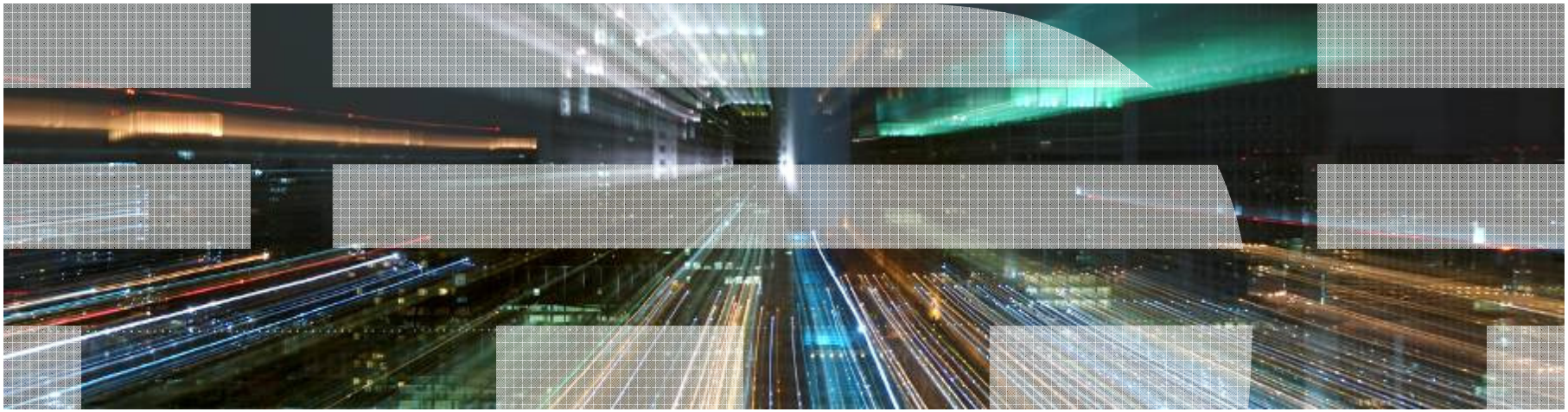
Alex Sorin, IBM Research - Haifa

Voice Analytics for Dementia Assessment

DemAAL, September 2013, Chania, Greece



Acknowledgements

- The experimental work presented below is supported by Dem@Care FP7 project partially funded by the EC under contract number 288199

- The author would like to thank

  Aharon Satt, Orith Toledo-Ronen and *Oren Barkan* from IBM,

  Ioannis Kompatsiaris and *Athina Kokonozi* from CERTH,

  Magda Tsolaki from Greek Association for Alzheimer Disease and Related Disorders (GAADRD)

  for their contribution to the experimental study presented in this lecture

Outline

- Voice-based assessment of dementia - Vision
  - Motivation
  - Use cases
  - Protocol-based assessment, vocal tasks
  - Trainable system for voice-based dementia assessment
  - Prior art
- Voice-based assessment of dementia – Feasibility study
  - Voice data corpora collected in Dem@Care
  - Discriminative power of some vocal features
  - Evaluation setup and results
  - Future work

Voice-based assessment of dementia
Vision

Voice-based dementia assessment – looks feasible and attractive

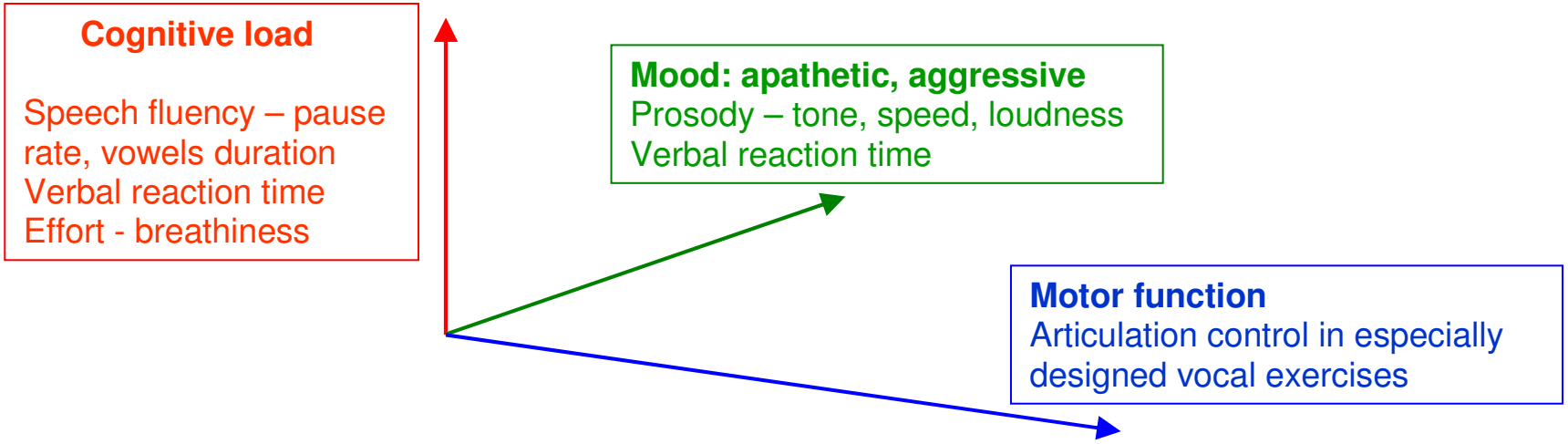
- Various types of dementia significantly affect human speech and language*) **). Therefore speech can be considered as a source of information for dementia assessment
- Dementia affects speech at two levels
  - Linguistic level - spoken content, what is said
  - Paralinguistic level – beyond the spoken content, how the person speaks
- Modern automatic speech processing and machine learning techniques could enable extraction of dementia relevant information from the speech audio signal and interpretation of this information in terms of presence and strength of dementia
- Audio capturing is relatively easy and cheap. The assessment can be done remotely if needed, e.g. over the phone

*) Appell, J., et al, "A study of language functioning in Alzheimer patients", Brain Language 17: 73-91, 1982

**) Reilly, J., et al, "Cognition, language and clinical features of non-Alzheimer's dementias: an overview", Journal of Communication Disorders 43(5): 438-452, 2010

Focus on the paralinguistic level

- Spoken content analysis would require speech-to-text transcription using an Automatic Speech Recognition (ASR) system
- ASR accuracy achievable on spontaneous conversational speech is not high enough for our purpose
  - We can expected ASR word error rate of 40% and even higher
- We focus on the paralinguistic information – how the person speaks rather than what he/she says



Voice assessment in ambient monitoring

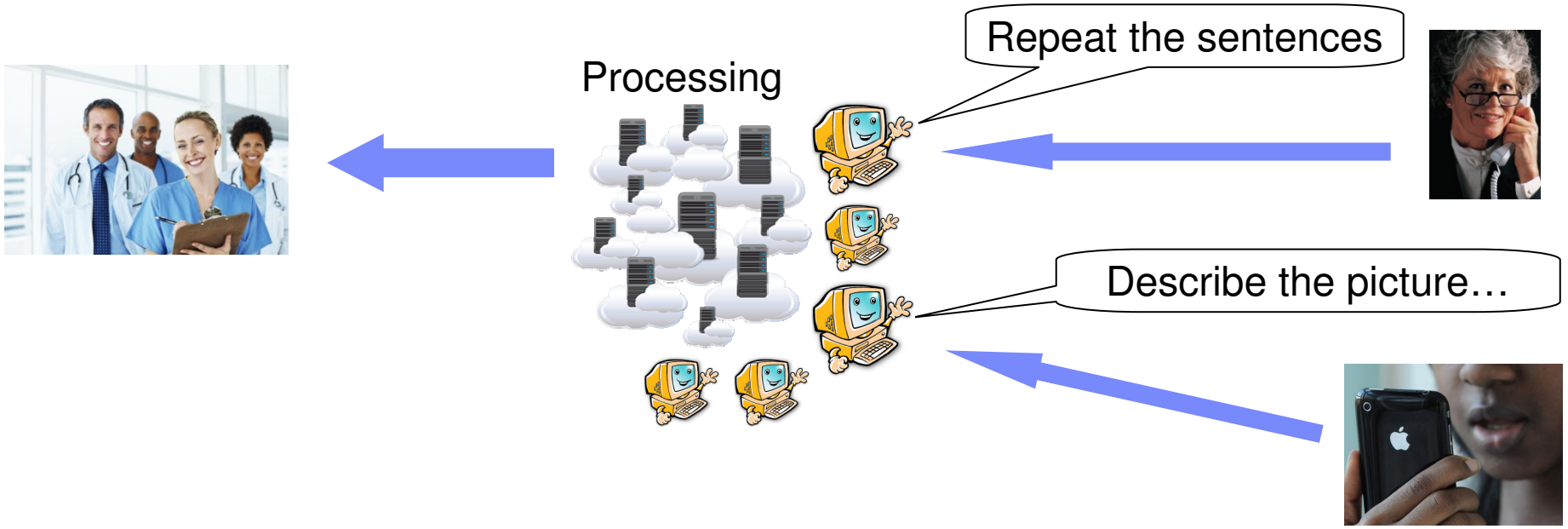
- Continuous status assessment of person with dementia in home settings
- Analysis of face-to-face and over-the-phone conversations with others
  - Part of Dem@Care Home use case
- Audio capturing with fixed wall-mounted microphones is complicated when high audio quality and good coverage of the living spaces are required
- Audio capturing with wireless lavalier microphone worn by the monitored person seems preferable
  - Compact, easy to operate, long battery life, good coverage of space, high quality audio
- Uncontrolled natural environment poses great challenges
  - Context ambiguity: is it a face-to-face conversation, phone call or a TV talk program?
    - Wearable microphone reduces the uncertainty but also attenuates speech of the counterpart
  - The person habits and life conditions are unique. Assessment in absolute terms is difficult. We only can detect abnormal patterns and trends for the monitored person
- Voice should be combined with video (and possibly other sensors) data to resolve the situational context ambiguity and improve the assessment reliability

Protocol-based voice assessment – our focus

- The person being examined is asked to perform a set of pre-defined spoken exercises (vocal tasks)
- The tasks are designed with the goal to pose certain (cognitive and/or motor) challenges to a person with dementia
- The audio can be recorded by different microphone types (table-top, headset, lavalier) or by a smartphone depending on the use case
- Controlled environment facilitates the audio analysis and enables assessment in absolute terms
- Use cases:
  - A part of multi-sensor clinical assessment, provides objective "second opinion" to the clinician – Dem@Care Lab scenario
  - Computer- or caregiver-guided multi-sensor self-assessment at home – a part of Dem@Care Home scenario
  - *Remote automatic voice-based assessment using Interactive Voice Response service or a smartphone application*

Vocal tasks

- Some tasks are suggested by prior art, others are borrowed from standard psycho-neurological tests or speech/language pathologist practices
- Below are some examples
  - *Verbal description of pictures*
  - *Repeating sentences uttered by a human/automatic agent*
  - *Diadochokinetic test – fast repeated uttering of a multi-syllable token, e.g. pa-ta-ka—pa-ta-ka—pa-ta-ka--…*
  - *Counting backward*
  - *Reading out a text passage*
  - *Verbal fluency test – naming as many as possible items belonging to a predefined category, e.g. animals, within a limited time interval, e.g. 1 min*
- Other tasks can be devised – it is an exploratory work



- Scalable service for early diagnosis and monitoring of dementia through voice/speech
- Immediate feedback to the clinician/caregiver
- Subsequent selective referral to the clinic
- Mass-screening of the population at risk without an extra-burden on the healthcare system

Trainable system for voice-based assessment of dementia

- Feature extraction step – the audio signals recorded during the vocal tasks execution are analyzed and *useful parameters* indicative of dementia are derived from them and aggregated to a *vocal feature vector*

- Prediction step - the feature vector is mapped to an integral *dementia rating* ranging from 0 (healthy) to 1 (severe) by means of predictive models trained on available data manually labeled using the best clinical practices
  - Examples of the predictive models include Support Vector Machine (SVM), Regression Trees, Gaussian Mixture Models (GMM)
  - Assessment on a continuous scale is especially important for monitoring of people with diagnosed dementia

Trainable system for voice-based assessment of dementia – cont'd

- Availability of the manual labels on a continuous scale is an unrealistic hope. The manual labels available for the training will be rather in terms of diagnosis, e.g. healthy, MCI. The continuous rating at the system output can be derived from the internal confidence measures calculated by the predictive model used, e.g. a distance from the separating hyperplane in SVM, or a likelihood value in GMM

- Availability of the vocal tasks audio recordings and associated clinical data is crucial for the feasibility study and real solution development including training and evaluation

Vocal features relevant for dementia assessment

- Modern techniques of speech signal analysis enable extraction of multiple parameters indicative of contribution of different parts of the human speech production mechanism
  - Spectral parameters - vocal tract (oral and nasal cavities)
  - Voicing, pitch – vocal folds
  - Glottal pulse parameters – glottis
- Many of these parameters are being explored for voice emotion and speaker trait recognition
- There are studies using some of these parameters for detection of diseases that damage motor functions and thus influence speech production
  - Parkinson's disease
  - Laryngeal pathologies

Vocal features relevant for dementia assessment – cont'd

- Although voice pathologies (e.g. dysarthria) may be associated with severe dementia, they do not appear at early stages of dementia
- Hence we are interested in speaking behavior characteristics indicating speech disfluency resulting from cognitive deficit
  - Number and duration of pauses – more pauses, longer pauses indicate cognitive load
  - Number and duration of vowels – people subconsciously prolong vowels and use filled pauses (*UH, UM, ER, ERM*) to gain more time for thinking
- Pauses can be identified using Voice Activity Detection techniques
- Vowels belong to the broad class of *voiced sounds* characterized by quasi-periodic movements of vocal folds and therefore by quasi-periodic waveform. Voiced consonants (e.g. 'b', 'd') are also voiced sounds but they are relatively short.
  Voiced segments of speech can be identified using Pitch Estimation techniques

Short summary of the Prior Art

- Telephone cognitive interviews for elderly – manually administered and scored. Mainly at the linguistic level – # of correct answers
  - Telephone Interview for Cognitive Status (TICS) – studied by NIH
  - Brief Test of Adult Cognition by Telephone (BTACT) – studied by Brandeis University; includes paralinguistic elements measurement using a sound editor

- Several studies reported correlation between dementia and certain vocal features. However, the expected performance of an end-to-end voice-based dementia assessment system was not clearly demonstrated

- TRIL Consortium (Technology Research for Independent Living). Reported a study on vocal tasks and certain speaking behavior features for cognitive decline assessment. Articulated the idea and performed an experimental work on the IVR-based telephone cognitive interview

Voice-based assessment of dementia
Feasibility Study

Data collection efforts in Dem@Care project

- Who participate in the data collection - elderly people who passed standard clinical neuropsychological assessment
- What data is collected – audio recordings of the tasks execution and anonymized clinical & demographic data of the participants used as a reference
- General multi-sensor data collection at the University Hospital in Nice, France – on-going
  - A rich protocol that includes certain vocal tasks
- Speech data collection by Prof. Robert at the University Hospital in Nice – on-going
- Speech data collection at GAADRD Center at Thessaloniki – completed
  - Below we present an experimental study performed using this data corpus

Dem@Care Greek Data corpus

89 participants performing a set of vocal tasks were recorded at GAADRD Center at Thessaloniki using a simple headset microphone

Group	Persons	Males/Females	Age mean	Age range
Control	19	4/15	67	56-84
MCI	43	12/31	73	52-88
Early AD	27	3/24	72	54-84

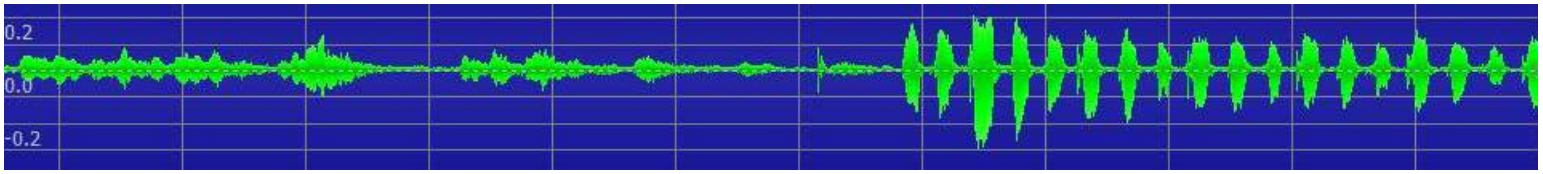
Task #	Task description
1	Verbally describe a picture while looking at it
2	Look at a picture, then describe it from memory
3	Repeat a short sentence. Done 15 times, with different sentences
4	Pronounce repeatedly and fast the sequence of 3 syllables: *pa-ta-ka*

The goal of the experimental study on the Greek Data

- Setting up a trainable experimental system for dementia diagnosis based on vocal features extracted from the recordings of a simple and short protocol
  - In this study we tried dementia detection instead of dementia rating on a continuous scale
- Finding of discriminative vocal features
- Evaluation of the diagnosis accuracy

Diadochokinetic (DDK) test

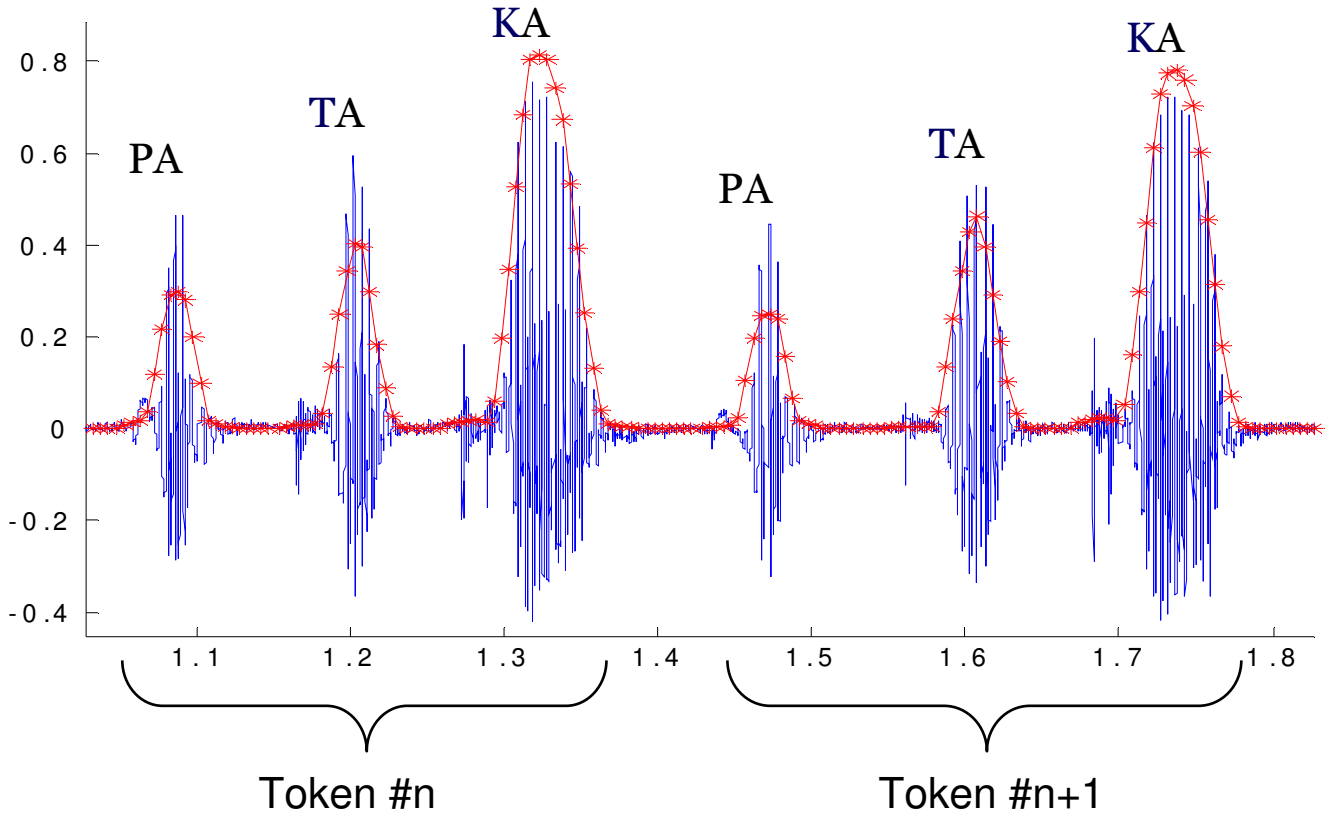
- DDK tests are used by speech-language pathologists for assessment of motor speech impairments, e.g. dysarthria
- Quick and accurate production of rapid, alternating sound tokens involving different parts of mouse, e.g. “pa-ta-ka—pa-ta-ka—…”
- Usually measurements are done manually and include the DDK rate statistics



- Our hypothesis and research direction:
  - The DDK test may challenge both **motor** and **cognitive** control over speech production
  - Motor and cognitive faults deteriorate the temporal regularity of the audio signal normally expected in this type of utterances
  - We can develop a regularity measure of the DDK performance
  - Is the DDK regularity useful in distinguishing between Control/MCI/AD groups?

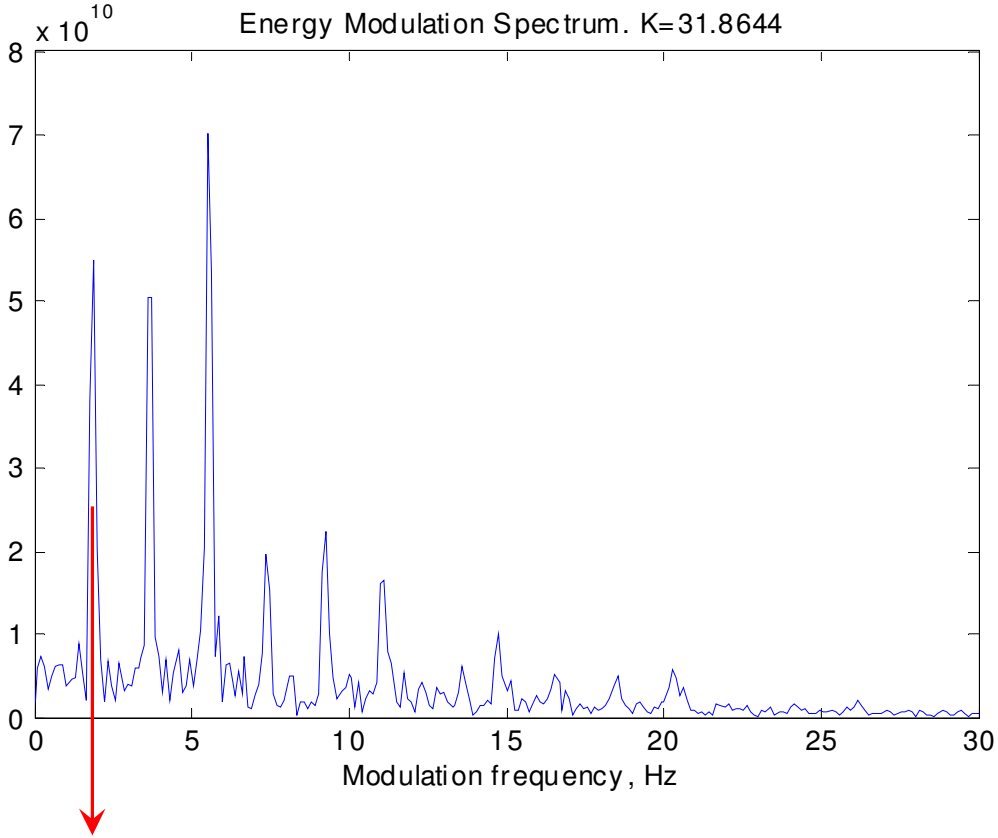
Developing DDK regularity measure: 1. Waveform Envelope

Build a waveform envelope in the form of instantaneous energy contour to filter out local details and short-term periodicity represented by the pitch cycles



Developing DDK regularity measure: 2. Energy Modulation Spectrum

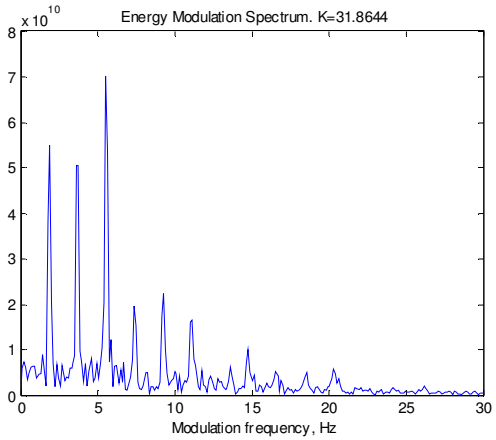
Long-term Fourier Transform of the Envelope

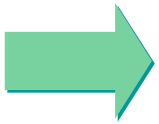


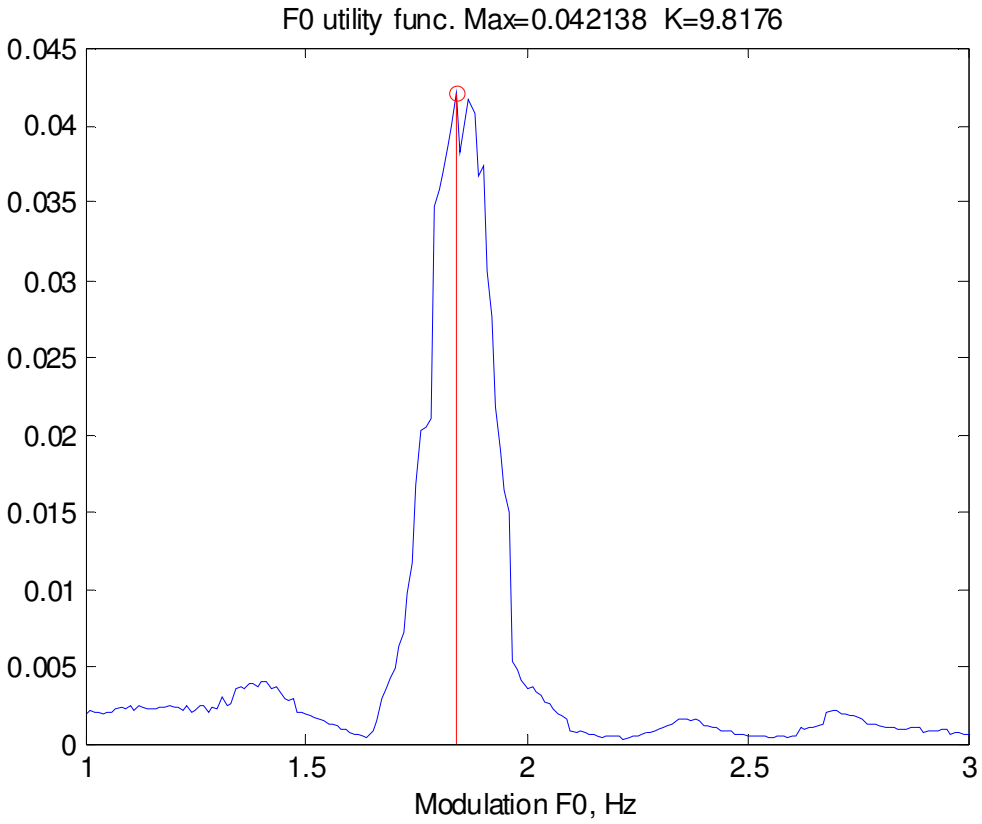
F0=1.8 Hz, i.e. 1.8 tokens per second on average

Harmonic structure indicates long-term periodicity of the waveform envelope

Developing DDK regularity measure: 3. Utility Function



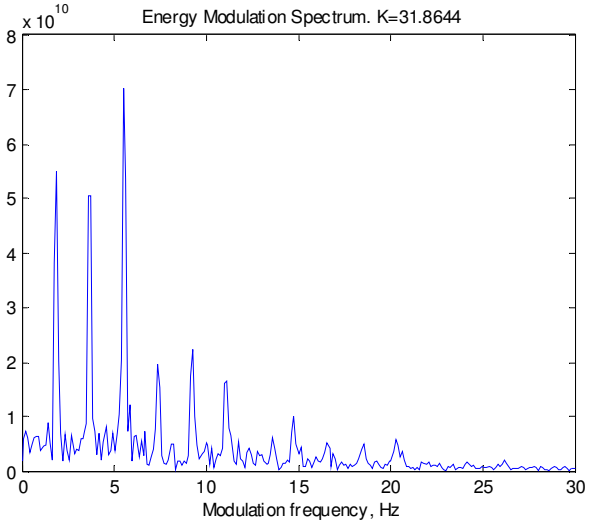


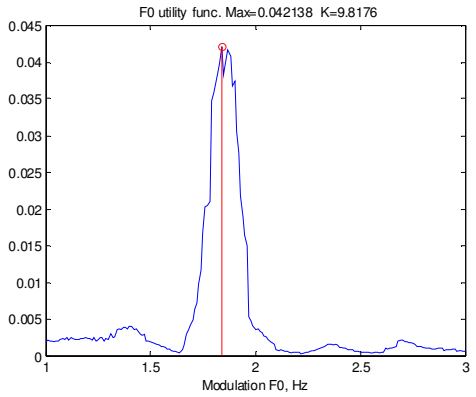


- U(F) is proportional to the integral magnitude of the Modulation Spectrum concentrated near the multiples of F Hz
- Sharpness of the Utility Function represented by its maximal value indicates the harmonicity of the modulation spectrum

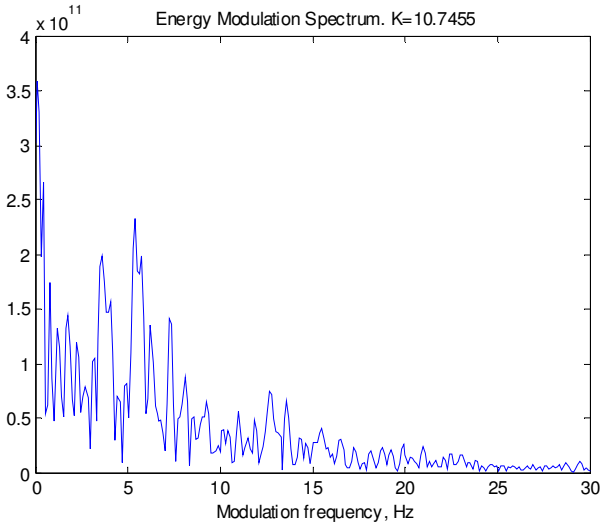
DDK regularity analysis examples

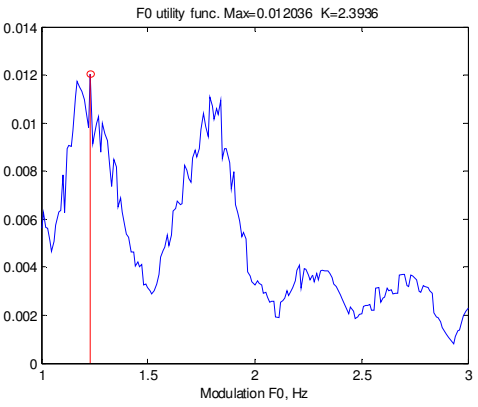
Good performance



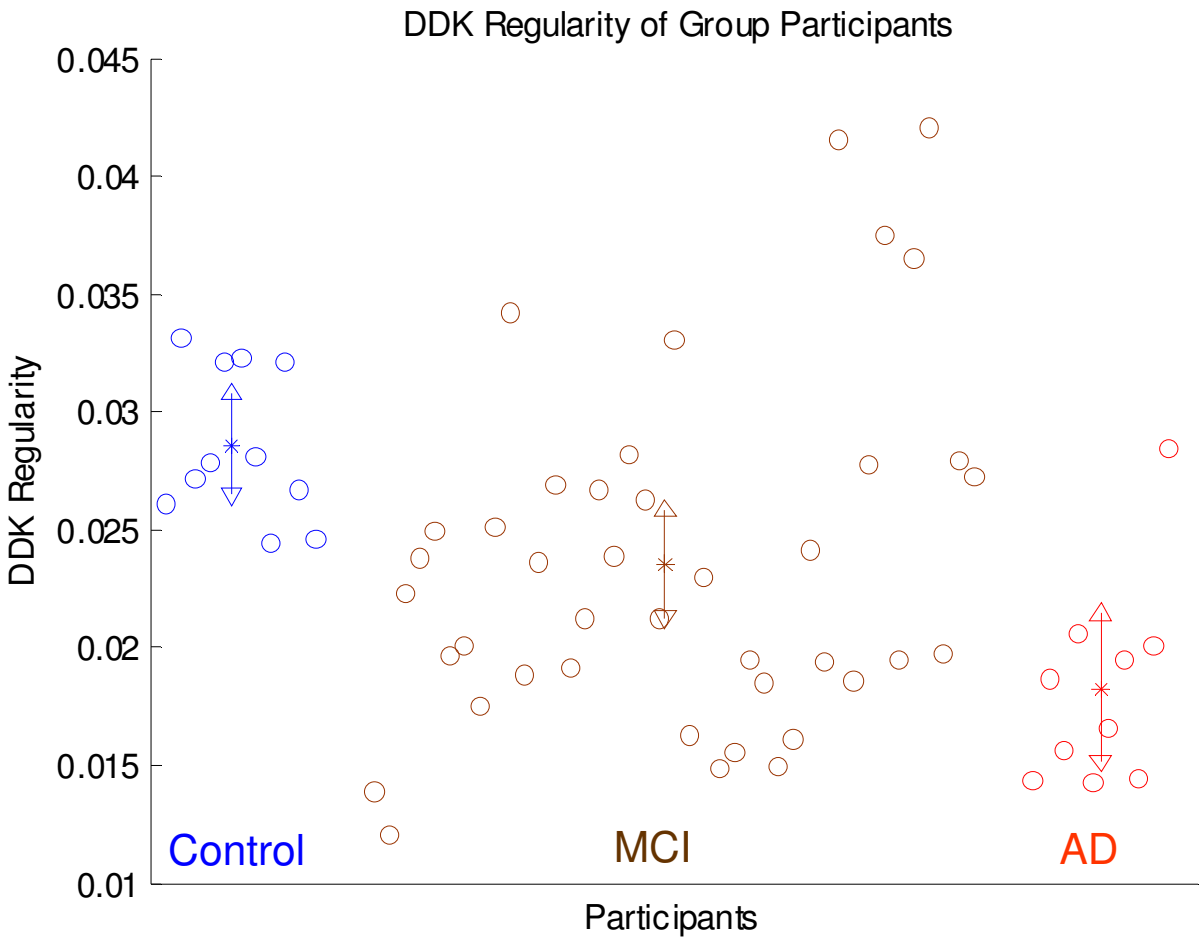


Poor performance





DDK regularity mapping for a subset of the Greek data



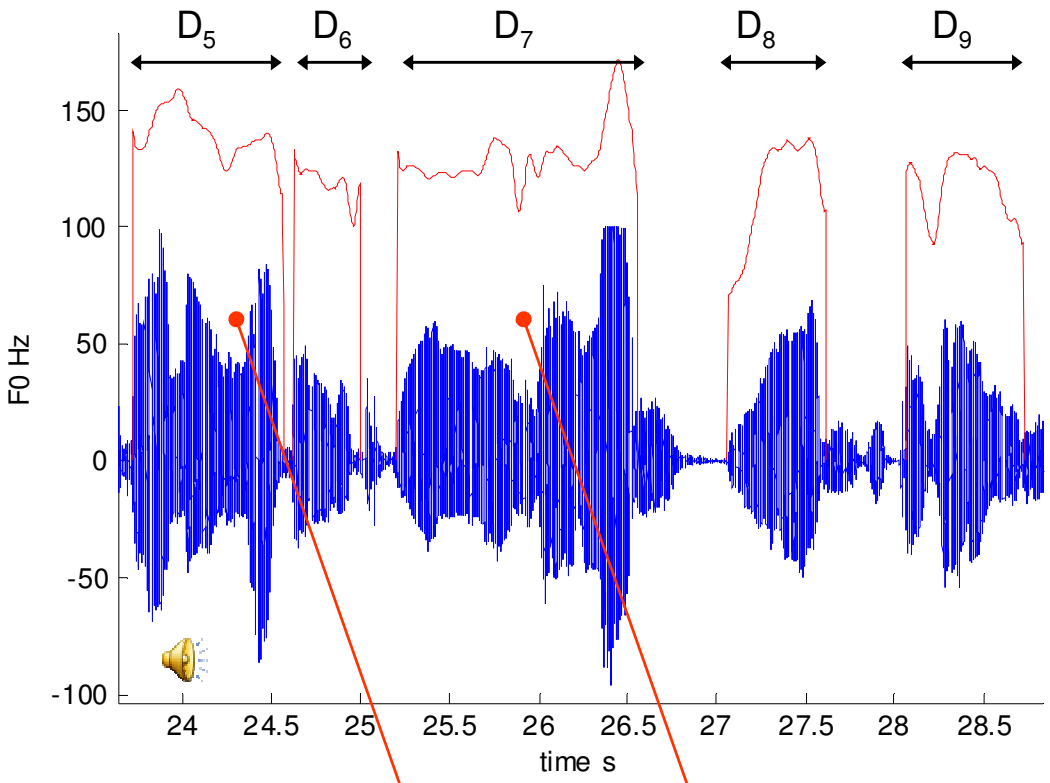
✓ *Statistically significant inter-group differences with T-test pvalues:*

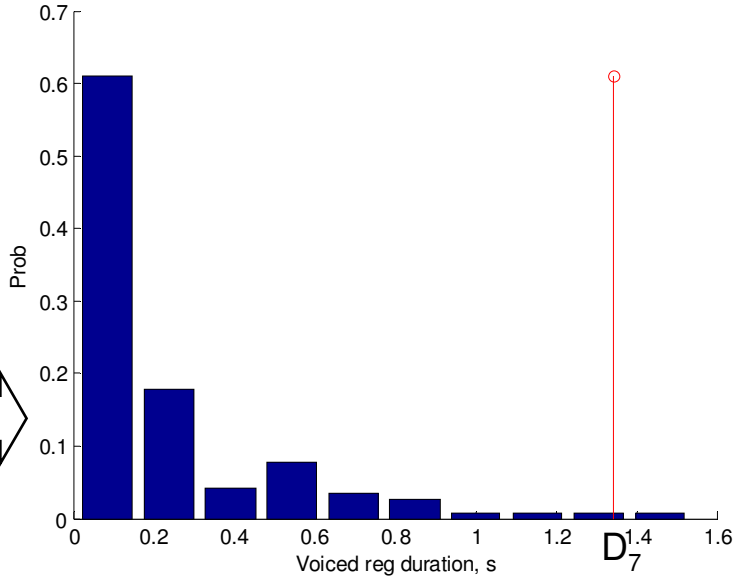
Control vs. MCI p<1.45%

Control vs. AD p<0.05%

MCI vs. AD p<1.65%

Voiced regions duration analysis in the picture description task

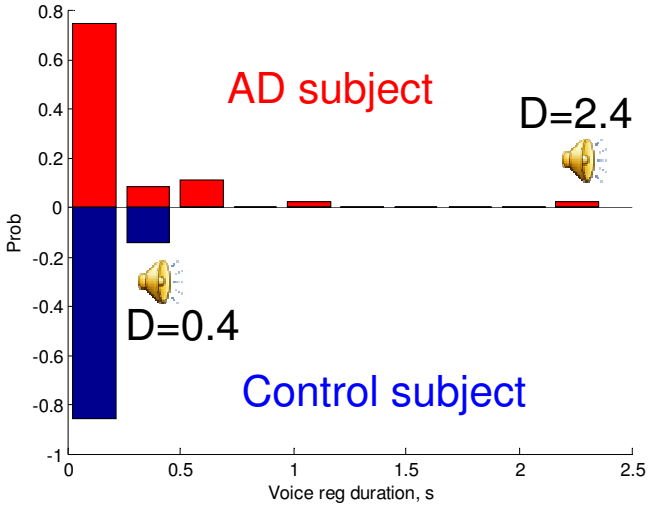




This sample contains unusually long voiced regions,
e.g. D7 ≈ 1.4 s

Maximal voiced region duration

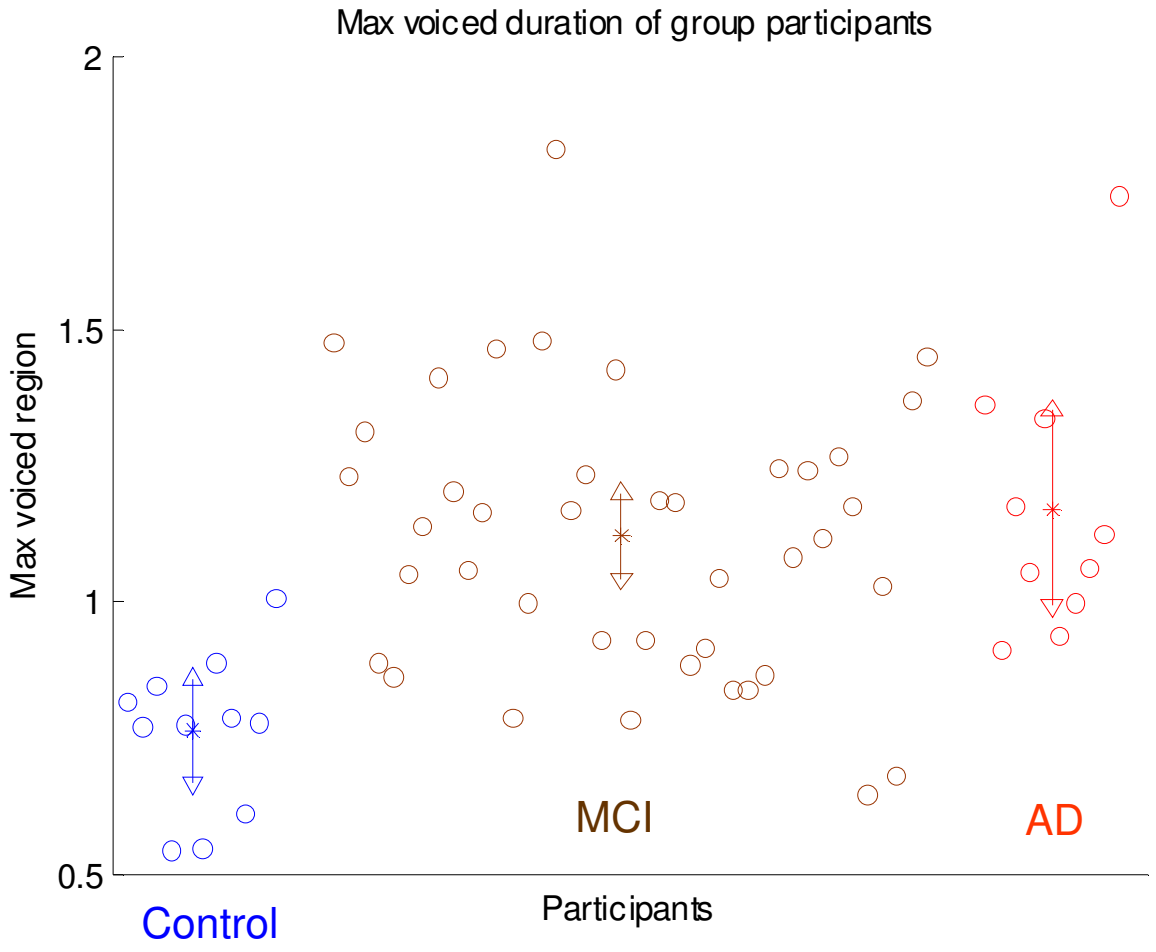
- Observations suggest to use the maximal voiced region duration or a high percentile of the voiced region duration distribution



- Possible reasons for long voiced regions in PwD's speech
  - Cognitive – subconscious prolongation of vowels to gain more time for thinking
  - Motor – "swallowing" unvoiced consonants unites separate voiced regions together

- This feature seems applicable to any conversation

Max voiced region duration mapping for a subset of the Greek data



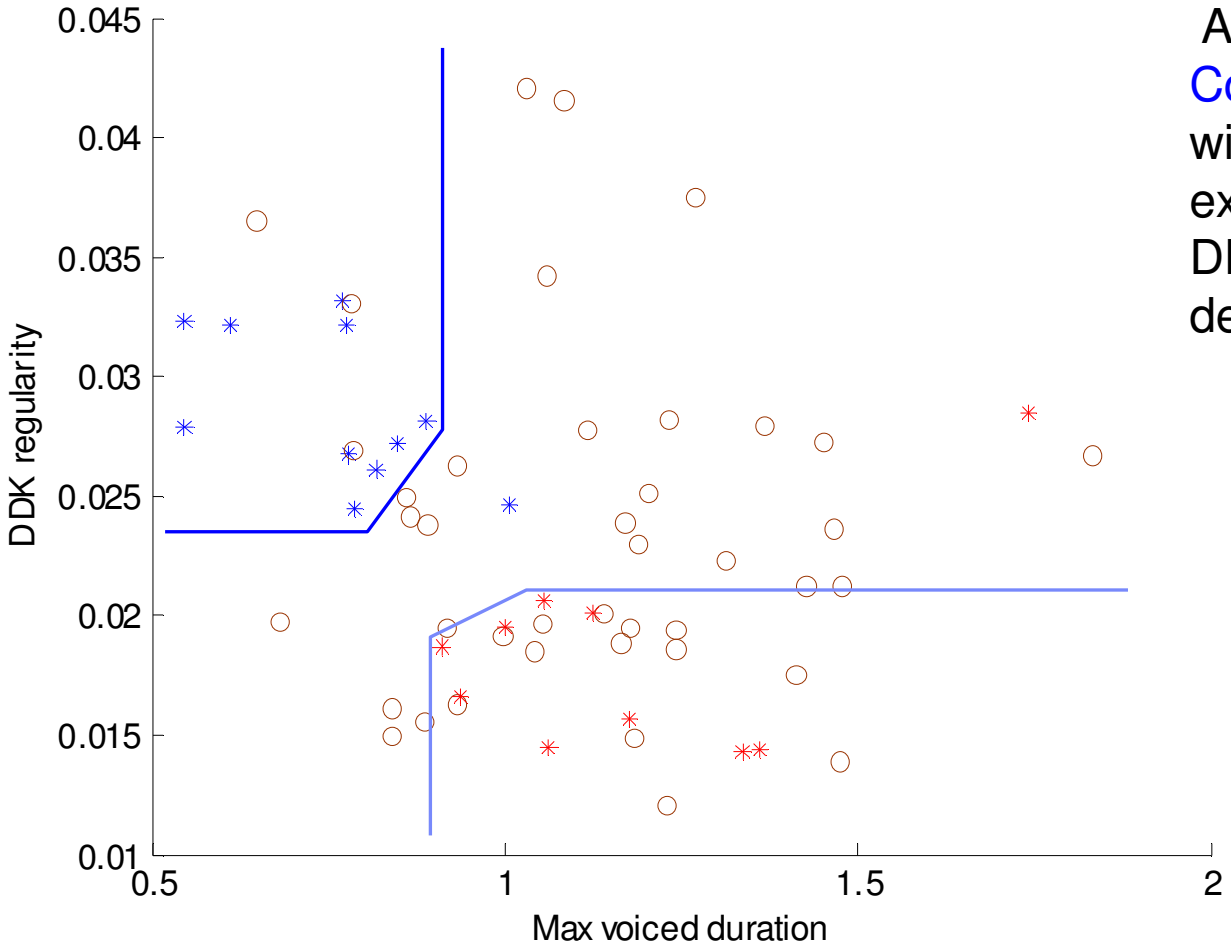
✓ *Statistically significant differences with T-test pvalues:*

Control vs. MCI p<0.05%

Control vs. AD p<0.05%

- The difference between MCI and AD is not statistically significant

2D mapping at “DDK regularity – max voiced duration” plane



A separation between Control vs. MCI vs. AD with only two vocal features extracted respectively from the DDK test and picture description task

* Control subject

* AD subject

o MCI Subject

Evaluation on the Greek corpus – feature extraction

- Tasks 1 and 2: verbal description of pictures (immediate and deferred)
  - Based on duration statistics of speech segments, pauses, voiced segments, unvoiced segments
  - The most discriminative features: stdev of voiced/unvoiced segment duration, temporal regularity of voiced segment durations, percentage of total pause duration, stdev of speech segment duration, total description duration
- Task 3: repeating sentences
  - Statistics of the normalized sentence duration (participant/interviewer), stdev and average of verbal reaction time, average distance to the reference sentence in the log-spectral space after optimal DTW alignment
- Task 4: diadochokinetic test
  - Average token duration, stdev of the token duration, total number of tokens uttered, average number of errors/pauses per token, DDK regularity (described above)
- All the features were aggregated in a single 20-dimensional feature vector representing the participant

Evaluation results on the Greek data corpus

- Support Vector Machine (SVM) classifier with Gaussian radial basis function
- 4-fold cross-validation repeated 10 times with random sampling
  - 40 instances of the classifier were trained and evaluated
- Accuracy is measured at the Equal Error Rate working point
  - P(class1|class2) = P(class2|class1)
- Joint paper by IBM, CERTH and GAADRD
  Satt, A., et al, "Evaluation of speech-based protocol for detection of early-stage dementia", In Proc. of INTERSPEECH-2013, Lyon, France, Sept, 2013

	Class 1	Class 2	Accuracy & 95%-confidence interval
Test 1	Control	MCI & AD	82% ± 6%
Test 2	Control	MCI	83% ± 6%
Test 3	Control	AD	84.5% ± 6%

Next steps

- Fully automate the feature extraction process
- Reach above 90% percent accuracy in healthy-vs.-MCI classification
  - Enrich the repertoire of the vocal features and use automatic feature selection techniques
    - In particular extraction of ASR-based features for constrained tasks like repeating sentences and reading aloud seems feasible and useful
  - Try other classification techniques including Artificial Neural Networks, Deep Belief Networks, Random Forest Classifier
  - Evaluate contribution of individual vocal tasks and devise new tasks
  - Enrich the protocol including selected non-vocal elements inspired by or adopted from the standard neuropsychological tests
    - These non-vocal tasks should suit GUI capabilities offered by smartphone devices
- Develop assessment rating on a continuous scale
  - Regression instead of classification