Two Stage Audiovisual Speech Filtering: Towards an Intelligent, Fuzzy Logic Based System

Andrew Abel, Amir Hussain, aka@cs.stir.ac.uk, ahu@cs.stir.ac.uk
Cognitive Signal Image and Control Processing Research (COSIPRA) Laboratory (http://cosipra.cs.stir.ac.uk)
Computing Science and Mathematics, University of Stirling, Scotland, FK9 4LA

Background

Human speech is multimodal in terms of both production and perception, meaning that there is a relationship between the audio and visual components of speech. We have developed a multi-modal audiovisual speech filtering system that utilizes both visual and audio speech information (both camera and microphone inputs) to filter speech in noisy environments. We have also implemented a fuzzy logic switching system to determine a suitable processing approach to use for each frame of speech, depending on the availability of visual information and the level of the input audio signal. This system is presented as a potential new direction for hearing aid and listening device development.

Proposed System

• Audio Feature Extraction
  A microphone array receives a noisy audio signal and extracts the Power Spectrum and Phase from each.

• Visual Feature Extraction
  A Viola-Jones face detector and lip tracking are used to automatically identify and track lip features.

• Visually Derived Wiener Filtering
  Visual information is used to produce an estimate of the noise free power spectrum using Gaussian Mixture Regression (GMR). This is then filtered using visually derived Wiener filtering.

• Audio Beamforming
  This audio-only approach takes the pre-processed filtered speech and cancels further noise.

• Fuzzy Logic Controller
  Audio and visual input variables are fuzzified and fuzzy inputs are used to automatically determine the suitable processing option (audiovisual, audio-only, no processing) on a frame-by-frame basis.

Audiovisual Speech Filtering

• Lip Tracking
  A Viola-Jones Region of Interest (ROI) detector and a lip detector is used to automatically identify the lip region and extract the DCT features for processing, as shown below.

• Two Stage Audiovisual Speech Filtering
  Two stage speech filtering operates by first applying visually derived Wiener Filtering, using a visual estimation of the noise free signal. Audio only filtering is then applied to this pre-processed signal. To demonstrate this, clean speech as shown in (a) below, is mixed with babble to produce noisy speech (b). Two-stage filtering removes much of the noise, as shown in (c).

This approach was found to be successful in environments with a very low SNR, but issues with distortion arose when tested in a high SNR environment, or when visual information was unavailable.

Cognitively Inspired Fuzzy Process Switching

• Variable Visual Information
  As shown above, not all lip images can be identified. It is inappropriate to perform visually derived filtering if the lip image is not available. Our online fuzzy logic approach ensures that visual information is only utilised if suitable, as shown below by a noisy example with variable visual information.

• Variable Audio Information
  The level of audio information can vary depending on the type of sound. The figure below shows the same good quality fuzzy input variable used, but with two different input noise sources. If the audio input is consistently noisy, then audiovisual processing is used at all times (see (e) below), but in a noise containing silences and transients, audio-only or no processing is used in a number of frames, as shown in (f) to produce filtered output speech.

Ongoing / Future Work

• Improved Visually Derived Filtering
  Make use of more sophisticated visually derived filtering, and improve audiovisual model training for better generalisation. Make use of improved visual features, possibly by combining several frames and exploiting audiovisual asynchrony.

• Refined Fuzzy Logic System
  Further use of fuzzy input variables such as Voice Activity Detectors to add sophistication to the system. Also, explore with tweaking and addition of rules to add flexibility and take account of more potential scenarios.

• Hardware Implementation
  Implement fuzzy logic speech enhancement system on dedicated hardware to create functional real-time prototype, potentially using Field-Programmable Gate Arrays (FPGAs).

Selected Publications