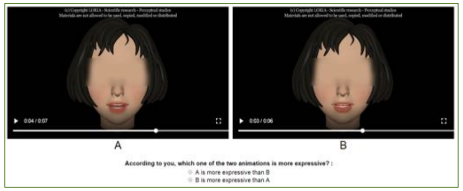
$$Z_{emo\_33} = Z_{emo\_mean} \text{ x } 0.33 + Z_{neutral} \text{ x } 0.67$$

$$Z_{emo\_67} = Z_{emo\_mean} \text{ x } 0.67 + Z_{neutral} \text{ x } 0.33$$

$$Z_{emo\_100} = Z_{emo\_mean}$$

- 210 videos *(6 emotions x 5 examples x 7 comparisons)*
- Presented in a random order
- 10 participants.




(c) Copyright LORIA - Scientific research - Perceptual studies
Materials are not allowed to be used, copied, modified or distributed

According to you, which one of the two animations is more expressive? :

○ A is more expressive than B
○ B is more expressive than A

Experiments – Experiment II Results

The recognition rates of the more expressive animation.

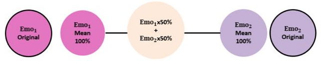
	0/33	0/67	0/100	33/67	33/100	67/100	100/orig
Anger	82%	94%	90%	94%	96%	88%	80%
Disgust	52%	80%	82%	92%	86%	70%	84%
Fear	58%	56%	80%	66%	72%	80%	86%
Joy	74%	92%	96%	90%	90%	90%	90%
Sadness	56%	70%	88%	74%	76%	86%	94%
Surprise	78%	92%	92%	90%	94%	86%	98%
Average	***66%***	***80%***	***88%***	***84%***	***85%***	***83%***	***88%***

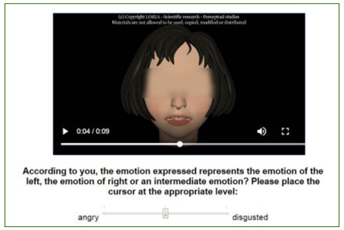
- Nuances were well perceived
- 0/33 the hardest to perceive

Experiments – Experiment III

Goal: Evaluate the ability of our system to generate blended emotion.

- 100 videos *(4 combinations x 5 videos x 5 examples)*
- $Z_{emo\_blend} = Z_{emo\_1} \text{ x } 0.5 + Z_{emo\_2} \text{ x } 0.5$
- Presented in a random order
- 12 participants.

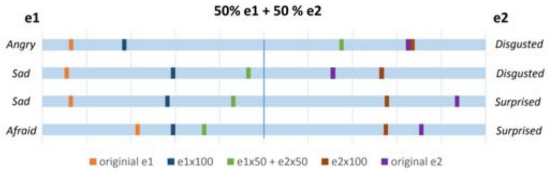



(c) Copyright LORIA - Scientific research - Perceptual studies
Materials are not allowed to be used, copied, modified or distributed

According to you, the emotion expressed represents the emotion of the left, the emotion of right or an intermediate emotion? Please place the cursor at the appropriate level:

angry — disgusted

Experiments – Experiment III Results



- The system succeeded in creating blended emotions that were correctly perceived as intermediate emotions
- Some emotions were more dominant than the others