Audiovisual Speech Processing:
Exploiting visual features for better noise reduction

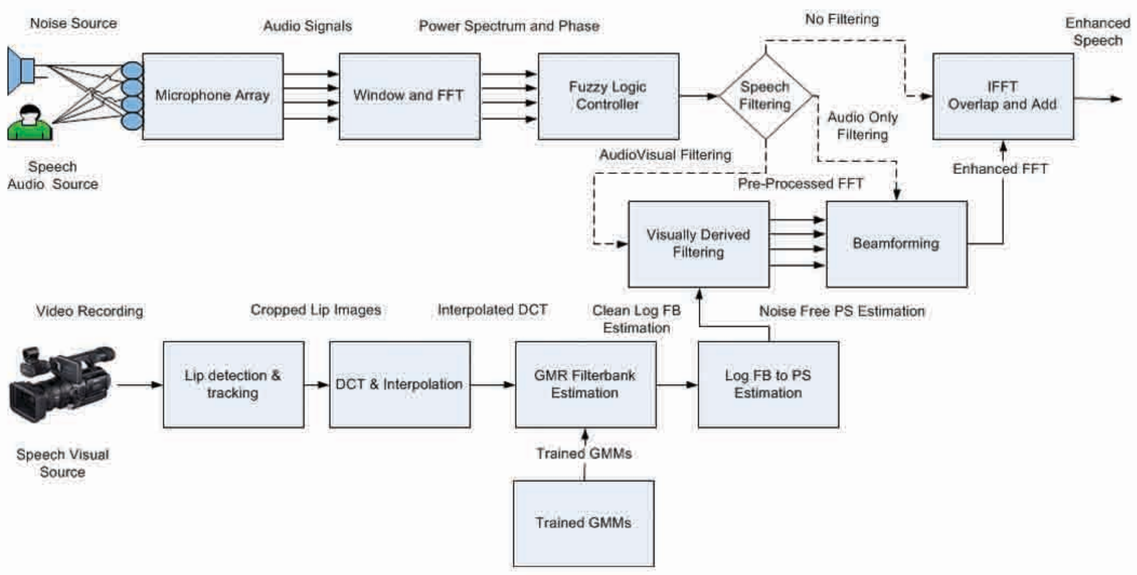
Andrew Abel[1], Ricard Marxer[2], Amir Hussain[1], Roger Watt[1], Jon Barker[2], William Whitmer[3], Peter Derleth[4]
aka@cs.stir.ac.uk, r.marxer@sheffield.ac.uk

(1) Computing Science and Maths, University of Stirling, Scotland, FK9 4LA
(2) Department of Computer Science, University of Sheffield, UK, S1 4DP
(3) MRC/CSO Inst. of Hearing Research - Scottish Section, Glasgow, G31 2ER
(4) Sonova AG, Switzerland, 8712 Staefa

Background

Human speech is multimodal in terms of both production and perception, with a relationship between audio and visual components of speech, and it has been shown that visual information has potential for hearing aid and listening device development. Preliminary systems have been developed to enhance noisy speech, by both filtering and resynthesis methods, and this project proposes to develop these techniques further and make better use of visual information.

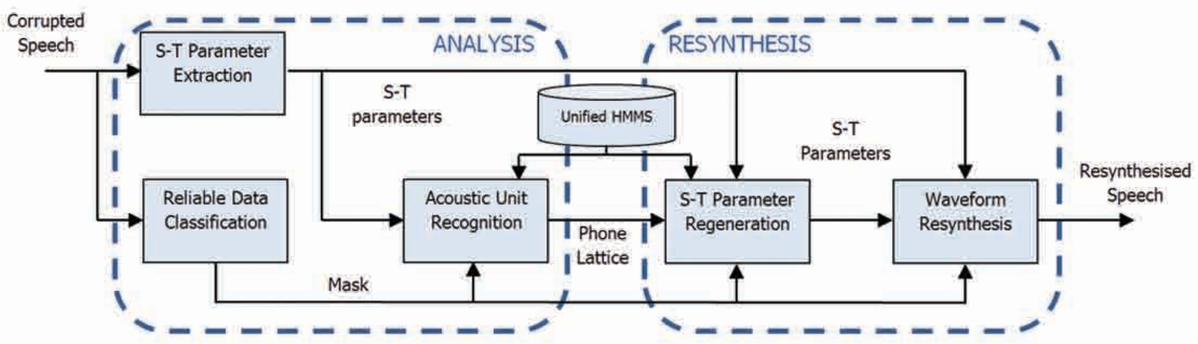
Audiovisual Speech Filtering



- Initially designed noise removal system. See [1] for more info.

An initial speech filtering system to remove noise from a mixed signal using lip images and produce filtered cleaned speech, see [1] for more info. The noisy audio signal (power spectrum and phase) and the equivalent lip region (DCT) features are extracted on a frame-by –frame basis. Visual information is used to produce an estimate of the noise free power spectrum using Gaussian Mixture Regression (GMR) to apply Wiener Filtering. There is also audio only beamforming applied, and a fuzzy logic controller to automatically determine the most suitable processing option (audiovisual or audio-only) on a frame-by-frame basis.

Audiovisual Speech Resynthesis



Missing-data speech analysis

Constrained HMM speech synthesis

- Initially designed speech resynthesis system. See [3] for more info.

Noise subtraction is not suitable in many real environments; noise source may even be speech. We can create an enhancement technique based on missing data speech analysis, followed by speech resynthesis. See [3] for more info. In missing data speech analysis, spectro-temporal information of an audiovisual noisy signal is extracted, enabling the creation of a mask of reliable speech data and generation of an HMM state sequence using trained HMM models of clean speech. This is followed by speech resynthesis, where using the HMM sequence and the mask, noise free data is identified, missing data synthesised, and a clean signal generated.

Selected Publications

[1] Abel, A.K., Hussain, A.: Novel Two-Stage Audiovisual Speech Filtering in Noisy Environments. Cognitive Computation, vol. 6, no. 2, pp 200-215, 2014.

[2] Abel, A.K., Hussain, A., Luo, B.: Cognitively Inspired Speech Processing for Multimodal Hearing Technology. IEEE CICARE 2014 (IEEE Symposium Series on Computational Intelligence), pp. 56-63, 2014.

[3] Barker, J., Shao, X.: Energetic and informational masking effects in an audiovisual speech recognition system. Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 3, pp 446-458, 2009.

Summary

- This new EPRSC project proposes to use lip reading to remove noise from speech, producing clean speech, with the aim of moving towards lip reading hearing aids

- The aim is to combine two previously developed speech processing systems, speech filtering and speech resynthesis to establish the best results.

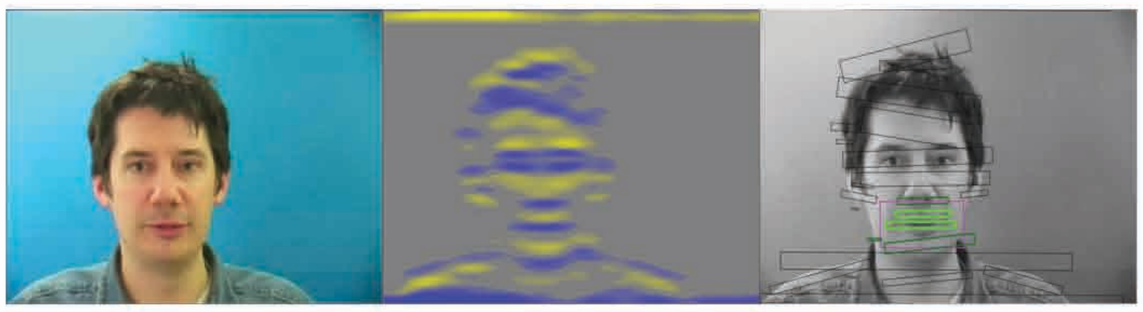
- Investigates psychologically motivated visual features, in this case Gabor features, as an alternative to conventional lip features.

- Current research focuses on using visual lip information to predict audio information using neural networks, a form of mapping problem with promising initial results.

Cognitively Inspired Gabor "Barcodes"

- Psychologically Motivated Image Features

Research shows that humans can extract the majority of speech information from horizontal edges of facial features. Applying horizontal Gabor features can identify edges and theoretically provide the same information at a much lower bandwidth. These horizontal edges can be seen as a form of "barcode".



- Implementation

Gabor filter applied to image (left), identifying edges and amplitudes. Results in thresholded image (centre). This can be further processed to identify patches and "barcodes", particularly around lip region (right). These are new visual features that can be used for speech processing.

Audiovisual Feature Mapping

- The Challenge

Given our visual features as above, can we use lips alone to predict an audio speech vector? In other words, given the shape of the lips, can we work out the sound?

- The Method

Using a large training set (35000 pairs of audio log FB and matching DCT image vectors), and machine learning (MLP neural network) to train audiovisual mapping with different combinations, different visual features, number of frames, network sizes etc.

Audiovisual Mapping Results

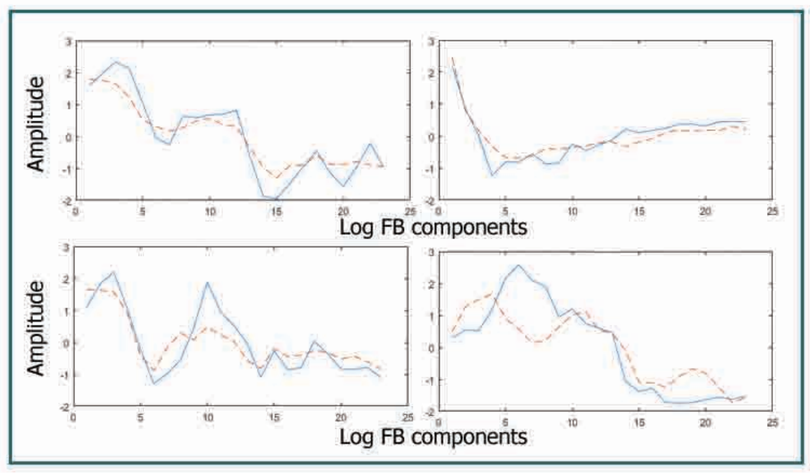
- Aim is to generate an MLP estimate of the log FB results using visual information. Some example results are shown here.

- Each figure represents a single example speech frame, showing audio log FB in (solid blue), and comparing it to equivalent estimation using 14 previous frames of visual information (red dashed line).

- Each figure shows that the visual estimate is a match for the actual estimate in a number of different audio utterance fragments to different extents.

- Implications for improved application to audiovisual processing, such as noise cancelling, or speech detection.

- Ongoing research!



Challenges

- Improved Use of Visual features

Investigate biologically inspired features, to determine the best utilisation for visual information.

- Refined Audiovisual Filtering and Resynthesis approaches

Improve basic filtering system with better use of visual information for speech estimation, and use visual information for masking and model training to improve resynthesis.

- Compatibility and Integration of Approaches

Develop common codebase and integrate different approaches to speech processing, compare both resynthesis and filtering with each other and identify strengths, weaknesses, and maximum compatibility

- Evaluation

Compare approaches to other audio-only approaches using the same dataset, identify scenarios where visual information provides clear benefits.

UNIVERSITY OF STIRLING
The University Of Sheffield.
sonova HEAR THE WORLD
MRC Institute of Hearing Research
CHIEF SCIENTIST OFFICE