UNIVERSITY OF STIRLING

Cognitive Computation for Speech and Language Processing

Dr Andrew Abel
University of Stirling, Scotland

University of Stirling – Scotland

COSIPRA Laboratory

UNIVERSITY OF STIRLING

- Cognitive Signal Image and Control Processing Research (COSIPRA)

- Cognitive Computation
  - neurobiology, cognitive psychology and artificial intelligence
  - Moving beyond traditional Machine Learning
  - Considering the world
  - Inspired by the brain, machines that can think and take decisions in response to stimuli
- Develop a deeper & more comprehensive unified understanding of brain’s cognitive capabilities:
  - perception, action, and attention;
  - Learning and memory;
  - decision making and reasoning;
  - language processing and communication;
  - problem solving and consciousness

Why Cognitive Computation?

- Promote a more comprehensive and unified understanding of diverse topics
  - perception, action, and attention;
  - learning and memory;
  - decision making and reasoning;
  - Language processing and communication;
  - problem solving and consciousness aspects of cognition.
- Increasing calls for the creation of cognitive machines, with 'cognitive' powers similar to those of ourselves:
  - are able to 'think' for themselves;
  - are flexible, adaptive and able to learn from both their own previous experience and that of others around them

COSIPRA Lab Techniques & Applications: Example Applications/Case Studies

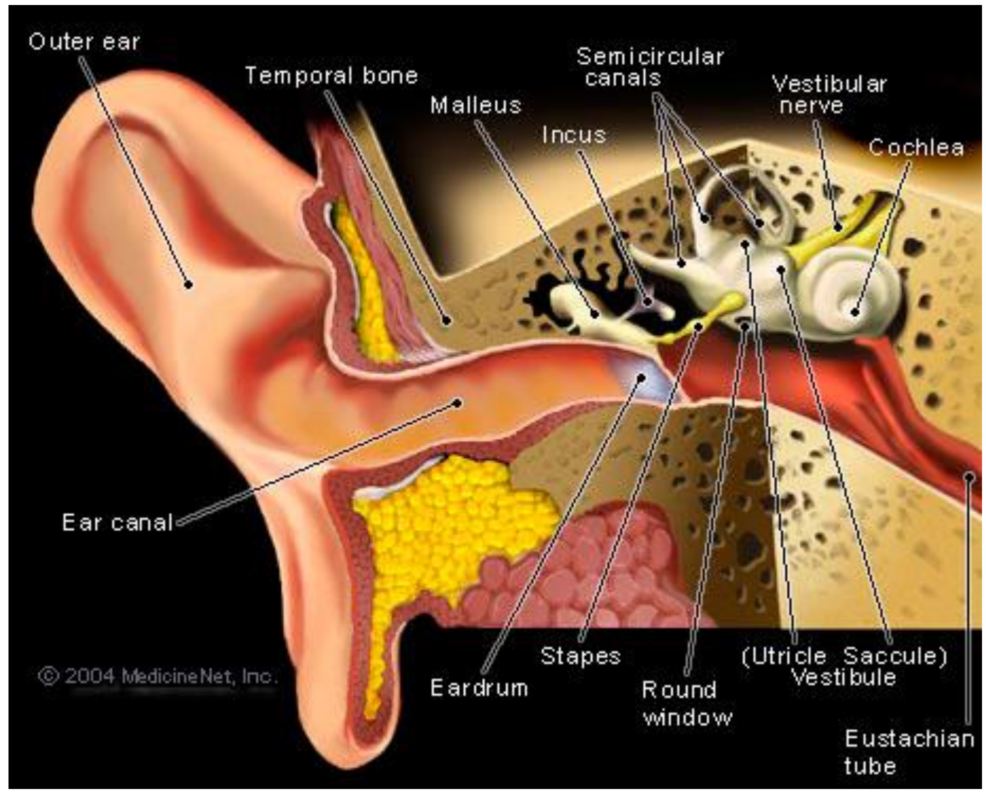
- Multimodal speech processing
  - cognitively-inspired multimodal speech communication capabilities
- Multimodal Sentiment analysis
  - realise 'emotional' cognitive machines for more natural and affective interaction & language processing capabilities in cognitive machine
- Cognitive Control of Autonomous Systems
  - realise action-selection and learning capabilities of our envisaged multi-modal cognitive machines
- Decision support
  - Provide intelligent analysis and estimation of cancer care patient support
- Fraud detection
  - Analyse words and meanings to seek truth and honesty

Multimodal Speech Processing

- Developing new and different approaches to speech filtering
  - Consider more than just audio
  - Make use of other modalities
  - Move away from traditional hearing aids
- Cognitive Inspiration
  - Will give overview
  - Demonstration of results
  - Discuss potential component improvements

Hearing – Mechanical Concept

- Sound pressure waves
- Causes fluid vibration in inner ear
  Hair cells send electrical impulses to brain
- Represent different frequencies



Audio-only speech filtering

- Hearing aids aim to compensate
  - limitations in hearing
    - auditory nerve
    - hair cell
    - inner ear damage
- Adjust frequencies to compensate
  - Amplify certain frequencies
- Added sophistication over time

Audio-only speech filtering

- Noise cancellation algorithms
  - Designed mainly to remove non-speech
  - Effective on broadband sound
- Directional Microphones
  - Only pick up sounds from set directions
- Frequency compression
- Programming to adjust settings
  - Buttons or automatic
- Detectors to determine filtering
  - Wind detectors
  - Level detectors etc.
- Two-stage approaches
  - Combining different methods
  - Directional microphones and noise cancellation

Limitations of hearing aids

- Many users do not use full range of potential settings
  - Dislike of directional microphones
  - Limitations of effectiveness of noise cancellation
- Improved results in lab conditions not matched by reality
- Industry research very advanced, looks for incremental improvements in audio only algorithms
  - Linking hearing aids
  - Improved array microphones
  - Rapid polarity changes

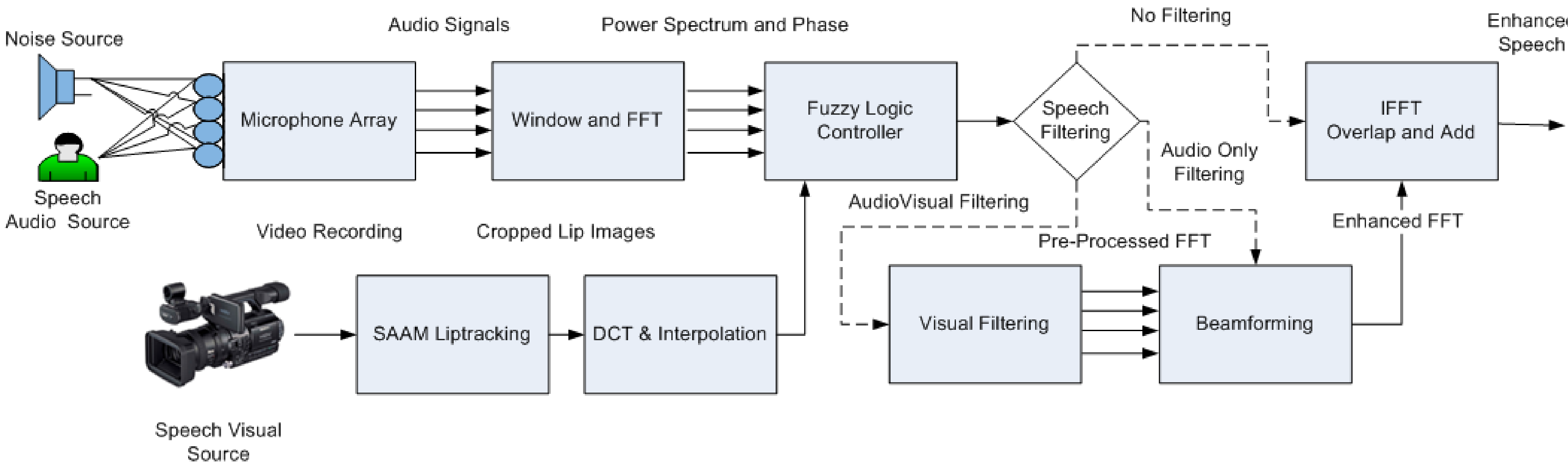
Multimodal Speech Cognition

- Hearing is not just mechanical, but is also cognitive
- Multimodal nature of perception and production of speech
  - There is a relationship between the acoustic and visible properties of speech production.
- Established since Sumby and Pollack in 1954
  - Lip reading improves intelligibility of speech in noise
  - Large gains reported
- Speech sounds louder when the listener looks directly at the speaker
  - Audio threshold to detect speech lower (1-2dB) if audio accompanied with lip gesture
  - Visual information can improve speech detection
- Audiovisual link seen in infants
  - Infants as young as two months can associate between audio and visual

Two-Stage Speech Filtering

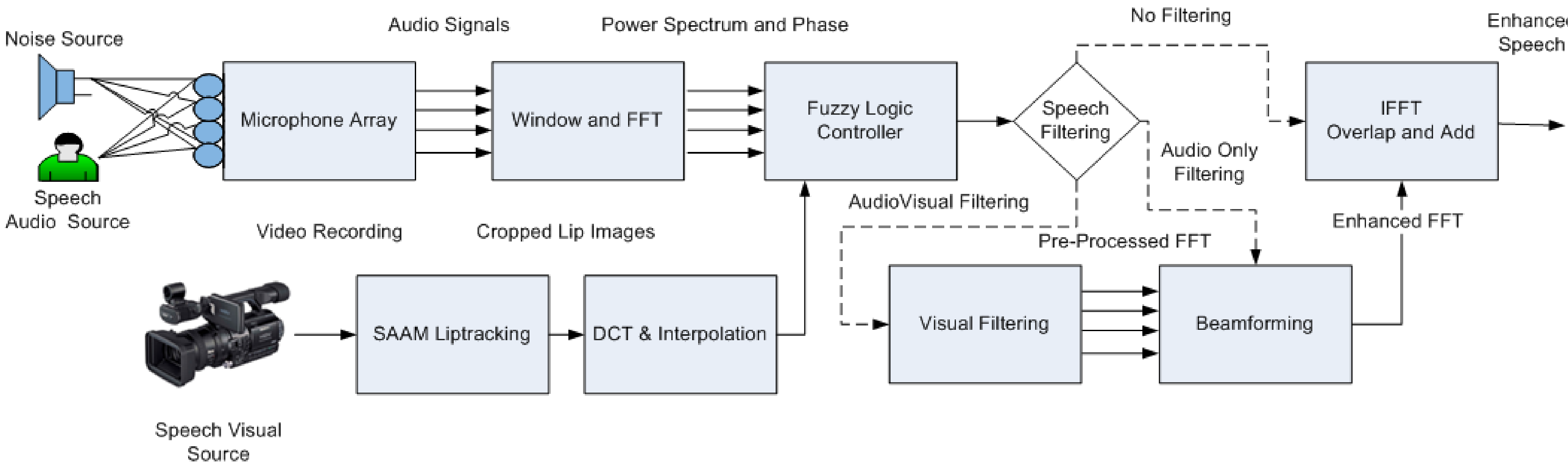
- Initially combines audio-only beamforming with visually derived filtering
- Adds a lipreading component to an audio-only system

Multimodal Speech Filtering



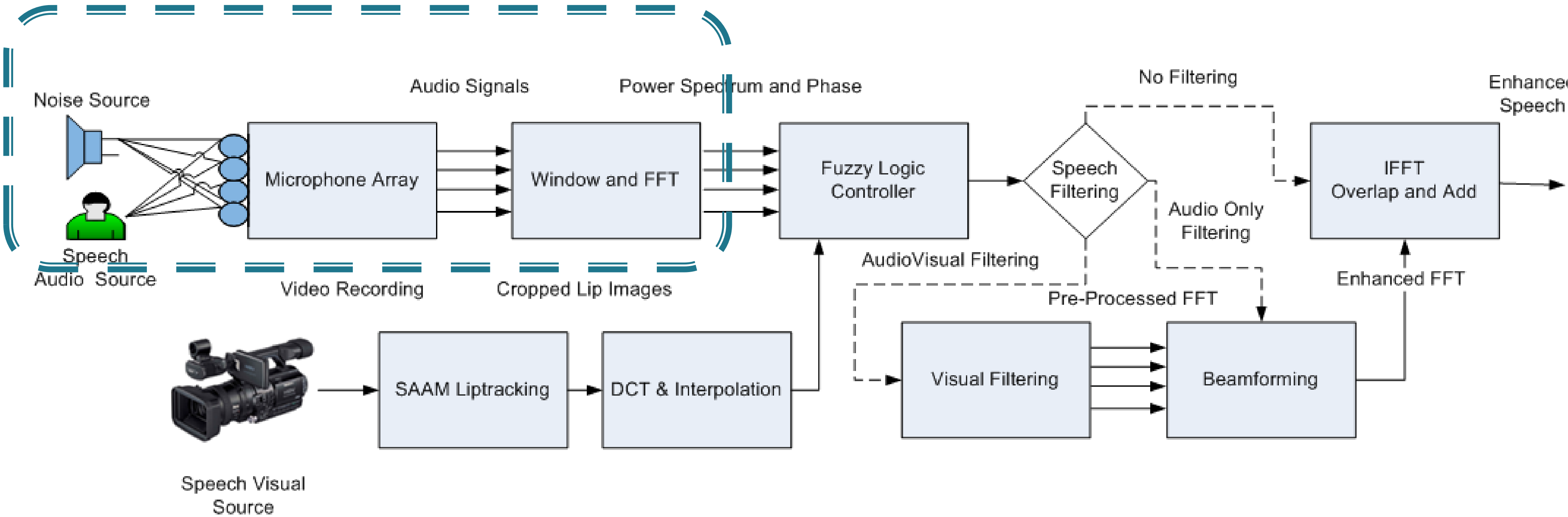
- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Multimodal Speech Filtering



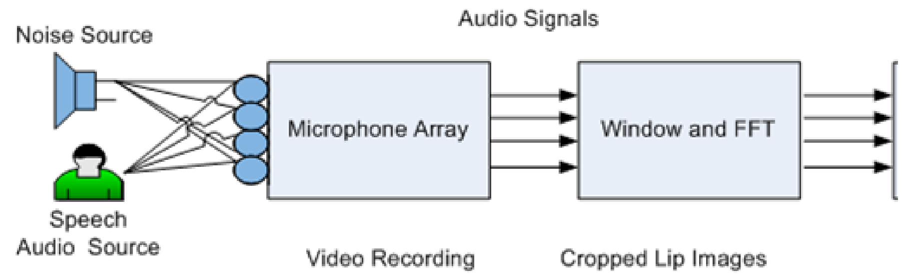
- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

System Components



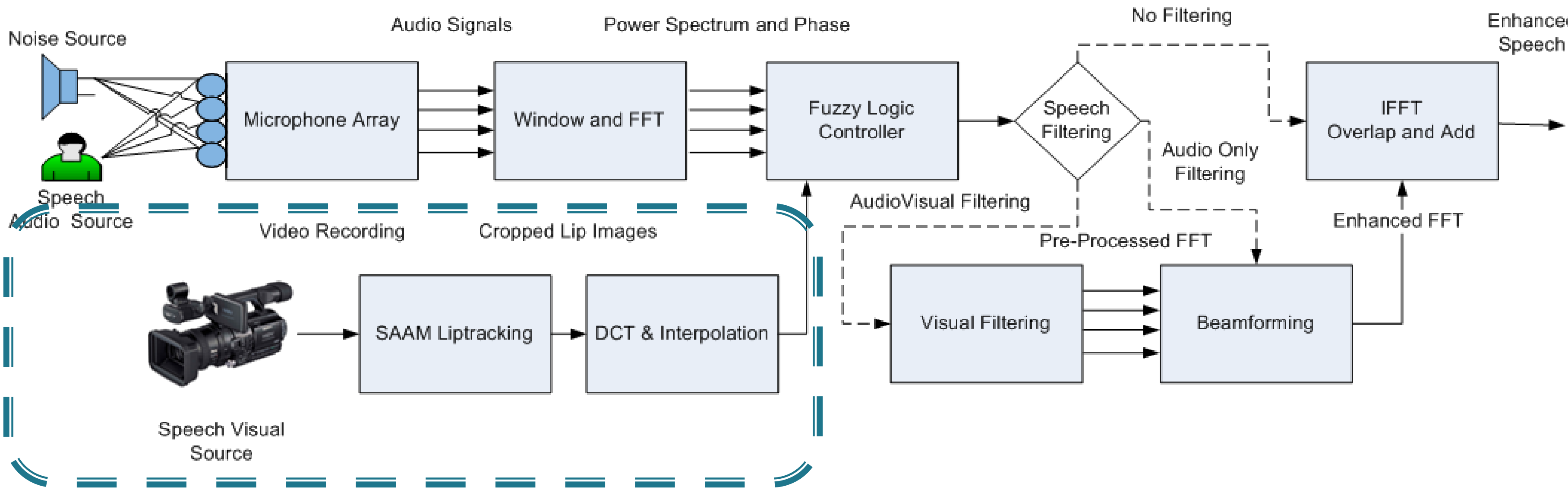
- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Audio Extraction – Current



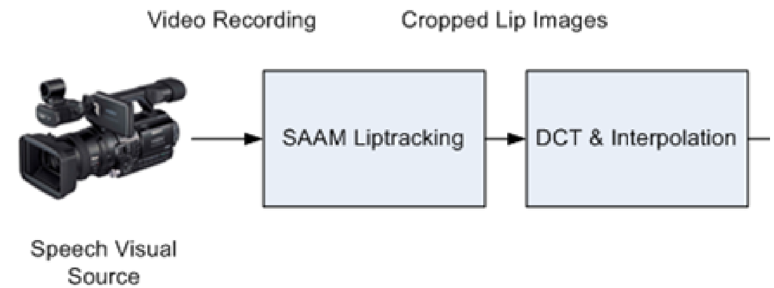
- Simulated Room Environment
  - Speech located at one location in room
  - Noise at a different location
- Microphones then read simulated noisy convolved speech mixture
  - Four microphone array used
  - Produces four noisy speech signals
- Each signal windowed into 25ms frames
  - Fourier Transform used to produce 128 dim power spectrum and phase of each signal

System Components



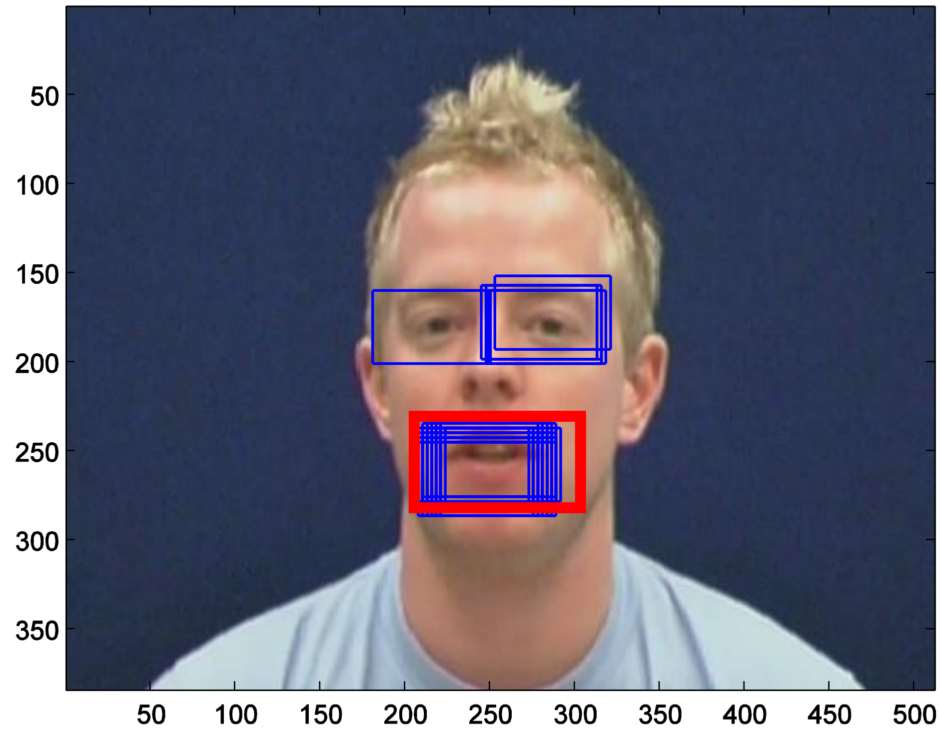
- Audio Feature Extraction
- Visual Feature Extraction
- Visual Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Visual Extraction – Lip Tracking



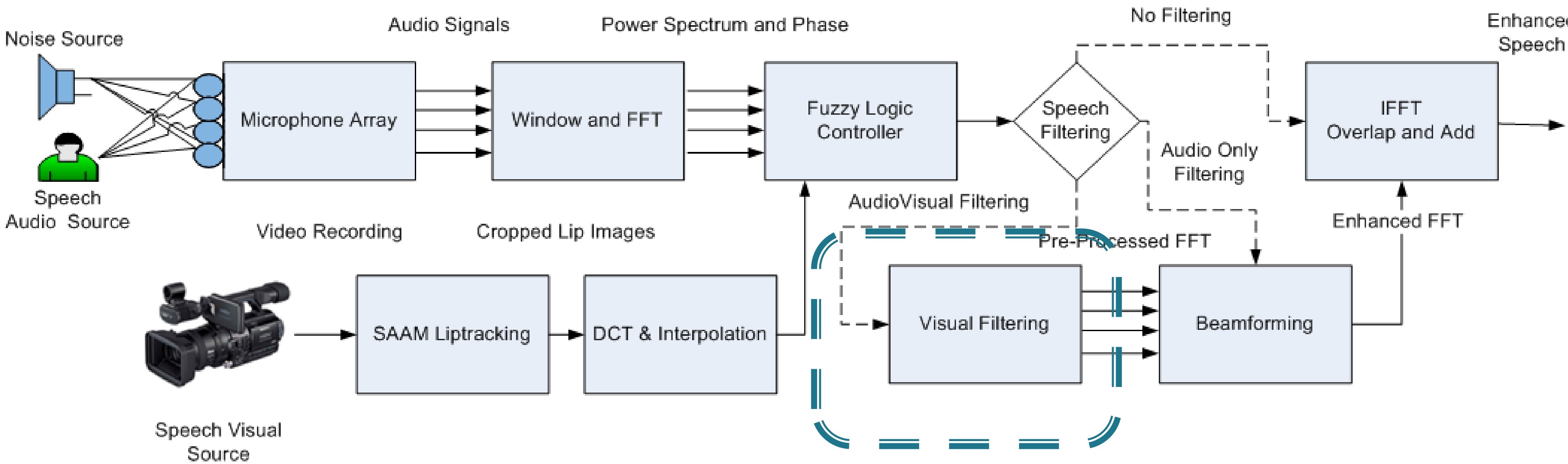
- Video Recordings
- Viola-Jones Lip Detector Used to locate ROI
  - Tracks each image automatically to get corner points of chosen ROI
  - We are using lips as the ROI
  - We extract DCT of mouth region from ROI of each frame
- Automatic lip tracking used to track ROI in sequence
- DCT extracted
- Interpolated to match audio frames

ROI Detection



Examples of lip tracking

System Components

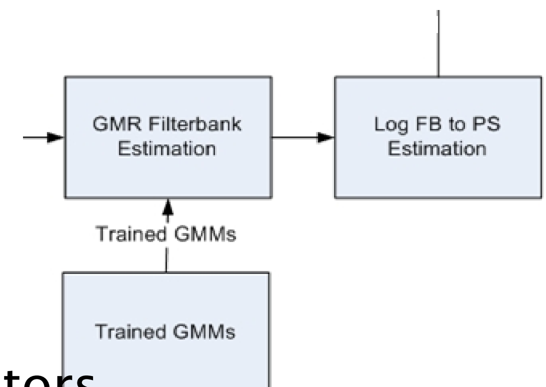


- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

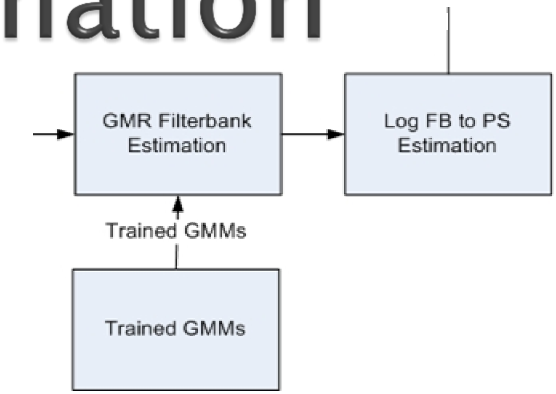
Audiovisual Wiener Filtering

- Wiener filtering
  - W(f) = Est(f) / Noi(f)
  - Estimation of noise free / noisy signal

- Carry out in frequency domain
  - Calculated power spectrum and phase of noisy
  - Estimated noise free power spectrum from visual data
  - Want to modify power spectrum to produced enhanced value
- Noiseless log FB Estimation
- Uses GMM-GMR
  - originally used for robot arm training
  - Uses training data from GRID corpus
    - 400 sentences chosen from four speakers
    - Each sentence contains joint audiovisual vectors



Noiseless Speech Estimation

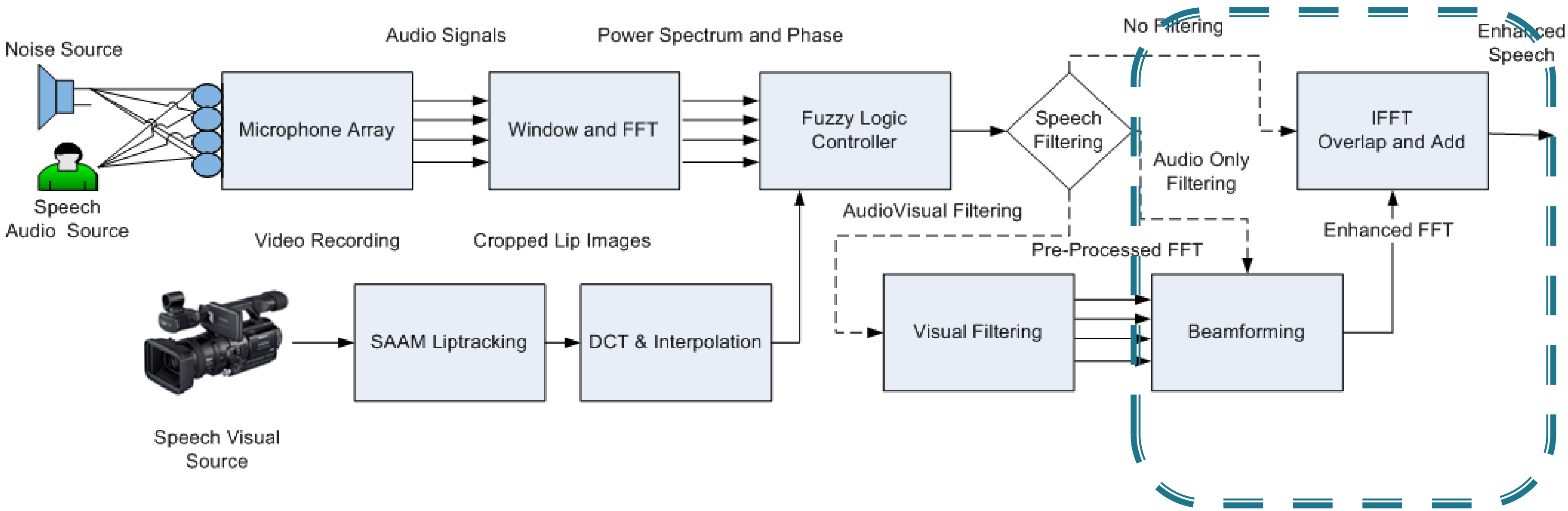


- Noiseless log FB Estimation
- Uses GMM-GMR
  - Originally used for robot arm training
  - Gaussian Mixture Models – Gaussian Mixture Regression
    - 8 components currently used
  - K-means clustering to initialise
  - EM Clustering to train
  - Uses training data from GRID corpus
    - 400 sentences chosen from four speakers
    - Each sentence contains joint audiovisual vectors
  - Allows us to estimate audio frames, given visual

Noiseless Speech Estimation

- Visual DCT vector input for each speech frame
  - GMM – GMR produces a smoothed estimate of equivalent audio
  - Attempts to predict speech fb vectors from visual information
- Power Spectrum Interpolation
  - 23 dim Log filterbank vector interpolated with Pchip
  - To create 128 dim PS estimate of noise free system
  - This can be used as part of a wiener filtering approach
- Currently still Early stage
  - Errors in estimation
    - Results in distorted speech
  - Result of using simple model and interpolation

System Components

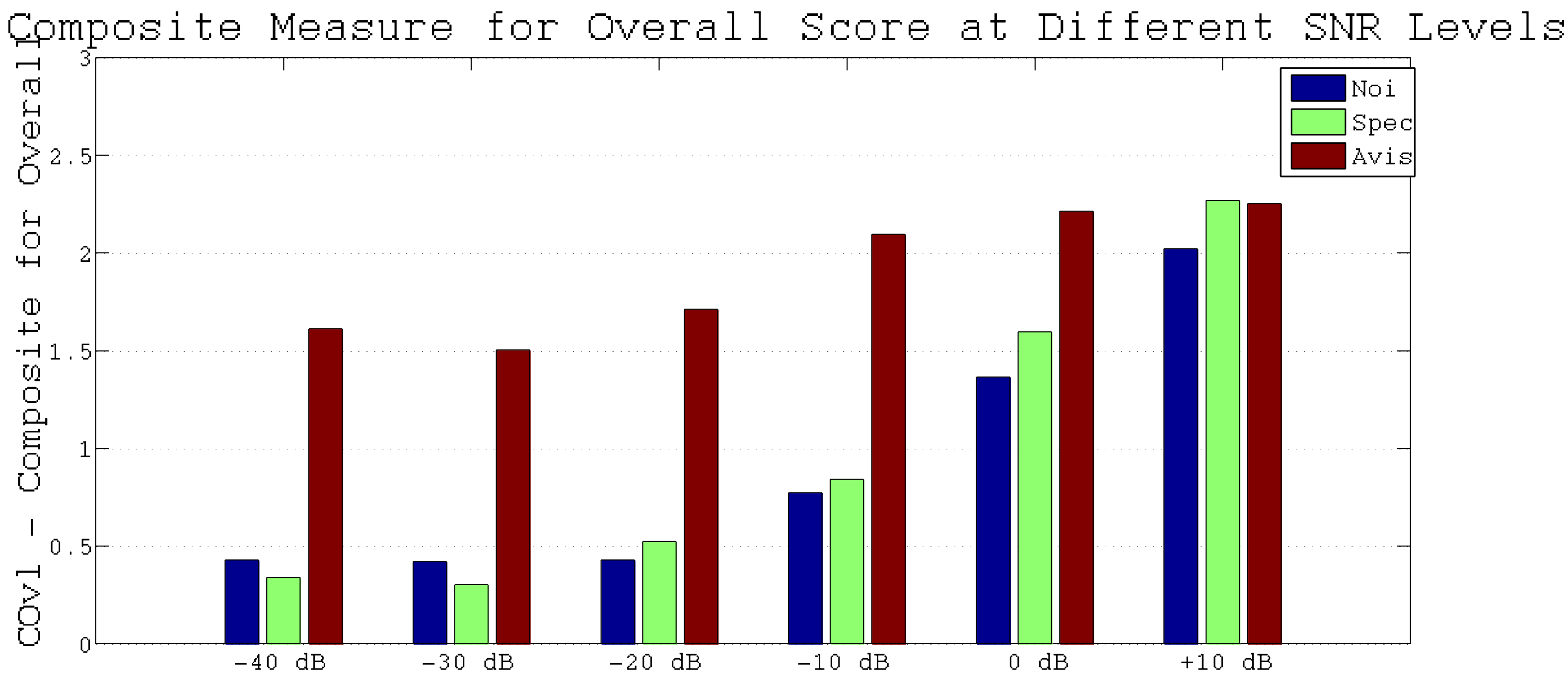


- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Initial Results – Objective Tests

- Composite objective measures
  - Combination of several measures (PESQ, SegSNR)
- Compare to noisy speech and audio-only spectral subtraction
- Consider in very noisy environments
  - SNR from −40dB to +10dB
- Test Data
  - 20 sentences from the GRID Audiovisual Corpus, taken from four speakers
  - Aircraft cockpit noise added to speech sentences
- Comparison
  - Three versions of each sentence considered
  - Noisy speech with no processing (Noi)
  - An audio only spectral subtraction approach (Spec)
  - Our audiovisual system (Avis)

Results – Objective Tests



Results –Objective Tests

- Significant improvement found at very low SNR levels (-40dB to -10dB)
  - Unfiltered speech and spectral subtraction produce very poor results
  - Audiovisual filtering produces much better results
- Higher SNR levels (0dB, +10dB)
  - Audiovisual filtering continues to outperform other measures

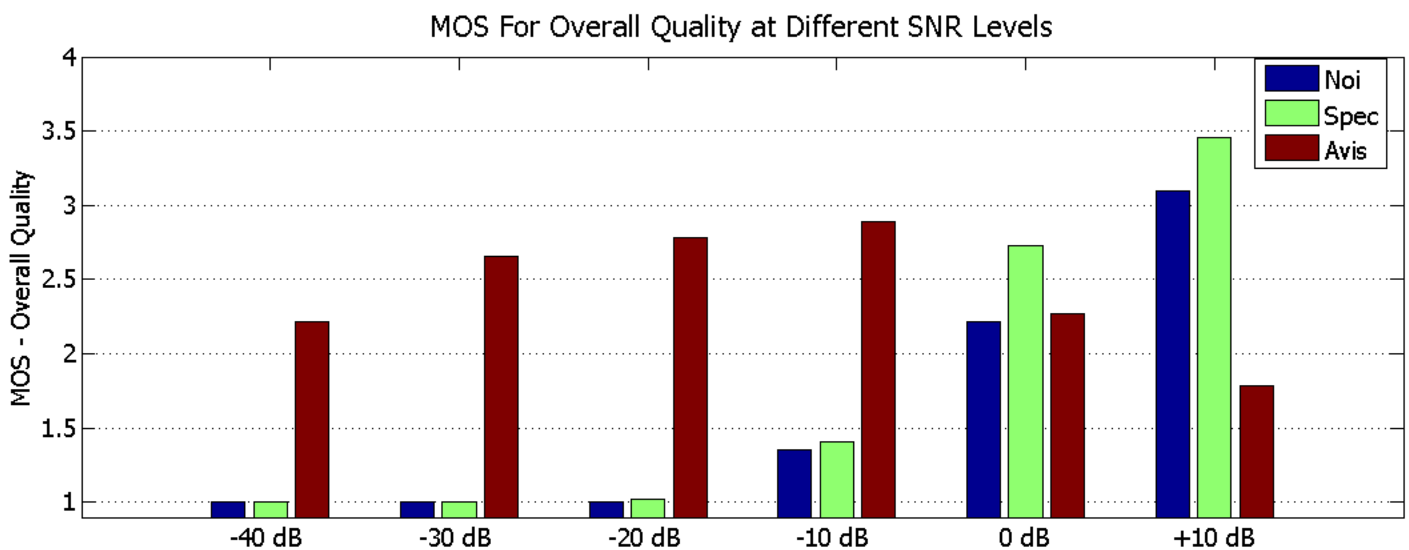
- Overall, Audiovisual the strongest performer
  - Particularly at low SNR levels

- This improvement less prominent when the noise level is lower
  - At +10dB, objective overall score almost identical for noisy, audiovisual, and spectral subtraction
  - Suggests that our system is best at very low SNR levels
    - Environments where conventional approaches might struggle

Results – Subjective Tests

- Primary aim of this work is to enhance speech for human listeners
  - Therefore, listening tests using volunteers to score speech subjectively carried out
  - Assess value of objective measures
- Criteria follows procedures proposed by ITU-T
  - International Telecoms Union Recommendation P.835
- Listener Evaluation
  - Listener listens to each sentence
  - Scored from 1 to 5
  - Results of this assessment used to produce a Mean Opinion Score (MOS) for each criteria
- Listeners listened to each sentence and scored them
  - Same dataset as objective tests
  - Mean Opinion Scores found

Results – Subjective Tests



Results-Subjective, Overall

- At very low SNR levels
  - Spectral Subtraction ineffective
  - Audiovisual results strongly preferred by listeners
  - Performs very well
  - Big improvement seen in terms of preference
- In less noisy environments
  - Audiovisual filtering performs very poorly
  - Significant speech distortion introduced
  - Reflected in very low listener scores
- Very strong overall scores at low SNR levels
  - Our system shows a big improvement in these environments
  - Outperforms audio only measure significantly
- Less strong at high SNR levels
  - Primary problem is the level of speech distortion introduced
  - Audiovisual performs very poorly
- Other limitations identified

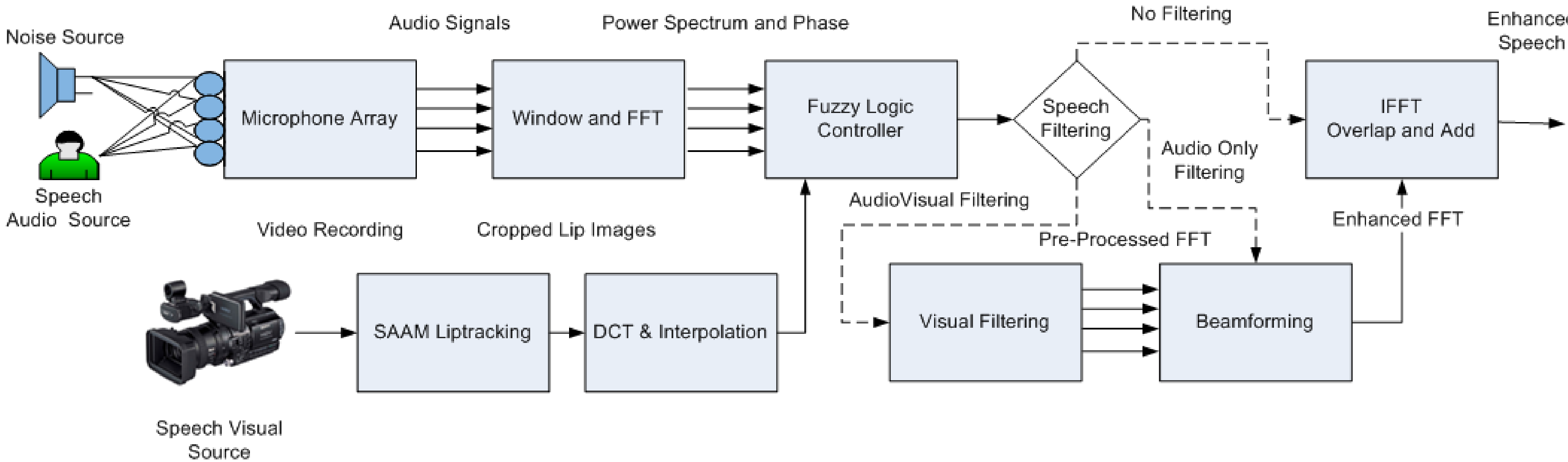
Cognitively Inspired Process Switching

- As stated, more than just lipreading
  - When visual cues used without accurate lip data (dubbing similar audio over lips)
  - Gain in speech intelligibility reported
- Also demonstrated by the well-known McGurk effect
  - Audiovisual illusion (demonstrated by dubbing a phoneme video with a different sound)
  - Often, a third phoneme is heard
  - For example, a visual /ga/ combined with an audio /ba/ is often heard as /da/.
- People do not stare at lips all the time
  - Focus on eye region predominantly
  - More use of lips in noisy conditions
  - Similar experiments on primates
    - gaze focused on eye region, focus on lip region during speech

Cognitively Inspired Process Switching

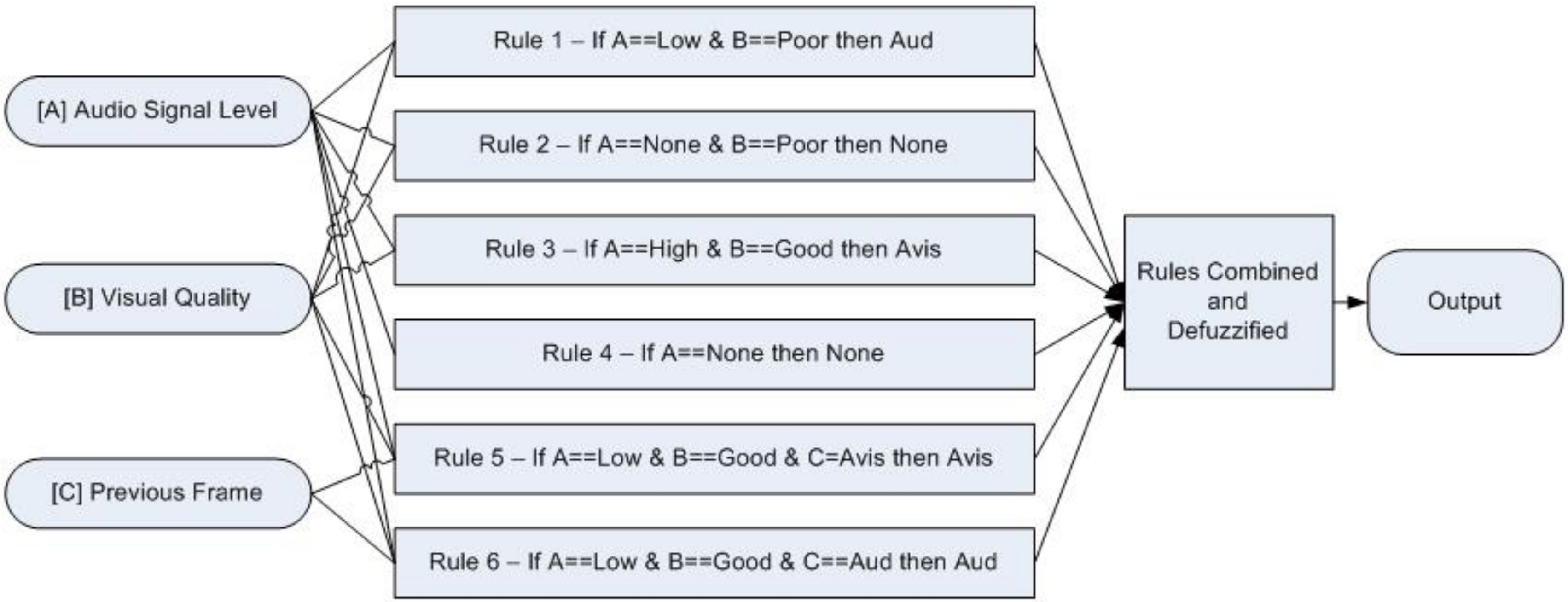
- Audiovisual Filtering can process a noisy signal using lip information
  - Only effective in certain conditions
  - Other times, may introduce additional distortion or not be needed
- Other situations, audio only best
  - Quite noisy, Stable source
  - No visual information available
- Sometimes, unfiltered speech produces better results

Fuzzy Logic Based System

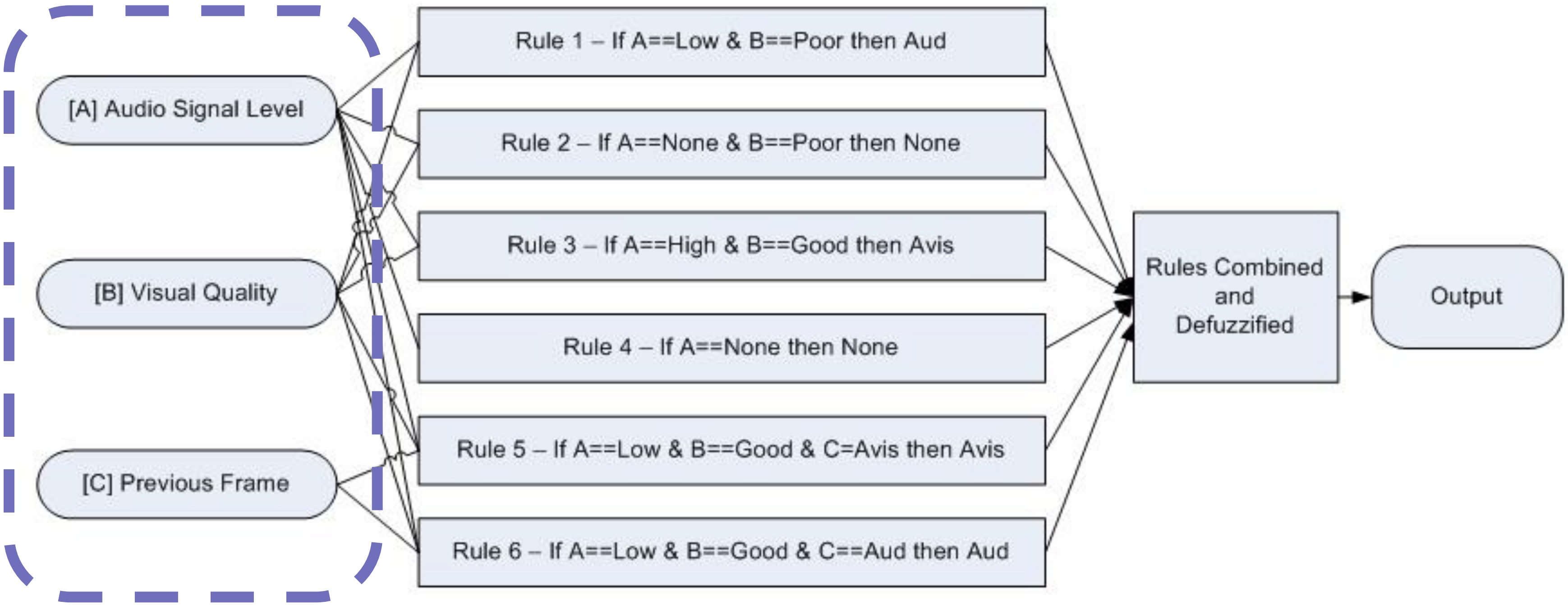


- Fuzzy Logic – Rule based system
  - No human control, all controlled by the system inputs
  - Able to adapt to changing audio and visual environment
  - In real world, noise and environmental factors can change
- Three approaches suited to different environments
  - Two stage filtering
  - Audio Beamforming only
  - No additional processing

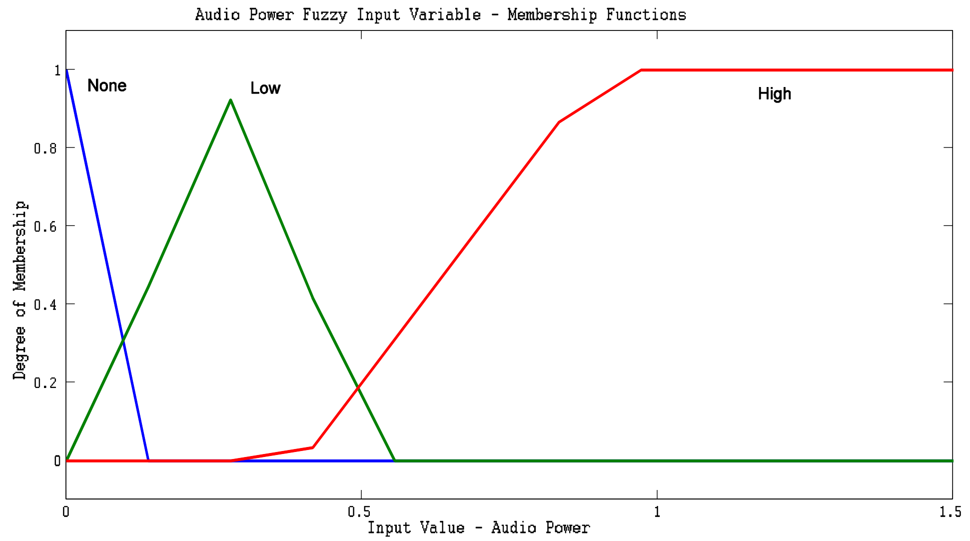
Fuzzy Logic Based System



Fuzzy System – Inputs

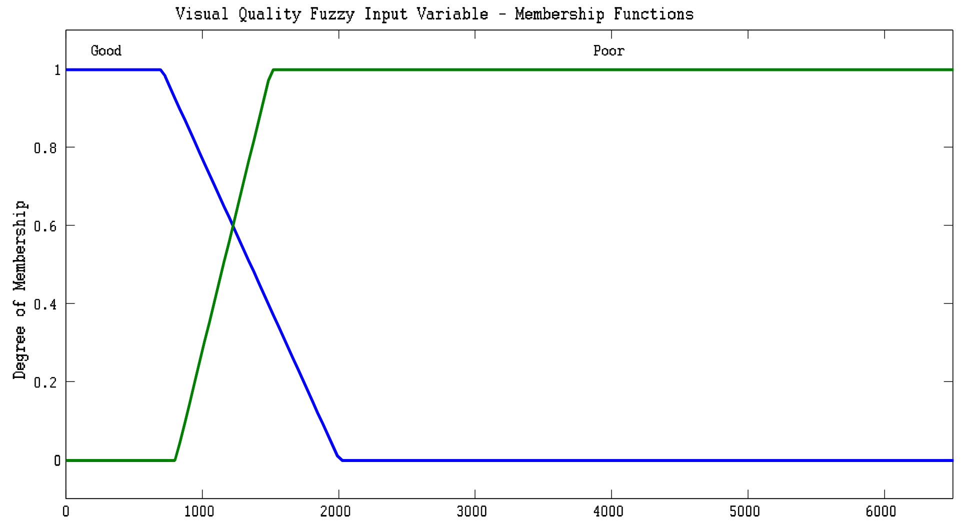


Fuzzy Input – Level Detector



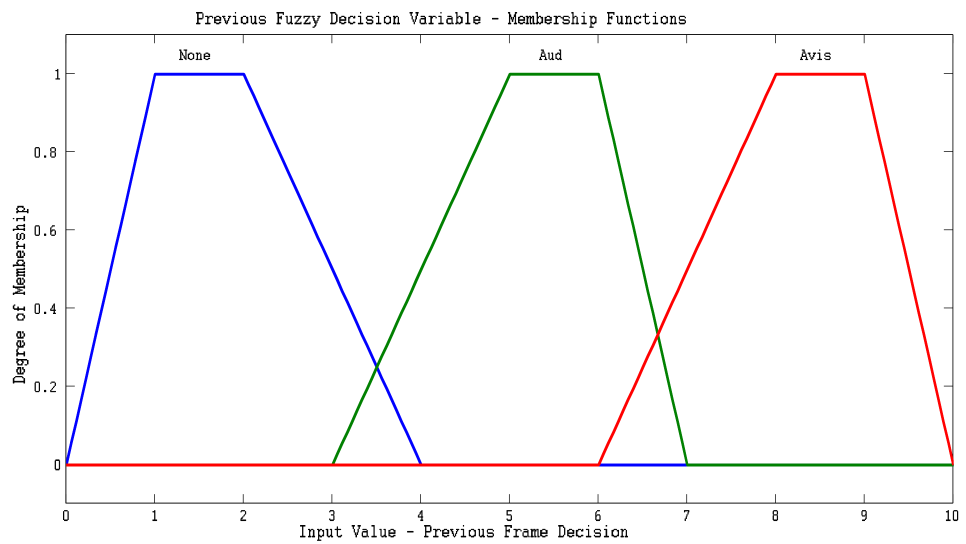
- As used in hearing aids
- Considers level audio power (i.e. how much activity) is in a frame

Fuzzy Input – Visual Quality



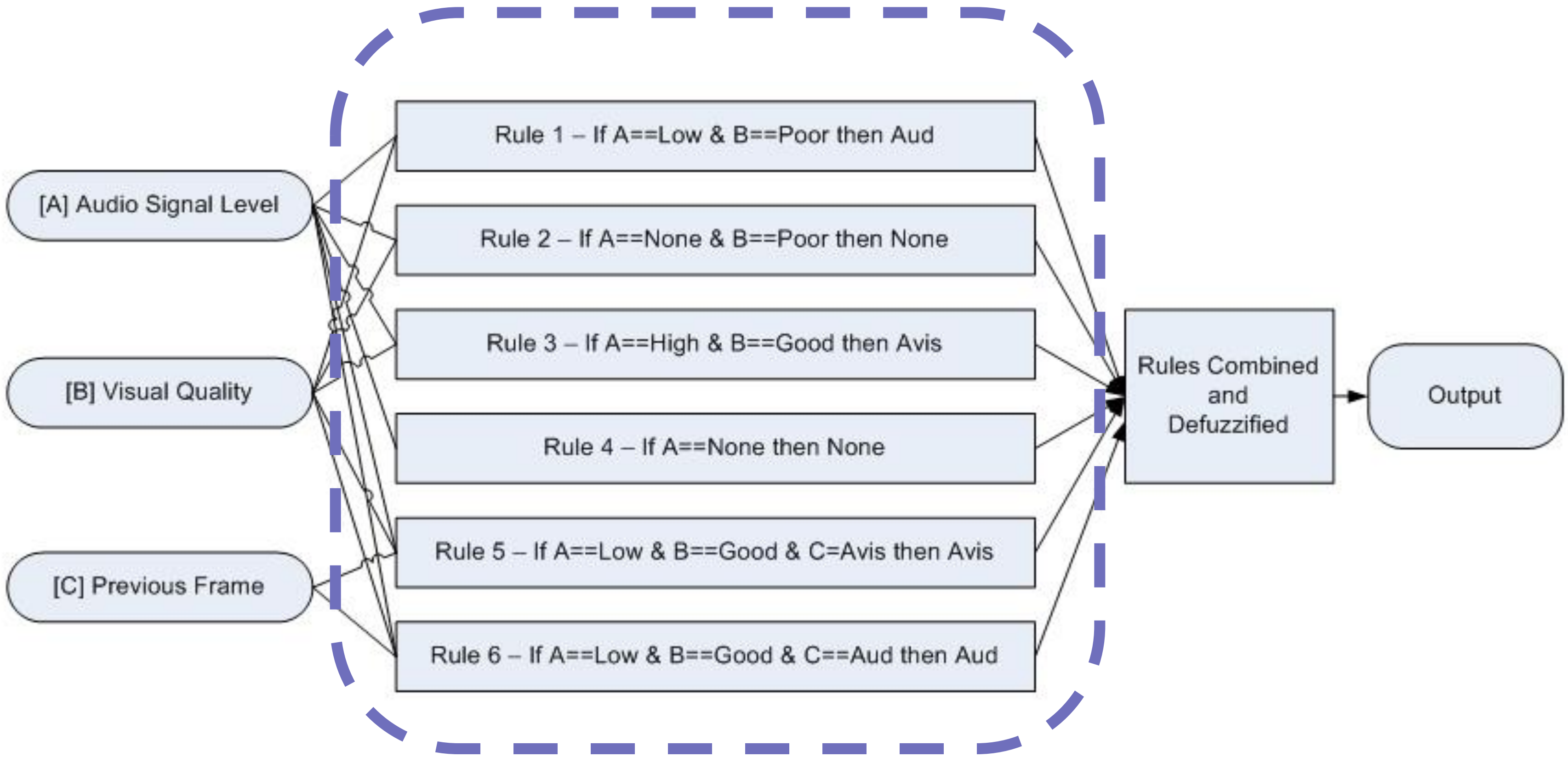
- Level of detail in each cropped ROI
- Absolute value of 4th DCT coefficient
  - Value varies image to image, but 4th coefficient value consistent
- Compared to moving average of up to 10 previous good values (takes account of drift)
- Poor quality result in greater difference from mean than good quality
  - Could be wrong ROI, or no ROI detected

Fuzzy Input – Previous Output



- Previous frame
- Takes output decision of previous frame
- Limits oscillation between frames

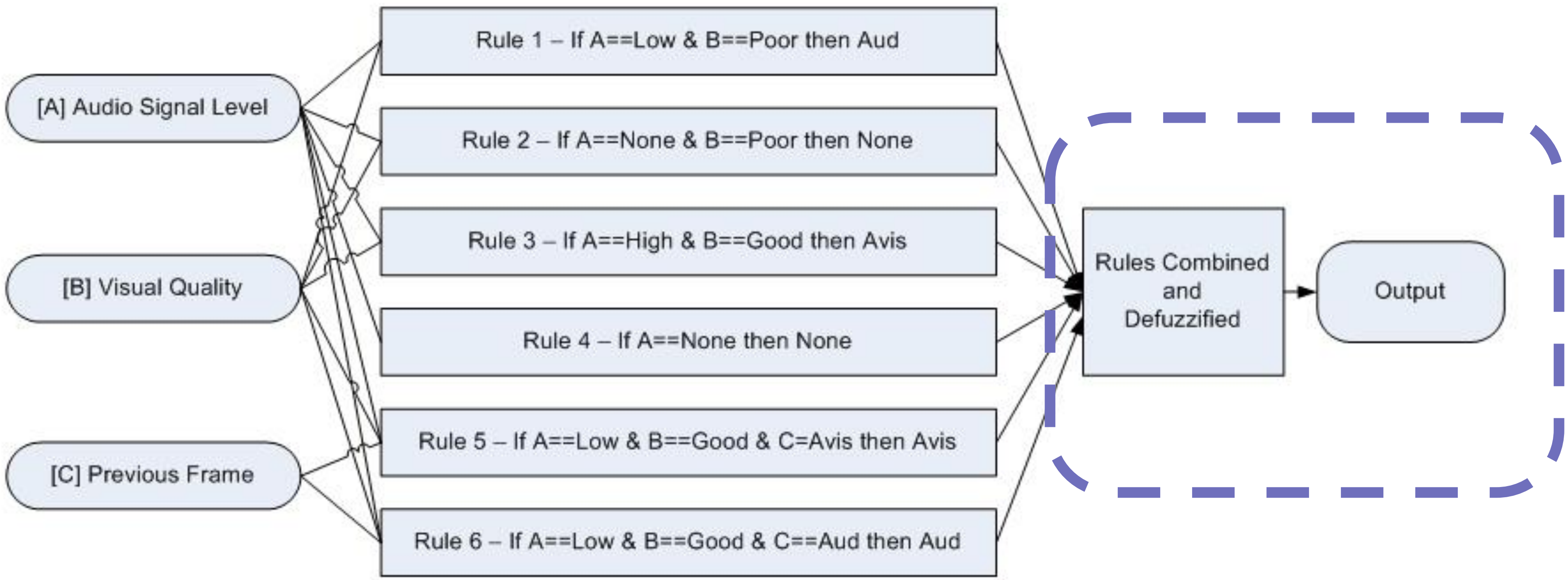
Fuzzy System – Rules



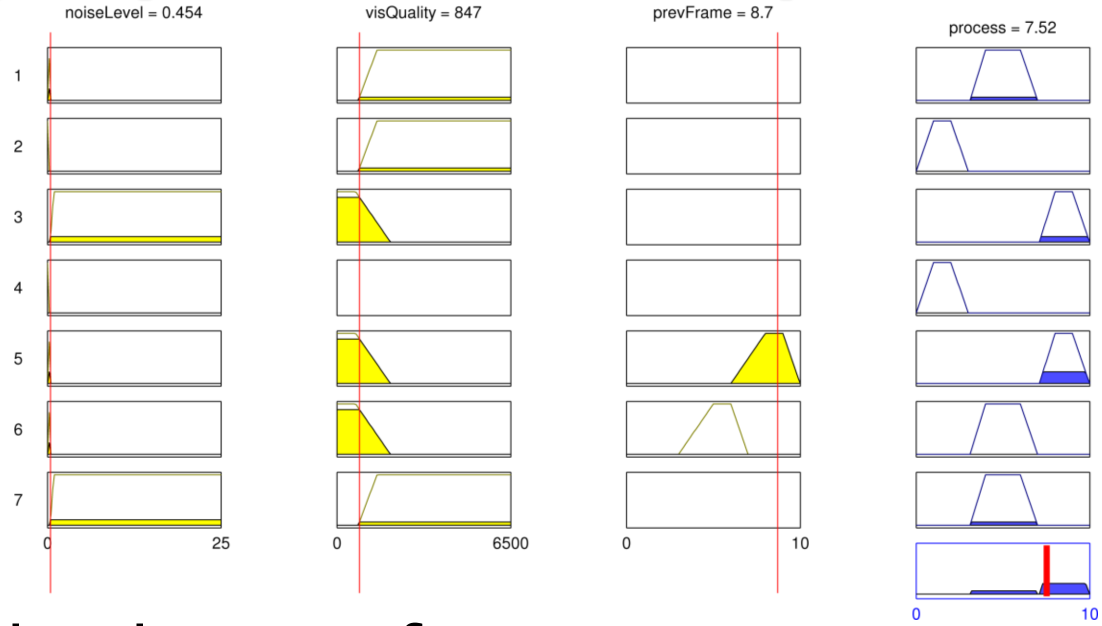
Fuzzy System – Rules

- Fairly common sense
- If very noisy, use visual information
  - But only if good quality visual information is available
- If less noise then use audio-only filtering
  - No need for visual information
- If very low noise, no processing at all
  - Keep background cues
- Why fuzzy?
  - Can be adjusted and tweaked
  - Not always clear which single rule is applicable
  - Thresholds may vary between users

Fuzzy System – Output

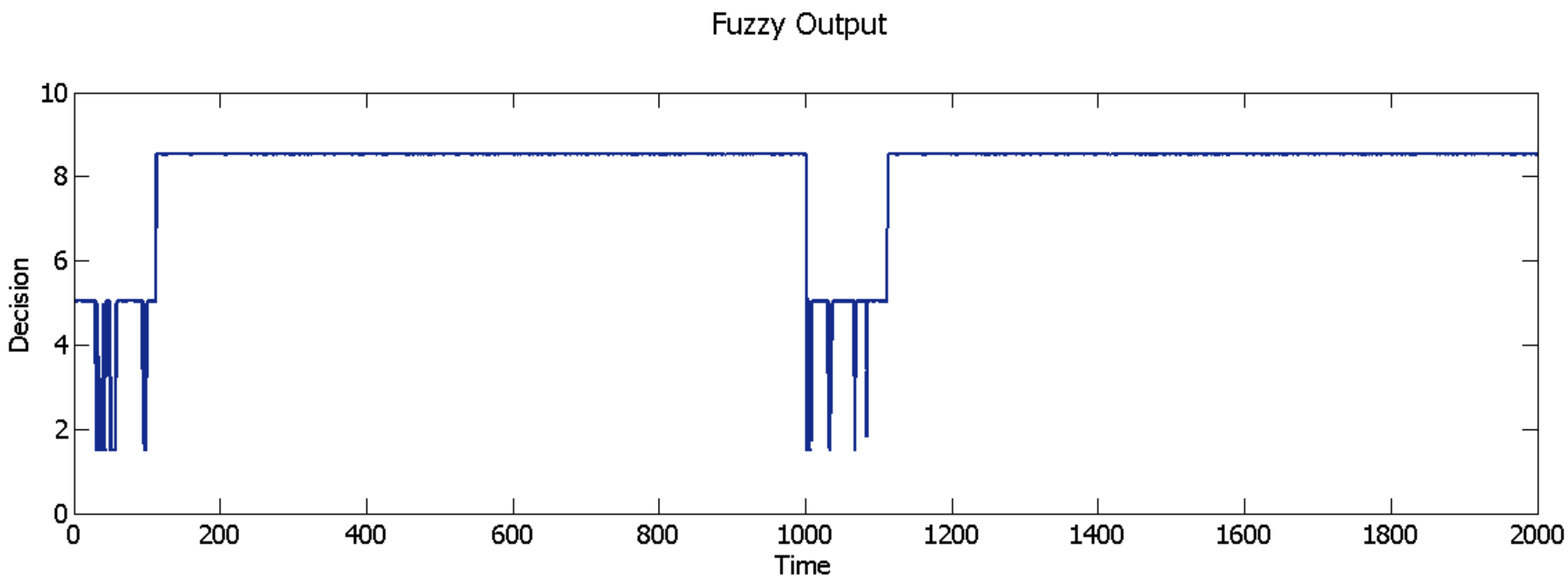


Fuzzy System – Output



- Several Rules may fire
  - E.g. could be threshold of audio and audiovisual
  - Fuzzy, so not part of crisp set
- Defuzzifying picks one final output for each frame
  - Audiovisual (high), audio (medium), none (low)

Fuzzy System – Output



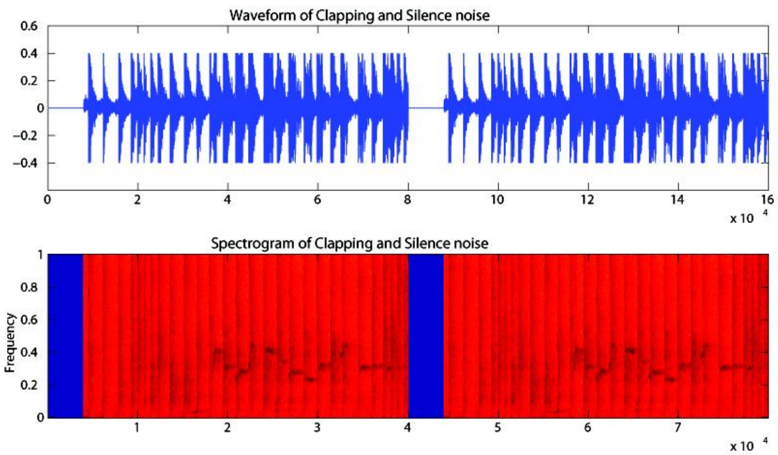
- Defuzzifying picks one final output for each frame
- Each frame is then filtered
- Fuzzy output at each frame used

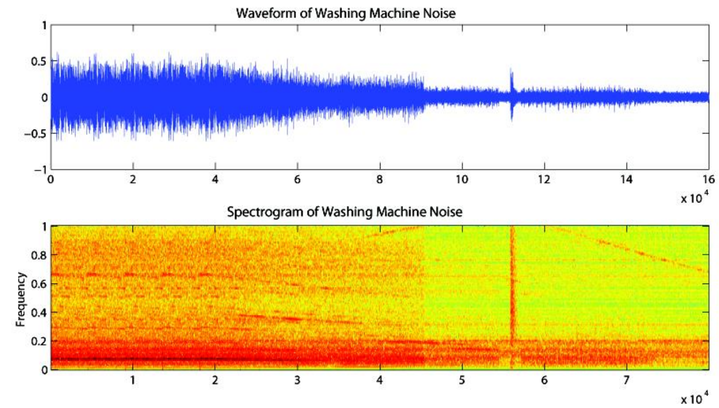
Testing with Custom Corpus

- Corpora in literature not sufficient
  - Limited quantity of “bad” data
  - Generally shot in clean environment
  - Consistent audio and visual noise
- Custom corpus recorded
  - Scottish volunteers
  - Mix of reading and conversation tasks
  - Emotional speech
  - Audio and video files available

Testing: varying visual data

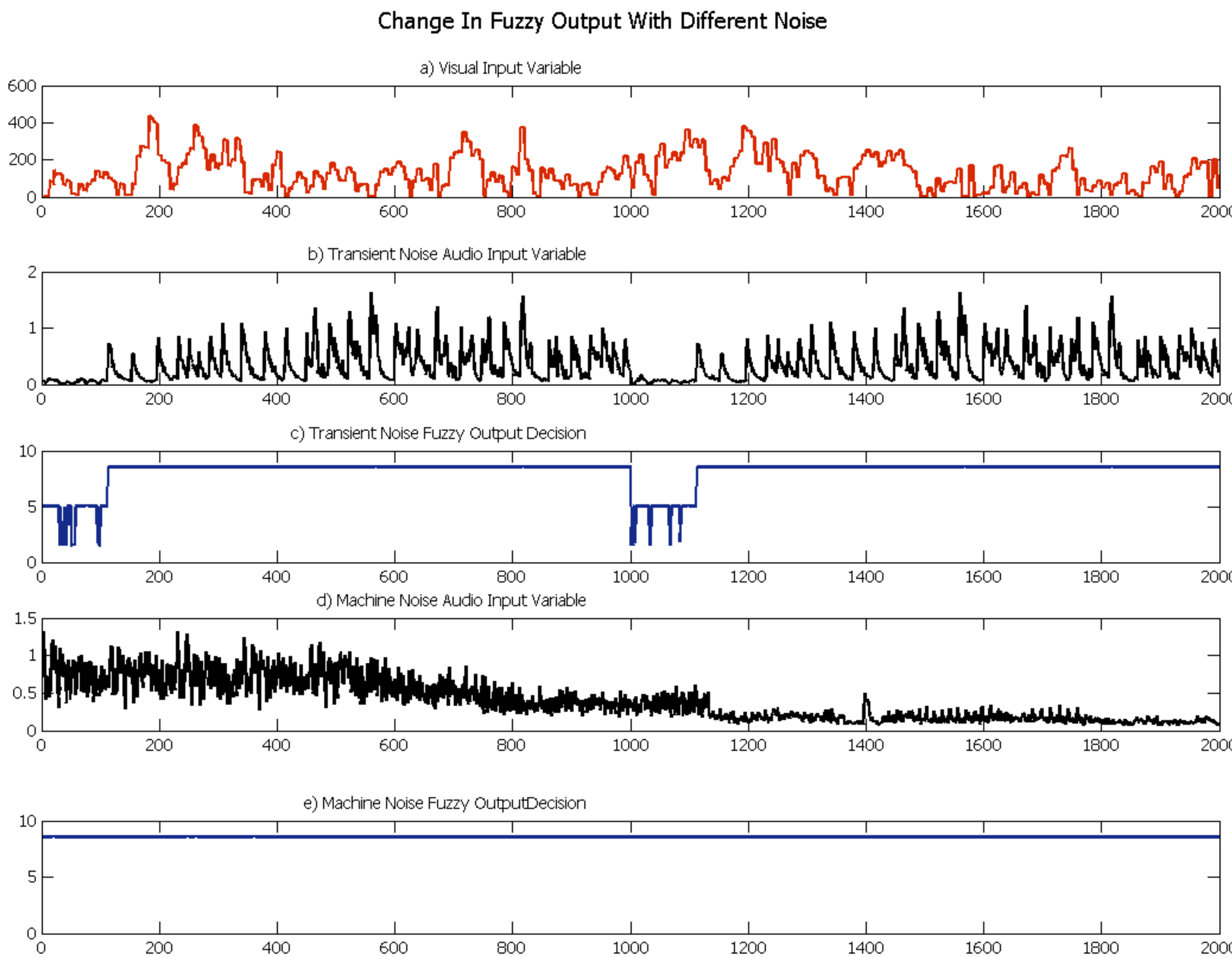
Testing – Different Noise Mixtures





- Speech and noise mixed in a simulated room environment
- Two noise types tested, broadband and transient
- Assume good quality visual information at all times

Results: Different Noise Types



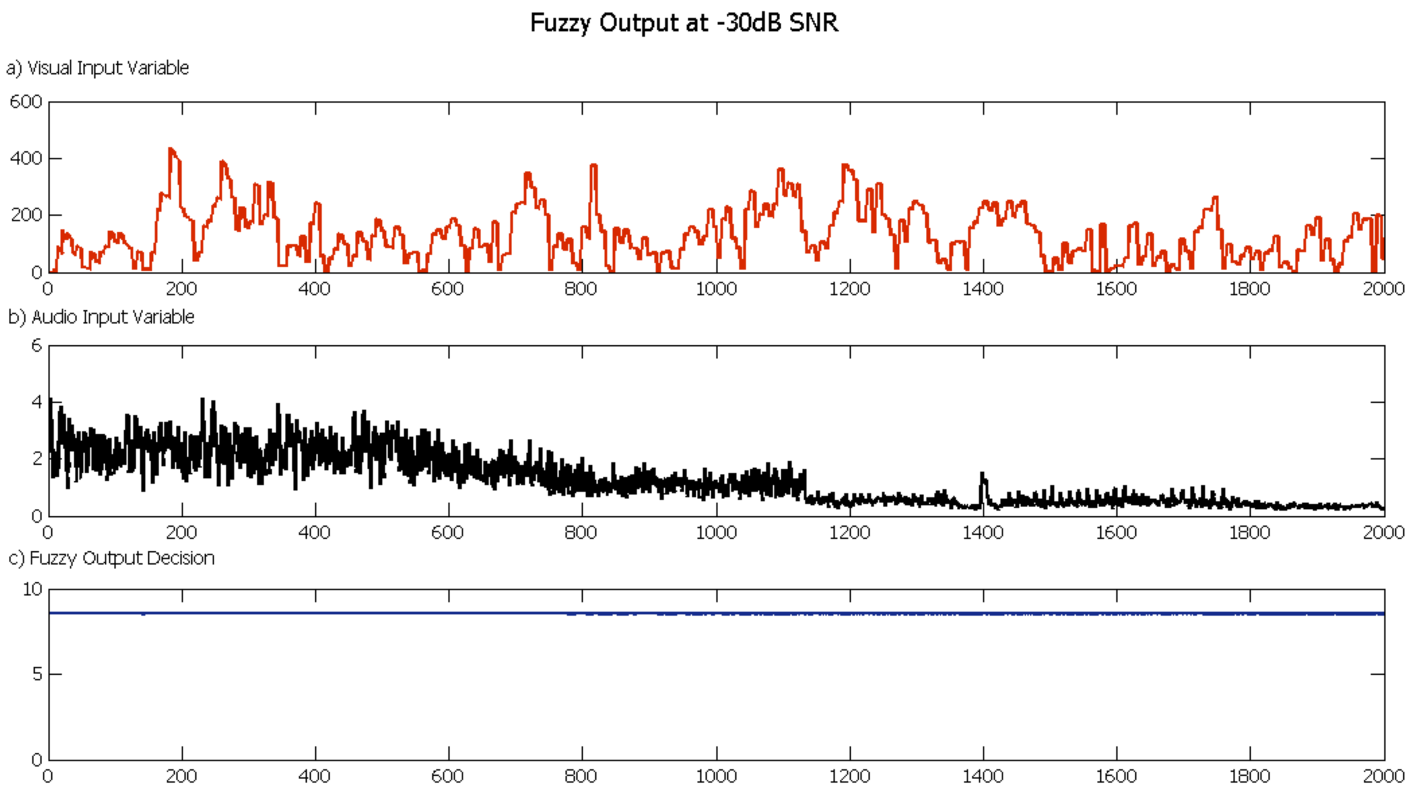
Results: Different Noise Types

- When the same audio and visual speech information is combined with different types of noise, the fuzzy decision is different.

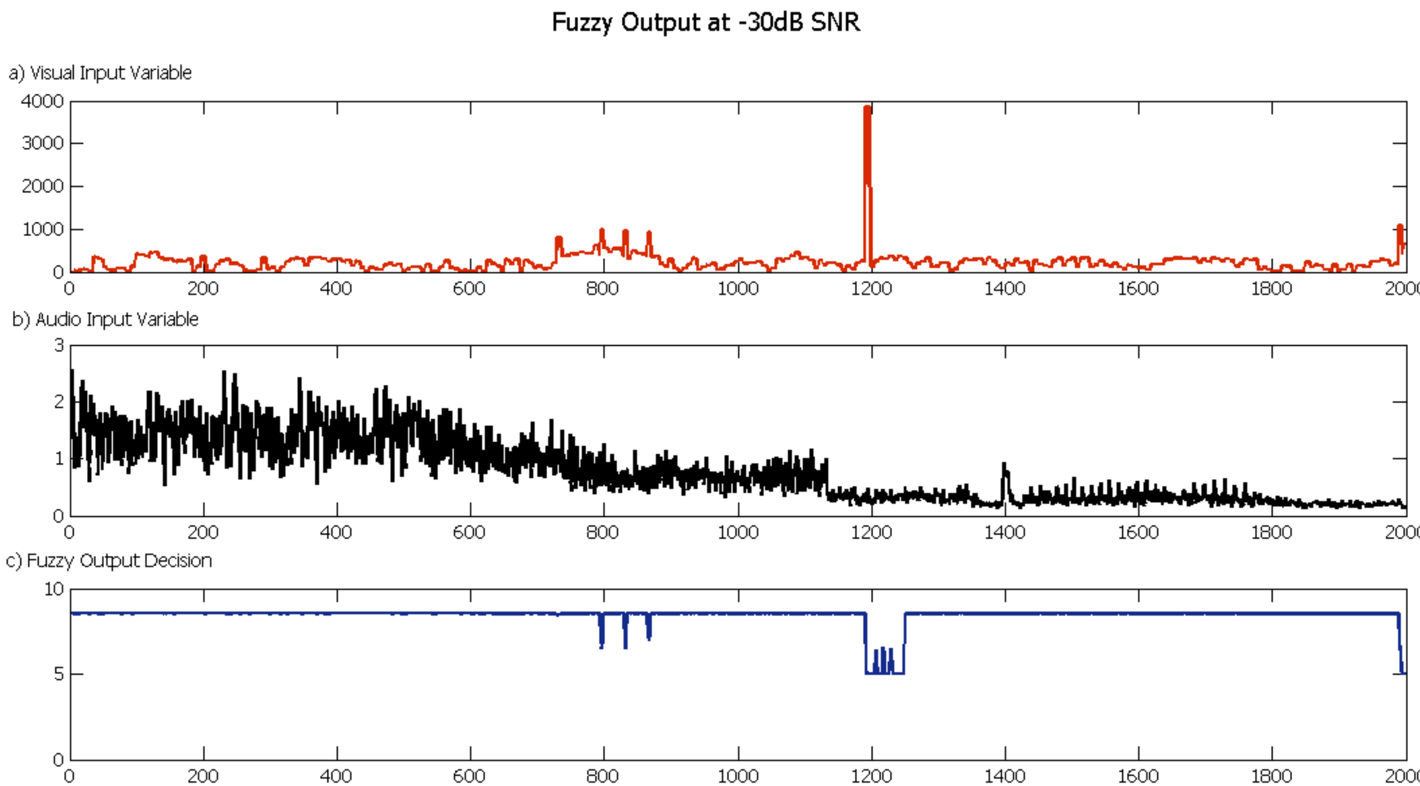
- What about intermittent visual data?
  - It isn’t always good quality!
  - Difficult to find in common corpora
  - Test a number of sentences with the same noise level

Testing: Bad visual data

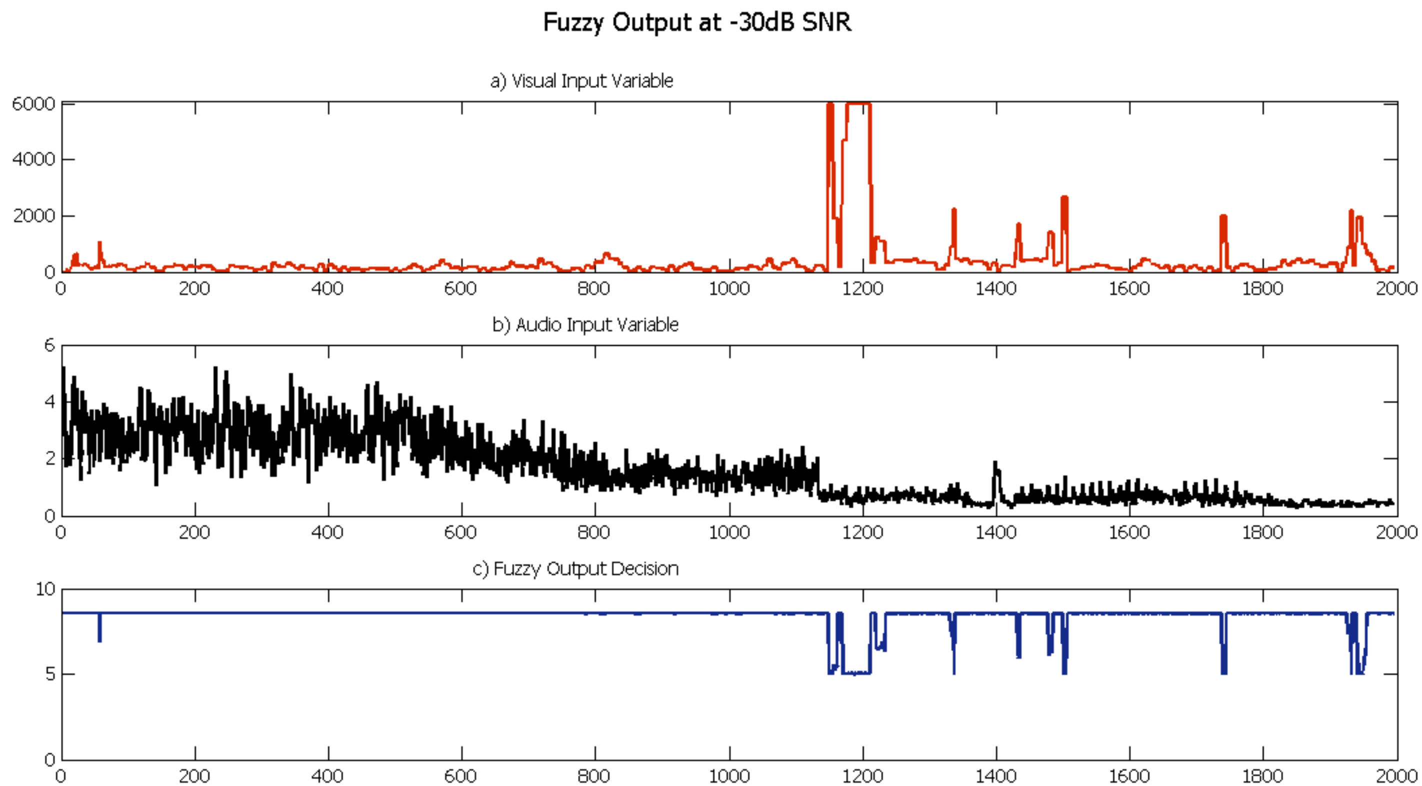
Results: Good Quality Visual Info



Results: Mostly Good Visual Info



Results: Poor Quality Visual Info

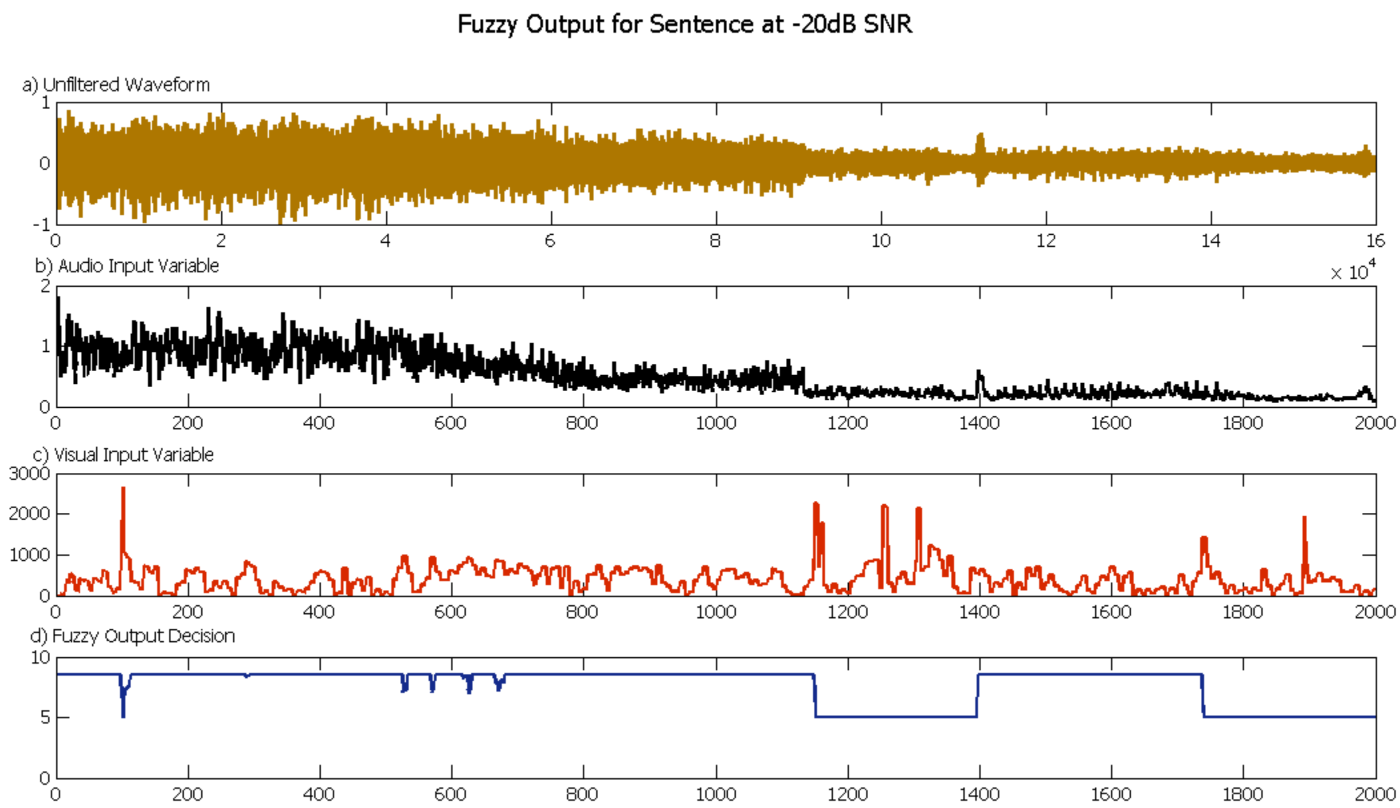


Results: Different Visual Information

- The system will only use visual information if it is available
- If the visual information is not good enough, then it has to rely on audio only
- The switching works

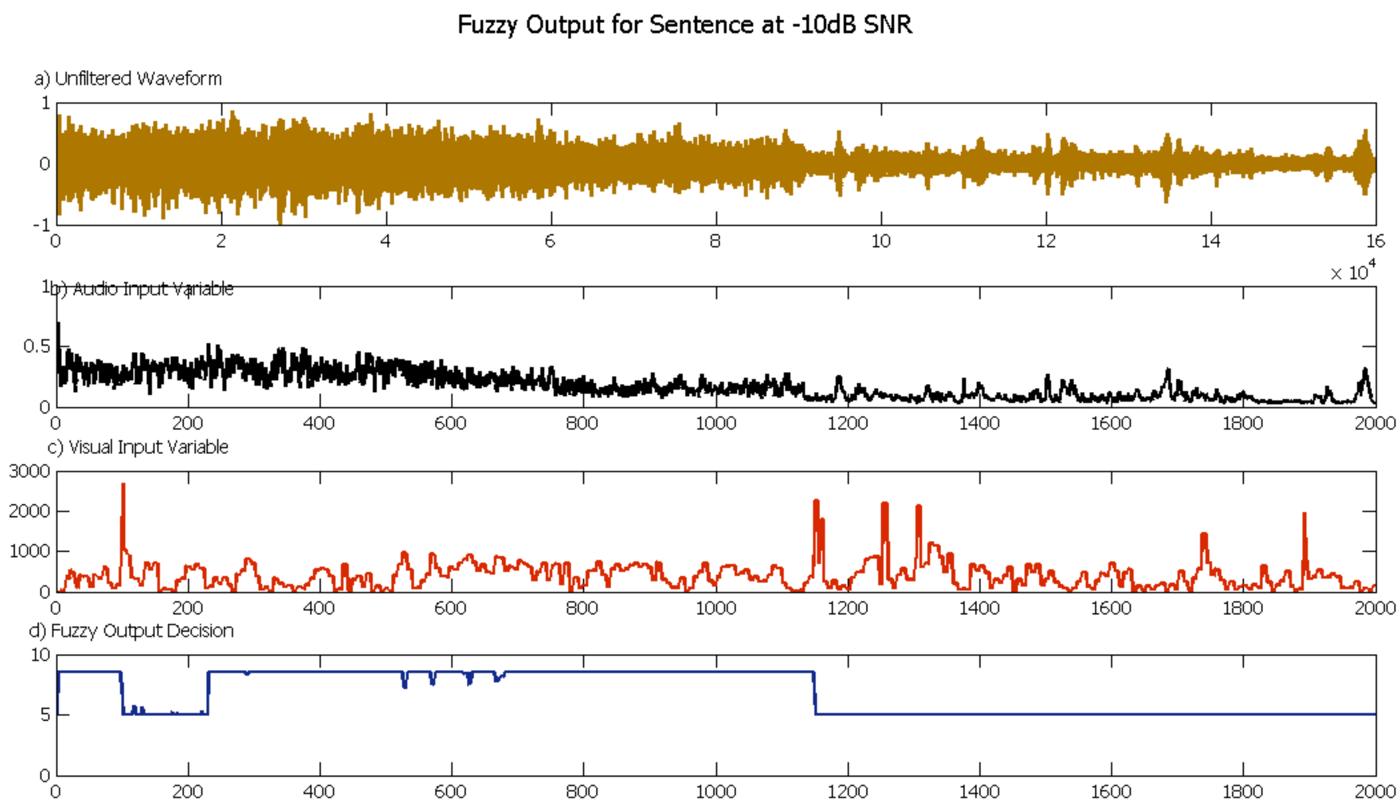
- What about different levels of noise?
  - Changing the SNR
  - Mixing speech and the same noise at different levels
  - Expect different outputs
  - Use of less visual information when less noisy

SNR of -20dB



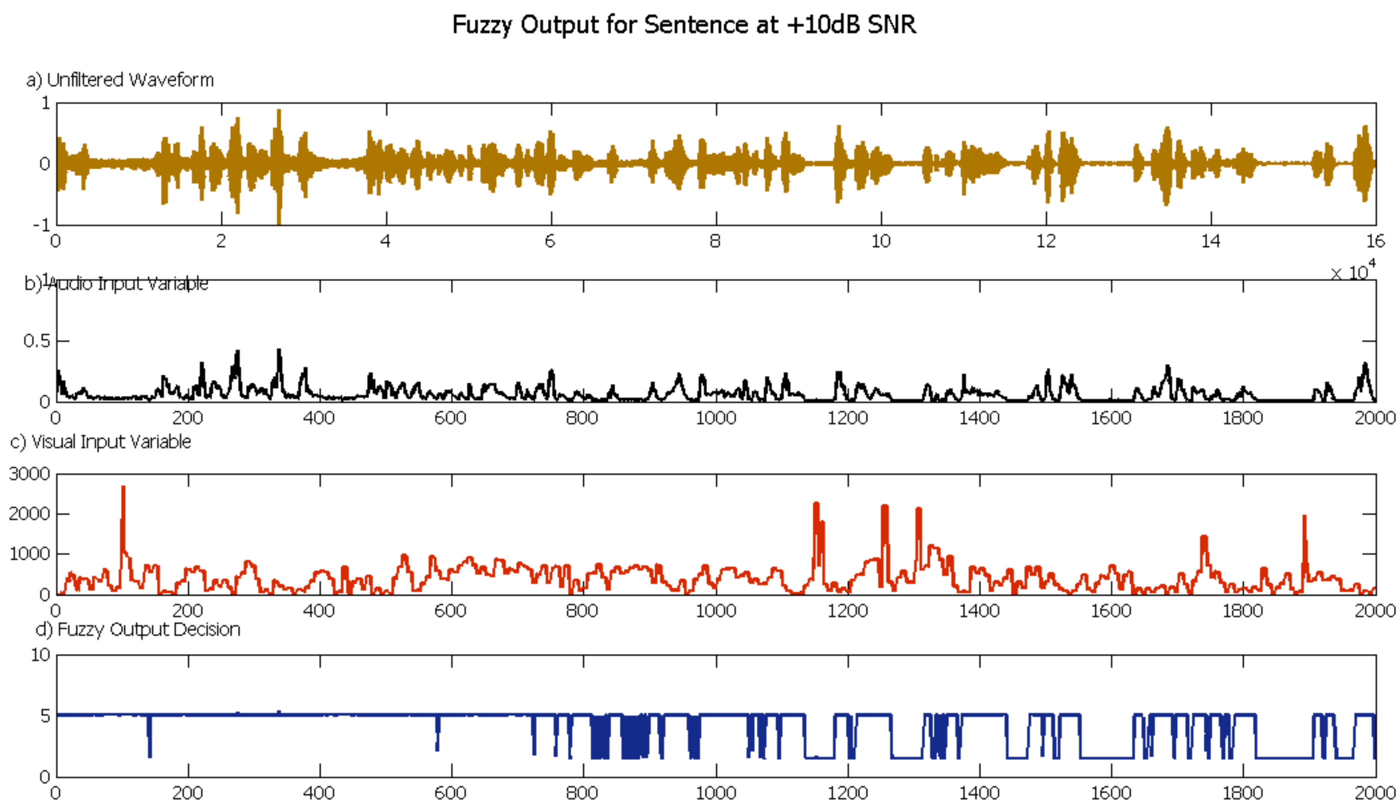
- Mainly audiovisual information when noisy

SNR of −10dB



- Less noise, less use of visual information

SNR of +10dB

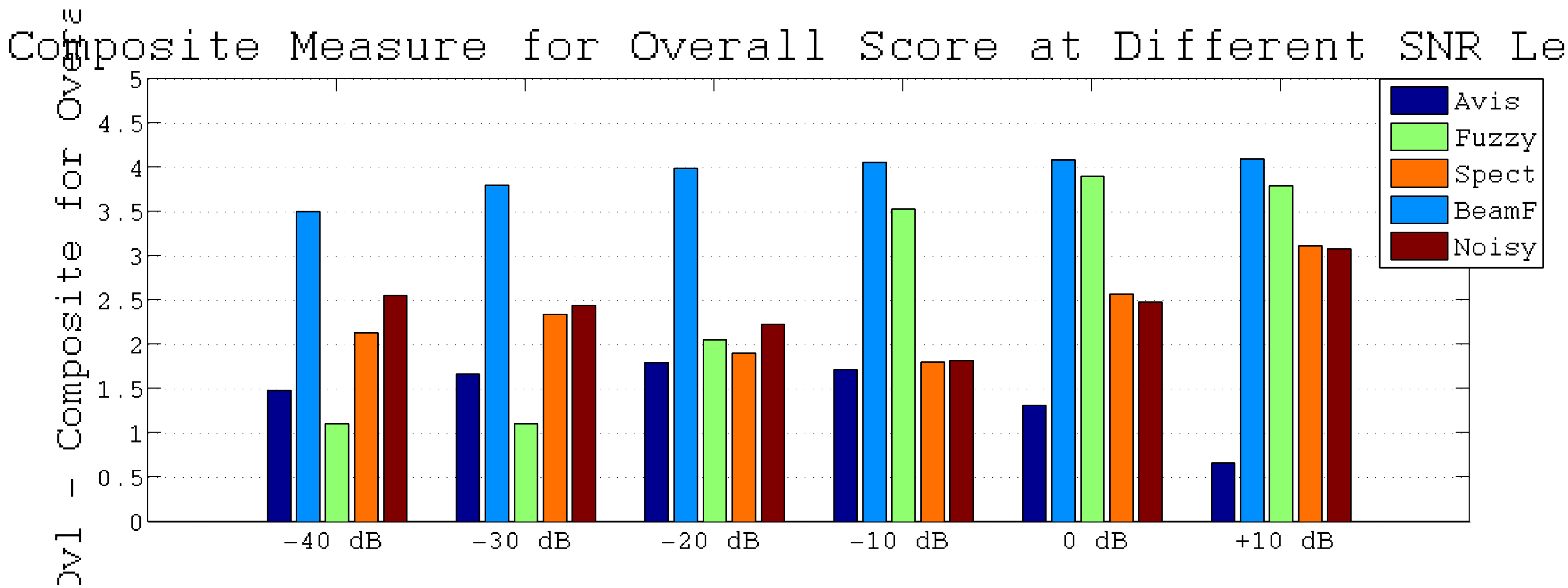


- Audio only or unfiltered
- Mirroring human processing

Fuzzy – Audio Results

- Currently limited
- Limitations with audiovisual model
  - Requires training with new corpus
- Beamforming artificially good
  - Currently using a broadband noise at a static source
  - Simulated room designed for the beamformer
- Other improvements also needed

Fuzzy – Audio Results



Fuzzy – Audio Results

- Currently limited
- Limitations with audiovisual model
  - Requires training with new corpus
- Beamforming artificially good
  - Currently using a broadband noise at a static source
  - Simulated room designed for the beamformer
- Other improvements also needed
- Shows that fuzzy switching works as expected
  - Uses avis in very noisy environments
  - Audio only in less noisy
  - Not identical though

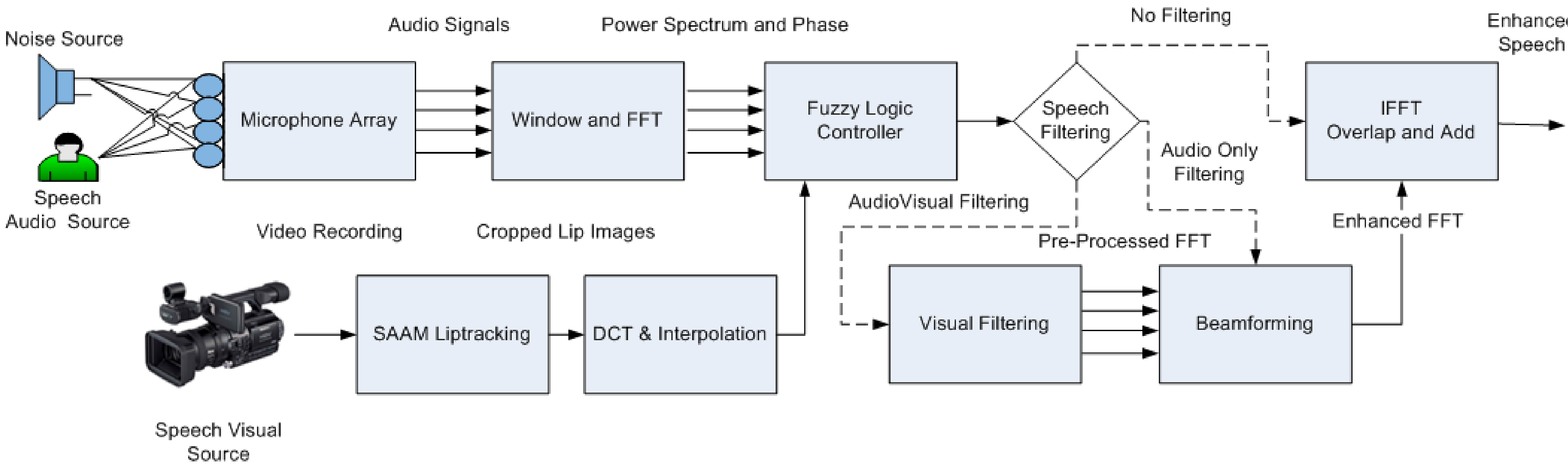
Fuzzy System – Summary

- Possible to build a multimodal fuzzy logic based Speech Enhancement System
  - A more flexible system
  - Cognitively inspired use of visual information
- Can solve problems with individual audio and visual filtering
  - Versatile with regard to different environments
  - Can overcome limitations of individual techniques
  - Can work with wider range of input data
- Use knowledge about the world
- Currently limited audio results

Detailed System Components

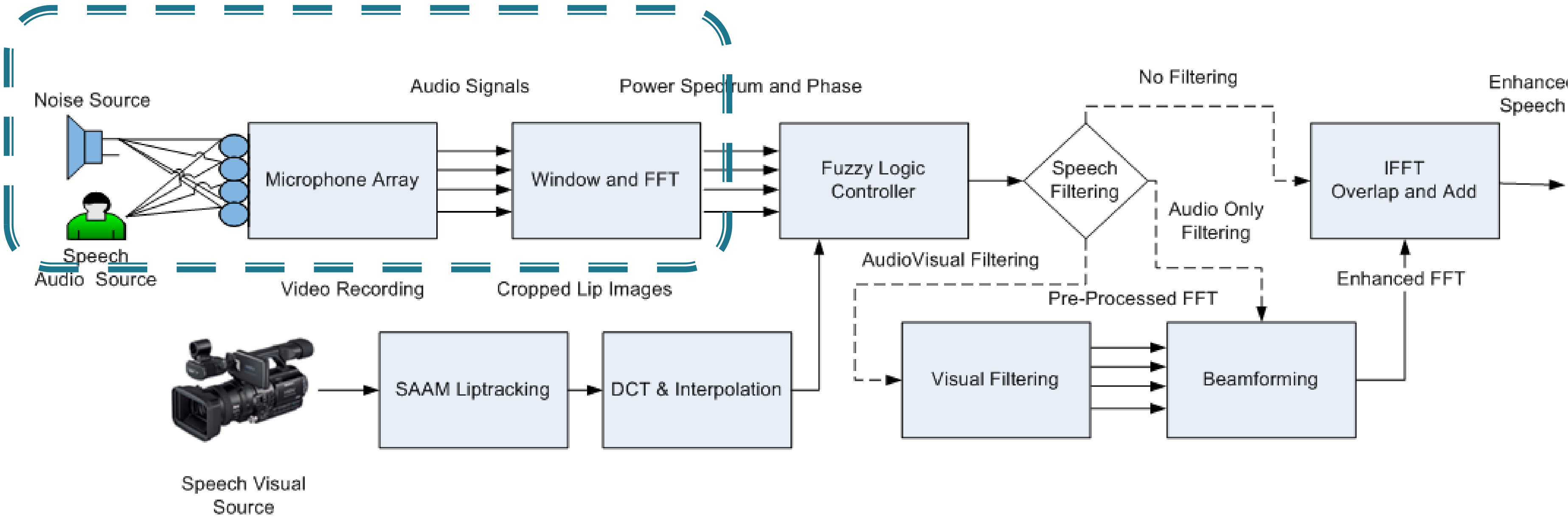
- Not a finished proposal…
- Individual components have been tested
- General framework is satisfactory
- Limitations with results due to early stage of implementation
  - Audio results of fuzzy logic limited due to audiovisual model used
  - Fuzzy logic results depend on limited knowledge of environment
  - Beamforming depends on simulated room mixing
- Much opportunity to upgrade individual components within the single framework

Multimodal Speech Filtering



- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

System Components

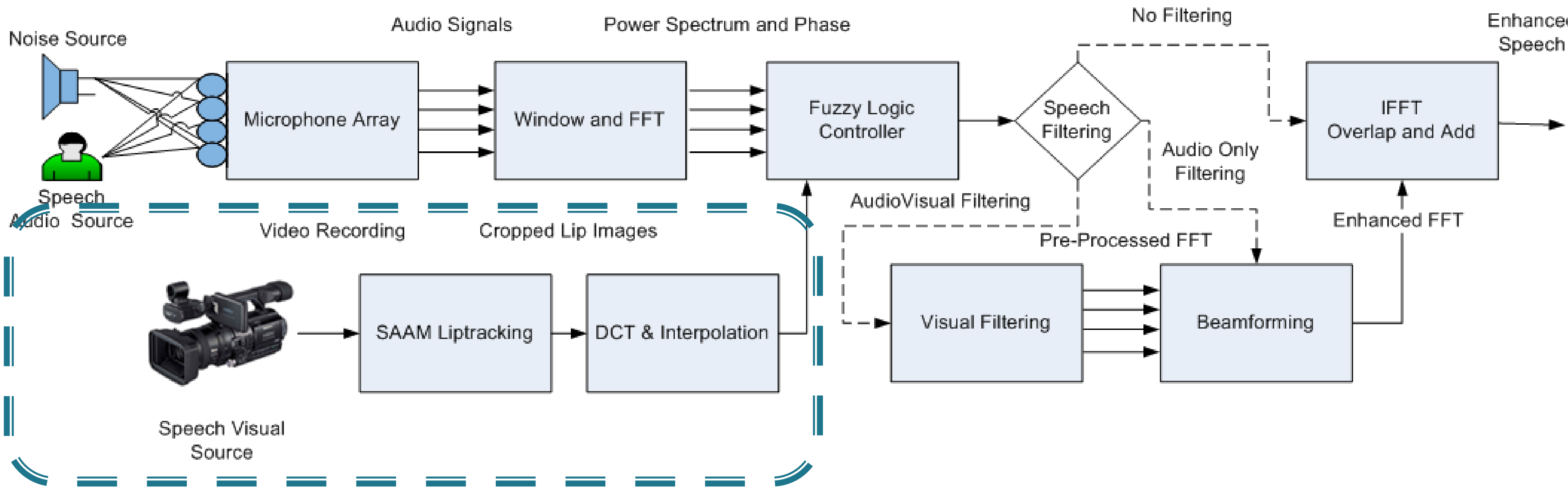


- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Audio Extraction – Potential

- Speech segmentation algorithm developed
  - Biologically inspired based on offsets and onsets
  - Taken from AN spikes
  - Tested separately successfully
  - Awaiting integration
- Improved use of modalities
  - Consider use of AN spikes as input
  - Implemented in other work, can integrate

System Components

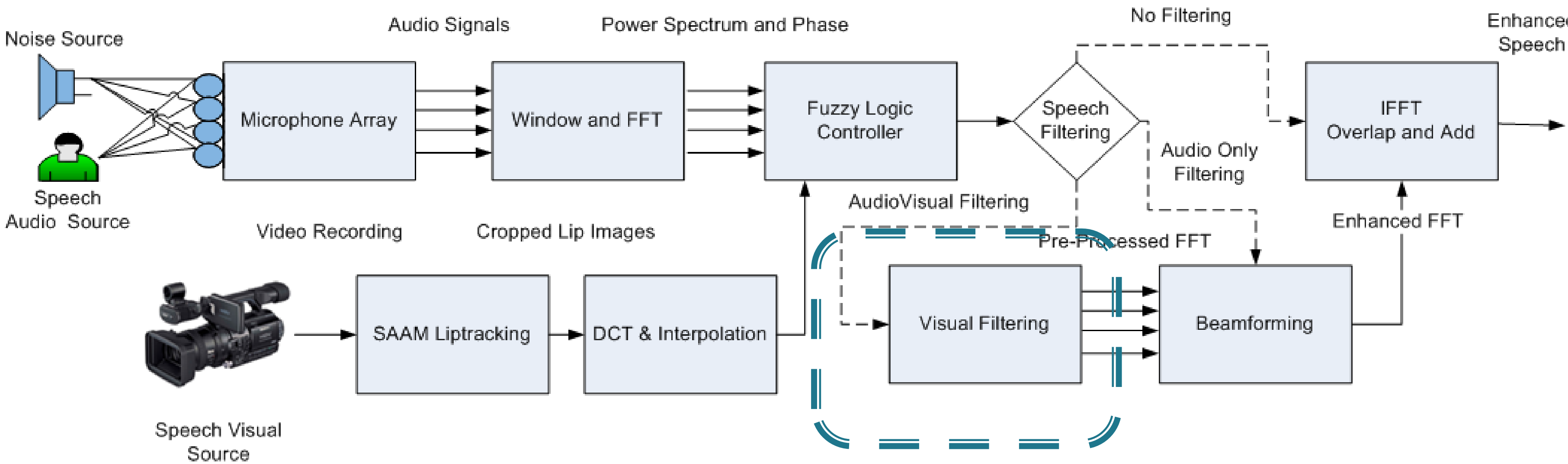


- Audio Feature Extraction
- Visual Feature Extraction
- Visual Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Visual Extraction – Potential

- Alternative Visual Processing options
  - Optical flow
  - DCT of optical flow
  - Shape models
- Additional visual modalities
  - Eye region (eyebrows etc.)
  - Body language
- Temporal element to processing
  - Sequence of frames

System Components

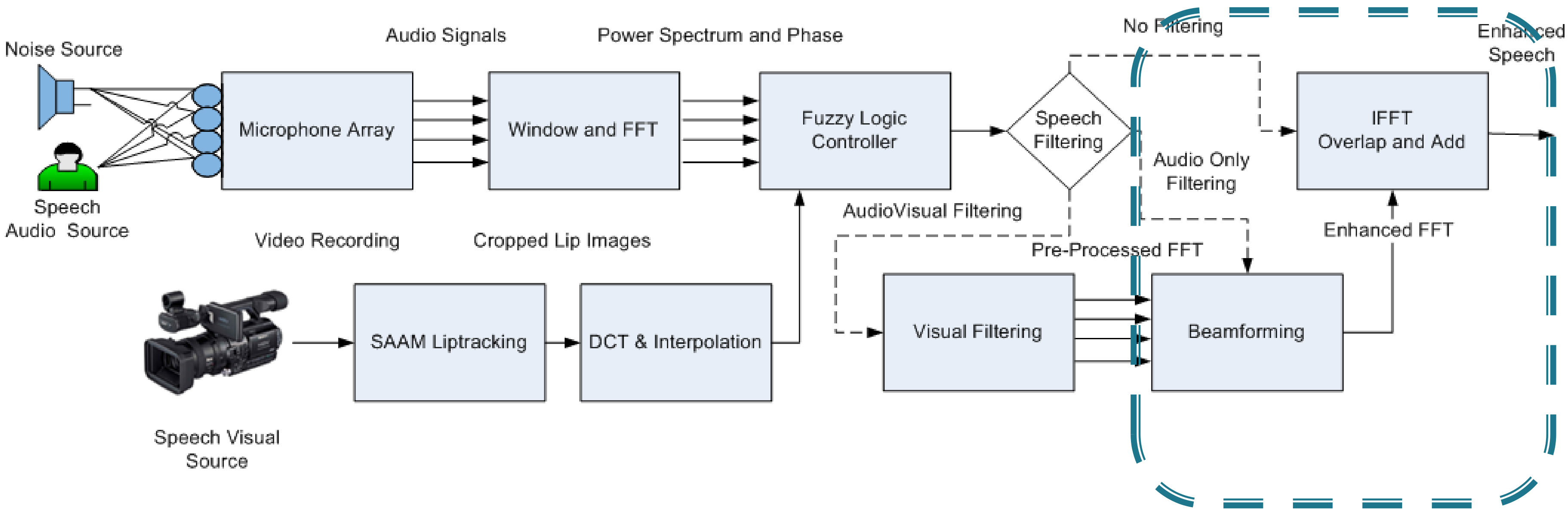


- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Audiovisual Filtering – Potential

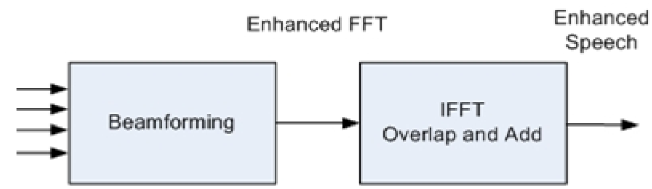
- This model is very basic
  - Is in urgent need of improvement
- Currently single gaussian model
  - Try Phoneme specific model
  - Segment specific model
- Improved machine learning approach
  - Neural network, HMM, Genetic Algorithm?
  - Have audiovisual data, overall framework
  - Need time and expertise...
- Different approach to filtering speech
  - Comparison of approaches

System Components



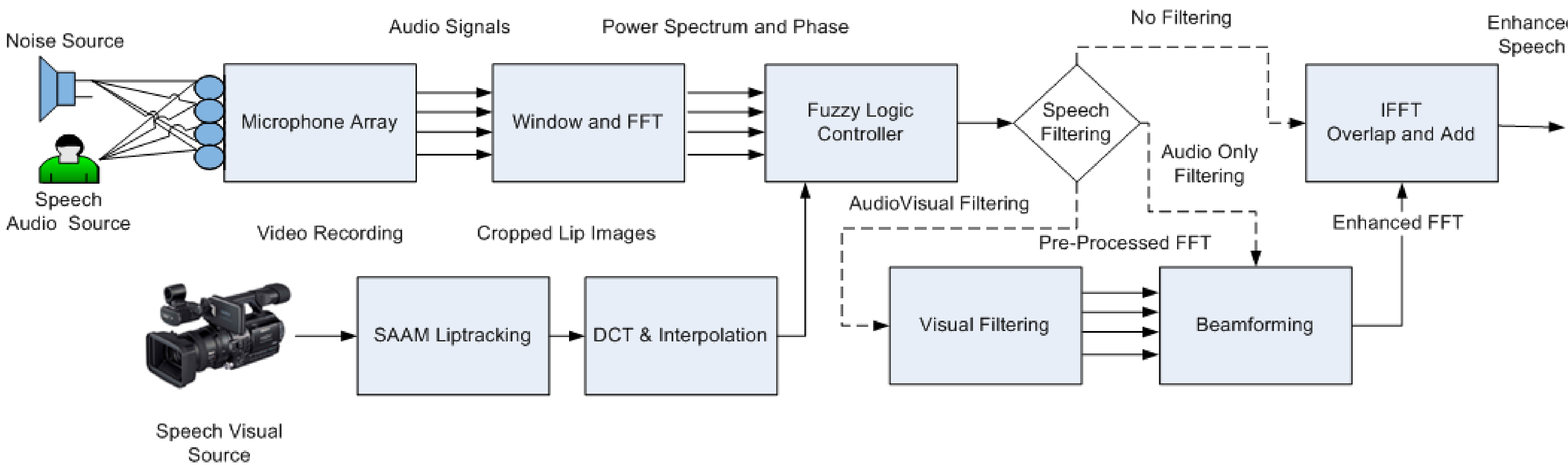
- Audio Feature Extraction
- Visual Feature Extraction
- Visually Derived Filtering
- Audio Beamforming
- Fuzzy Logic Controller

Beamforming – Potential



- Improve to more state of the art model
- Adjust programming to use with real room rather than simulated room
- Use visual information for directional filtering
  - Knowledge of source

Fuzzy Logic Based System



- Fuzzy Logic – Rule based system
  - No human control, all controlled by the system inputs
  - Able to adapt to changing audio and visual environment
  - In real world, noise and environmental factors can change

- Three approaches suited to different environments
  - Two stage filtering
  - Audio Beamforming only
  - No additional processing

Fuzzy Logic – Potential

- Concept functions well
  - Chosen detectors work
  - Results are satisfactory
- Additional detectors and rules could be used
  - Take account of any additional segmentation/modalities
  - Consider more information about environment and sources
- More than just fuzzy logic?
  - Make use of more modalities and outputs in a blackboard system

Hearing and Listening – the future

- More than lipreading
  - Hearing and listening depend on many factors
- Knowledge of language
  - Understanding of accents
- Context of conversation
  - Prediction of next words based on previous content
  - Overall mood
- Body language
  - Emotion, gestures, facial movements, volume

Hearing and Listening – the future

- Cognitively inspired
- Hearsay system
- Blackboard

Audiovisual – Conclusions

- Cognitively inspired filtering aims to design hearing systems that function in a manner inspired by humans
  - Take account of the environment when filtering
  - Combine multiple modalities
  - Switch between processing as appropriate
  - Ignore information when not useful
- A different direction to current hearing aid development
  - Overall grant in preparation
  - Framework has been presented and tested
- Much potential for upgrades of individual components within this framework

- Thank you all for listening
- Questions?

- Contact Details
  - Dr Andrew Abel (Speech filtering research)
    - aka@cs.stir.ac.uk
    - www.cs.stir.ac.uk/~aka