
Dutch Multimodal Corpus for Speech Recognition

A.G. Chițu and L.J.M. Rothkrantz

E-mails: {A.G.Chitu,L.J.M.Rothkrantz}@tudelft.nl

Website: <http://mmi.tudelft.nl>

Outline:

- Background and goal of the paper.
- How is the data recorded?
- What type of data is recorded?
- Conclusions.

What is the background and the goal of the paper?

- ICIS – Crisis management.
 - Supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.
- Facilitating communications among actors.
- Visual enhanced speech recognition.
 - people do it all the time: McGurk effect - Nature, 1976
- Not sufficient data resources.

What is the background and the goal of the paper?

- Lipreading process is insufficiently understood.
- Start discussions about a set of guiding rules for building audio-visual databases used in multimodal speech-recognition research.
- Underline a first set of directions and spot some places that could/(need to) be improved.

How is the data recorded?

What are the issues related with the current audio-visual datasets?

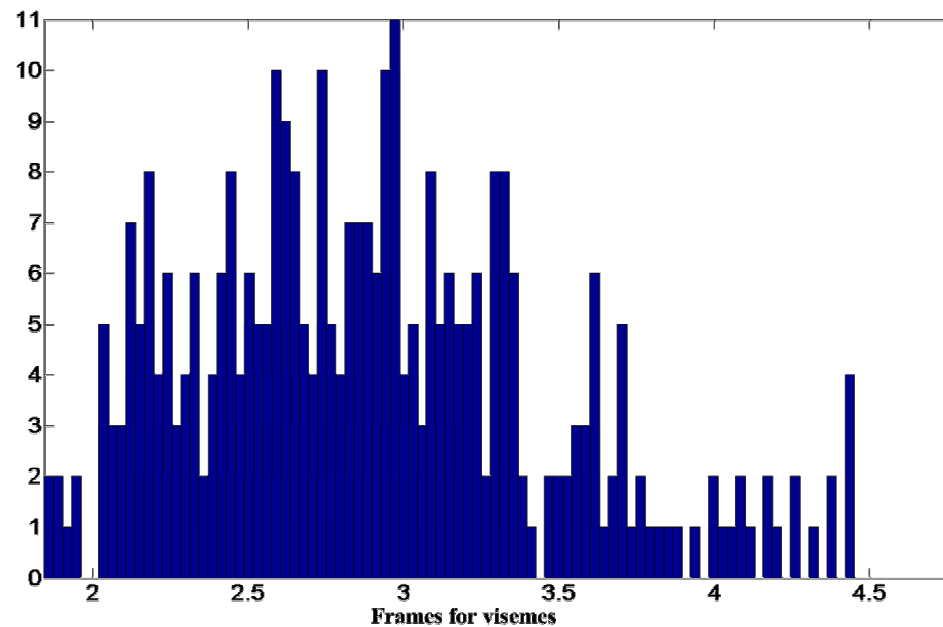
- Small number of respondents
 - » Unbalanced w.r.t. gender, age, race, dialect, etc.
- Poor language coverage
 - » Limited coverage of phonemes/visemes
- Poor quality of the recordings: esp. video
 - » Controlled/Uncontrolled environment?
- Not (freely) available

How is the data recorded?

- Video quality
 - Difficult to handle, stronger requirements from the recording device
 - Grayscale/Color
 - Resolution: 80x60 to 720x576
 - Frame rate: **24fps to 29.97fps**
 - Illumination, background
 - Region of Interest –ROI?

How is the data recorded?

- Video quality: poor coverage of the visemes



Fast speech rate vs. slow speech rate

How is the data recorded?

- Video device

Allied Vision Technology – Pike F032C, purely computer vision camera: max 206Hz b/w.

We recorded at 100Hz color images, setting chroma subsampling 4:2:2 and resolution $\frac{1}{2}$ PAL 384x288.

Controlled environment, chroma keying, ROI face level.

- Audio device

High fidelity MICS: NT2A Studio Condensators.

Stereo signal at 48kHz and 16bits sample size.

Controlled environment.

How is the data recorded?

- Video quality: dual view



Solutions:

1. Two cameras → Synchronization
2. One camera + mirror



How is the data recorded?

- Video quality: dual view

Solutions:

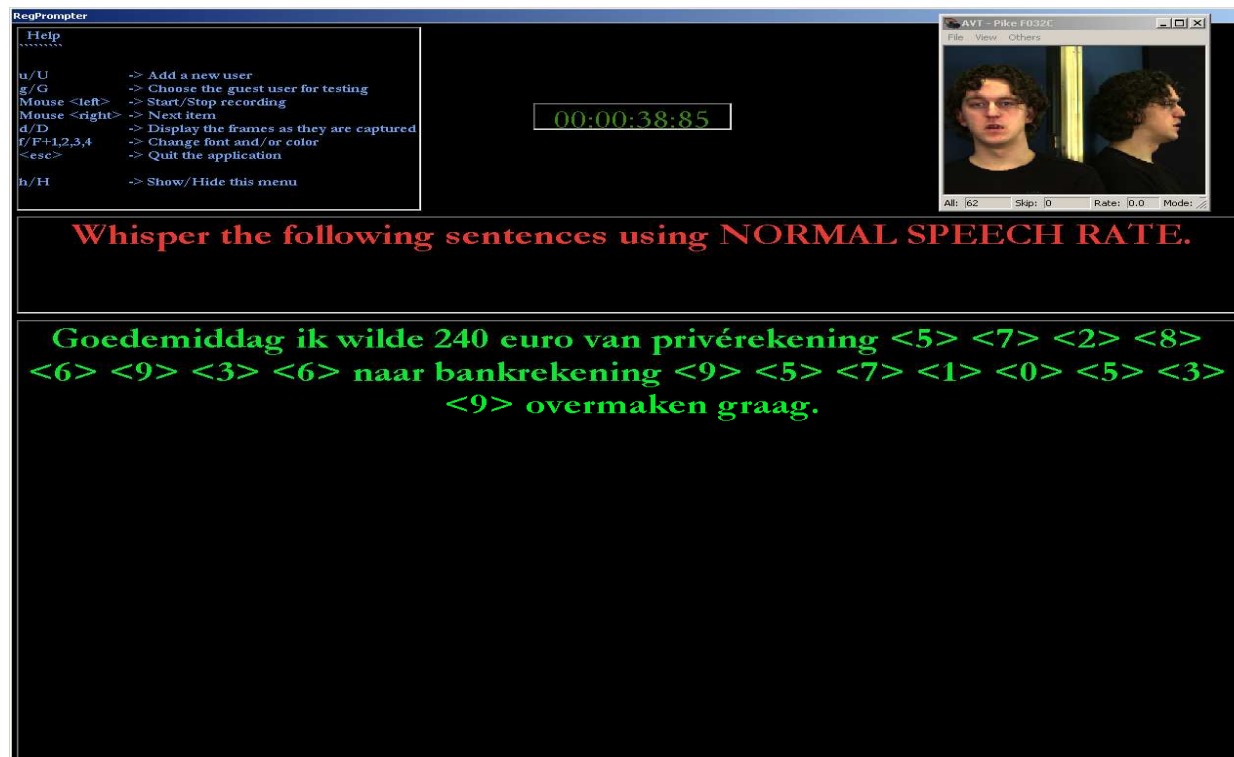
1. Two cameras →
Synchronization

2. One camera + mirror



How is the data recorded?

- Prompter → Speaker control of the devices



00:00:38:85

Whisper the following sentences using **NORMAL SPEECH RATE**.

Goedemiddag ik wilde 240 euro van privérekening <5> <7> <2> <8>
<6> <9> <3> <6> naar bankrekening <9> <5> <7> <1> <0> <5> <3>
<9> overmaken graag.

What type of data is recorded?

- Language content for lipreading
 - Start from DUTAVSC DB
 - » based on POLYPHONE for Dutch + Newspaper clips + banking application
 - » enriched by adding: short interview questions, “klink”-ers, many new connected digits series, many new random words.
 - » recorded in 4 styles: normal rate, fast rate, whispering, free answer interview.

Enrich the language coverage!!!

What type of data is recorded?

- Collected data: statistics
 - 1966 unique words
 - 427 phonetically rich sentences
 - 91 context aware sentences (banking application)
 - 72 different “klink”-ers
 - 41 simple open answer questions
 - 64 items per speaker
 - 16 categories w.r.t. content and speech style
 - ~45min per speaker

What type of data is recorded?

- Collected data: statistics
 - native Dutch speakers
 - hope for >100 speakers (we have already 60)
 - more than 5000 utterances
- Collected data: demographic
 - age
 - gender
 - education
 - address: region

What type of data is recorded?

- Emotional speech

#	Emotion	#	Emotion	#	Emotion
1	Admiration	8	Disgust	15	Indignation
2	Amusement	9	Dislike	16	Inspiration
3	Anger	10	Dissatisfaction	17	Interest
4	Boredom	11	Fascination	18	Pleasant surprise
5	Contempt	12	Fear	19	Sadness
6	Desire	13	Fury	20	Satisfaction
7	Disappointment	14	Happiness	21	Unpleasant surprise

Acted emotion based on reading emotional stories.

Second data corpus. The respondents were both naïve speakers and professional actors.

Capture passport view to observe all facial expressions and gestures

What type of data is recorded?

- Emotional speech

Dutch original
Emotie: “Bewondering” Vertelling: “Je loopt samen met een vriend/vriendin door een dure winkelstraat in Amsterdam en ziet in de etalage een jas hangen die je altijd al had willen hebben. Je droomt over wat je zou doen als je het geld had om deze jas te kopen. Je gaat voor de etalage staan en denkt...” Reactie: R1: Oooohhh.. R2: Dat ziet er goed uit. R3: Die zou ik graag willen hebben. R4: Was die maar van mij. R5: Zodra ik mijn geld heb, is die jas van mij.
English translation
Emotion: “Admiration” Story: “You walk together with your friend in front of a fancy store in <u>Amsterdam</u> and you see in the store’s window a coat that you always wanted. You dream of what you would do if you have had the money to buy the coat. You stand in front of the window and think...” Reaction: R1: Oooohhh.. R2: That looks so nice. R3: I would really want it. R4: That is for me. R5: As soon as I’ll have money that coat is mine.

What type of data is recorded?

- Emotional speech

Motivation → Influence of the affective state on the shown visemes and vice versa.

Problems → In real life there is not such thing as clean emotion: amalgam of emotions.

→ Assessing the quality of the acted emotions.

We plan to use questionnaires to elicit speakers' opinion, and to weight naïve speakers against professional actors.

What type of data is recorded?

- Different environmental conditions
 - While the recordings were performed in laboratory conditions with reduced noise level, we want to expose the speakers to different noise conditions.
 - It is not clear to us how is the speaking style changing with the environment.
 - It seems common sense that under heavy noise the speaker will change his speaking style to increase lip readability?! But how, in what sense and degree?

What type of data is recorded?

- Different environmental conditions
 - Hence we think of asking speakers to record while:
 - Very low feedback is possible – but silence
 - Very low feedback is possible – but heavy noise conditions
 - Half feedback and half noisy input.

Take home facts.

- Lipreading is a very active subject at this moment.
- At TUDelft there will be a new advanced data corpus available for multimodal speech recognition.
- Frontal and side view lipreading.
- High speed video recording.
- Multiple recording conditions are simulated.
- Emotional speech is included and facial expressions and gestures are recorded together with speech.

Comments

Thank you very much for your attention!

A.G. Chițu and L.J.M. Rothkrantz

E-mails: {A.G.Chitu,L.J.M.Rothkrantz}@tudelft.nl

Website: <http://mmi.tudelft.nl>