Multi-modal Speech Processing Methods: An Overview and Future Research Directions Using a MATLAB Based Audio-Visual Toolbox

Andrew Abel and Amir Hussain

Dept. of Computing Science, University of Stirling, Scotland, UK
aka@cs.stir.ac.uk, ahu@cs.stir.ac.uk

Abstract. This paper presents an overview of the main multi-modal speech enhancement methods reported to date. In particular, a new MATLAB based Toolbox developed by Barbosa et al (2007) for processing audio-visual data is reviewed and its performance potential evaluated. It is shown that the tool does not represent a complete and comprehensive speech processing solution, but rather serves as a standardised, yet versatile base to build upon with further research. To demonstrate this versatility, preliminary examples that make use of these computational procedures with an audiovisual corpus are demonstrated. Finally, some future research directions in the area of multi-modal speech processing are outlined, including future research that the authors aim to carry out with the aid of this newly developed audio-visual MATLAB toolbox, including toolbox customisation, and processing noisy speech in real world environments.

1 Introduction

Speech processing is a diverse field with a range of disciplines, such as speech recognition and speech enhancement. The goal of speech recognition is to have a machine recognise and respond to the speech of a user, whereas speech enhancement is focused on cleaning up a noisy speech signal, with developments such as audio blind source separation (BSS), which aims to solve the "Cocktail Party Problem" [1] by separating a mixture of speech signals. Recently, more research has been focused on multimodal speech processing. It is established that human speech is multimodal [2] [3], with complementarity and coherence between audio and visual signals. For example, it has been found that visual data can help to discriminate between similar sounds such as "f" and "s" [4] [5].

This paper is organized as follows: Section 2 presents an overview of several state of the art multimodal speech enhancement techniques, including kernel based separation, speech fragment decoding and Weiner filtering systems. A recently developed MATLAB Audiovisual Toolbox [6] will be reviewed in section three, which aims to ease and standardise the process of measuring, organising, and analysing speech. Section four will outline some preliminary results and future research directions.

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 121–129, 2009.

2 State of the Art

2.1 Kernel Based Separation System

Rivet et al [7] have developed a multimodal BSS speech enhancement system, with the aim of removing a speech source from a noisy background mixture. This expands upon previous related work, and is an extension of audio only based BSS solutions. Rivet et al propose to use visual information to complement the audio when BSS is unable to be carried out solely with the audio signal. To provide the convolutive speech mixture, two French speakers are combined. The target speaker utters a mixture of consonants and vowels, and the secondary speaker provides the background noise with well balanced sentences. The two speakers are combined to provide a more realistic convolutive speech mixture.

The speech is extracted by splitting the noisy audio signal into a number of kernels. At a given point in time, the visual vector is made up of the height and width of the mouth opening, and the audio signal is converted to the frequency domain with a Fourier Transform. These parameters are then used to find the permutation of the information that produces the best audio and visual data match, and subsequently the best demixing matrix to be applied to the input signal.

2.2 Speech Fragment Decoding System

Barker and Shao [4] have developed a multimodal speech fragment decoding technique. This is primarily a speech recognition system, but it also covers speech enhancement, and deals with both additive and convolutive mixtures. This system expands on audio only fragment-based BSS. The audio signal is broken up into fragments and grouped by frequency. Fragments dominated by the target speaker are selected, while fragments dominated by background noise are discarded. To provide more information than available in audio only approaches, visual information is used to aid with fragment selection.

The audio signal is passed through a filterbank to split into frequencies, and the visual signal is extracted by carrying out a Discrete Cosine Transform (DCT). Hidden Markov Models (HMMs) are trained to construct a limited word vocabulary, based on a corpus containing a number of small sentences. When attempting to recognise speech, this system works under the assumption that the speaker exists within the trained HMMs, and compares the fragments with a relevant HMM set. The recognition and enhancement stages are intertwined, due to the use of the same HMMs for both.

2.3 Wiener Filtering System

An audiovisual Wiener filtering system has been created by Almajai et al. [8], which attempts to provide an estimation of the noiseless speech signal in order to perform speech enhancement. This research makes use of sentences spoken by a single male speaker, which is then combined with white noise to create a noisy

speech mixture. The input audio signal makes use of filterbank vectors, and the visual information is extracted using Active Appearance Models (AAMs) [16]. Two different estimation methods are discussed, global, and phoneme dependent.

The global technique divides the input into frequency domain clusters, and visual information is used to select a single Gaussian Mixture Model (GMM) to apply to the speech mixture. The phoneme dependent method selects a different GMM for each phoneme uttered by the speaker. These are identified with trained HMMs, and the resulting GMM sequence is then applied to the utterance to create an estimation of the noiseless signal. This is compared to the input, and the resulting cleaned up speech is output.

This technique takes a nuanced approach to the content of speech when using visual data. It currently only deals with additive speech mixtures, rather than convolutive mixtures, but preliminary results described by Almajai et al [8] have shown that a lot of the added noise has been removed, with informal listening tests also reporting an improvement in quality.

2.4 Overview of other Multi-modal Speech Processing Approaches

There are many other multi-modal speech approaches that have not been discussed here, including prominent works by Scanlon and Reilly [9], and segment-based recognition research by Hazen el al [10]. Potamianos et al [11] summarise a wide range of speech recognition techniques, with particular concentration on the various visual feature extraction techniques. Another briefer summary is provided by Goecke [12], who review some audio-visual corpora with a focus on multimodal biometric authentication.

3 Audio-Visual MATLAB Toolbox

3.1 Overview

The Audiovisual Toolbox for MATLAB is a collection of functions developed by Barbosa et al [6] for measuring, organising, and analysing multimodal speech data. It makes use of the existing Signal Processing and Image Processing Toolboxes, and for optical flow analysis, OpenCv is required. Experiments are prepared by converting the data into the appropriate format, and then setting the parameters in an experiment file, inputting information such as the file location and appropriate time windows.

When the data is processed, the raw audio and visual data is converted into more useful values with the aid of a number of functions. These are returned as a matrix, grouping all information about a particular take in one location for ease of analysis. The audio signal is converted into Root Mean Square (RMS), Line Spectral Pairs (LSP), and fundamental frequency (F0) values. If optical flow analysis is carried out, a matrix of pixel movement is returned, along with a greyscale visualisation of movement. Tracked marker data is returned by tracking the movement of pre-defined markers, and performing Principal Component Analysis (PCA). There are also a number of data analysis functions available.

These include functions to resample the data, analyse correlation, window, and enframe signals.

3.2 Limitations

One useful refinement of this toolbox would be to increase the range of visual feature extraction techniques supported. The toolbox can only presently carry out PCA on tracked marker data, and also supports optical flow, has relatively little support for analysis of clean face corpora. It would be useful to update it in order to support some of the more sophisticated clean face techniques such as AAMs or DCT [13]. A suggested outline of additional toolbox functions is shown in Fig.1, showing the current visual techniques supported (3D and 2D Marker Trajectories, and Optical Flow), with proposed AAM modifications.

Active Appearance models were originally developed by Cootes et al. [16], and create models of visual features by making use of shape and texture information. This makes them very suitable for tasks such as the creation of face models, and the generation of appearance parameters for speech processing.

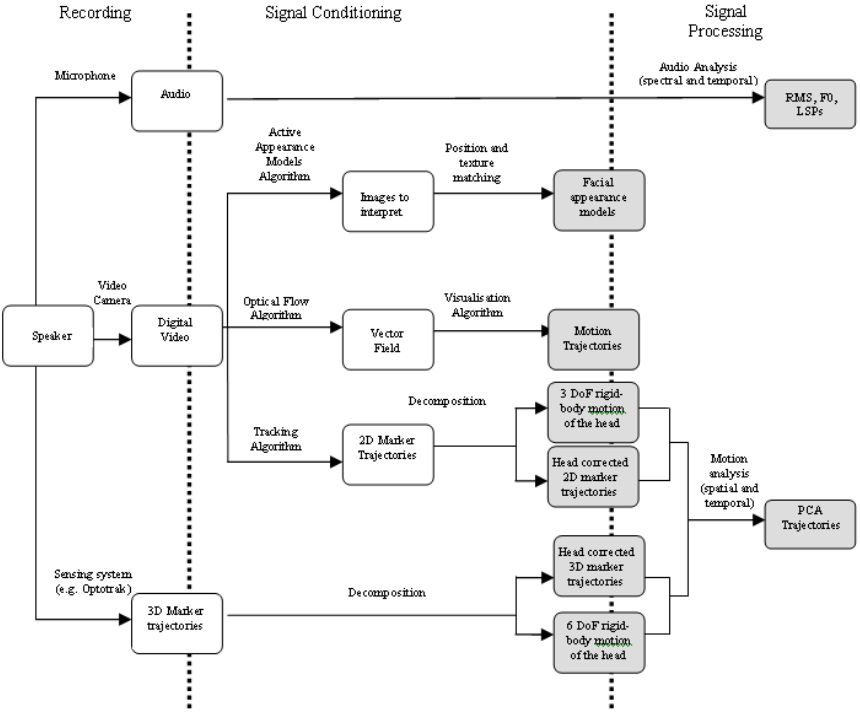

Fig. 1. Modified diagram showing recording, conditioning and processing of speech data, with suggested AAM enhancements. All shaded boxes represent signals that are stored and available to the analysis functions of the toolbox.

Fig. 2. Example of an AAM trained using selected images from the VidTIMIT corpus. Top row represents shape model only, bottom row shows reconstruction using both texture and shape information.

Existing AAM software can be adapted in a similar manner to the way optical flow is currently utilised. This means that the image modelling and parameter generation are carried out externally, but accessed via MATLAB. The parameters can be read in and converted to the same format as existing audio and visual features, making them compatible with the data structures of the toolbox. This means that the AAM can take advantage of the functions that the toolbox offers such as resampling and correlation analysis. Figure 2 shows an example of a shape and appearance model, trained using images from the VidTIMIT corpus.

There are some advantages to extending the toolbox in this manner. Firstly, although the toolbox is non corpus specific, it is limited to an extent in that if no tracking markers are present in a corpus, only optical flow can currently be used with it. As many corpora do not contain these markers, this reduces the usability of the toolbox. Adding support for AAMs and DCT means that precise visual features can be analysed with more detail than using optical flow, and that the toolbox can be used with a greater range of audiovisual corpora.

Additionally, as the toolbox is still under development, documentation is currently relatively sparse. Functions are inconsistently documented, with no comprehensive guides or tutorials available at this stage in development.

3.3 Benefits

The Audiovisual Toolbox provides a standardised and verifiable set of functions. This saves time as speech analysis functions no longer need to be custom written from scratch, and they provide a degree of automation to ease audiovisual speech analysis. It is also non corpus specific; it can work with a range of different corpora, and can provide a useful starting point to inexperienced researchers.

4 Preliminary Results and Future Directions

4.1 Preliminary Results

Although detailed experiments are at a very early stage, initial experiments have proven the compatibility and feasibility of VidTIMIT with the toolbox.

Fig.3 shows an example of the instantaneous correlation coefficient between the audio signal and lip movement. In this example a single sentence of a speaker from the VidTIMIT corpus is used to demonstrate correlation. The audio and visual signals are processed by functions available in the toolbox to produce the RMS amplitude of that speakers utterance and the pixel movement (attained by performing optical flow analysis, shown in Fig.4) of the image sequence. These

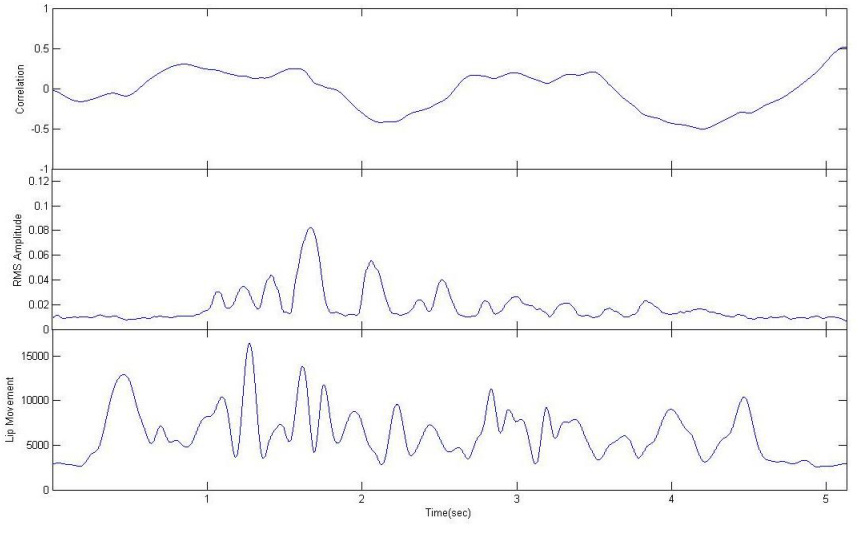

Fig. 3. Instantaneous correlation coefficient between lip movement and RMS amplitude of speech. (Lower plot above shows Lip movement, middle-plot shows RMS amplitude, and the top plot shows the Correlation)

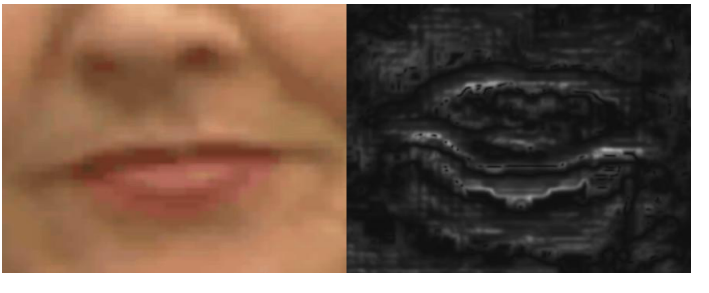

Fig. 4. Example of a cropped VidTIMIT lip image, and optical flow results

are resampled to allow for frame by frame instantaneous correlation analysis, as shown in Fig.3.

It can be seen that the correlation coefficient is usually significantly above or below zero, suggesting that parts of the utterance demonstrate a strong degree of audiovisual correlation. This shows the suitability of VidTIMIT for further audiovisual research. In addition, these initial results also demonstrate the flexibility of the Audiovisual Toolbox. The authors are not aware of any existing examples of this toolbox being used with VidTIMIT, and this work shows the compatibility of the toolbox with untested corpora.

4.2 Immediate Future Developments

For future development, the Audiovisual Toolbox will be used for in depth analysis of the VidTIMIT Corpus [14]. A number of experiments will be carried out, with the goal of deriving a suitable division of speech for HMM or Artificial Neural Network (ANN) allocation. Firstly, the effect of different speakers speaking the same utterance will be carried out, in order to identify speaker independent common visual information. The most suitable division of speech will also be investigated, with the merits considered of dividing speech by single phonemes, multiple phoneme groupings, or whole words.

A comparison of different corpora may be carried out, with VidTIMIT compared to the AviCar [15] corpus. This is to analyse the difference in speech between the quiet VidTIMIT environment, and the noisier automobile background noise provided by AviCar. Speakers talk differently in different environments, and these experiments will attempt to see if there is a significant difference in audiovisual correlation.

4.3 Long Term Future Developments

The completed speech analysis experiments will produce a set of speech utterance divisions that will serve as the basis for construction of an initial wiener filtering system, initially closely related to the one outlined in section 2.3. Once an initial system has been built and tested, it can be refined and improved, for example, by being extended to cover both additive and convolutive mixtures.

The existing audiovisual Wiener filtering system makes use of trained Hidden Markov Models to identify the suitable GMM to use. There is scope for experimenting with other methods of model selection such as Artificial Neural Networks or Genetic Algorithms, and the merits of alternative feature extraction methods could be considered. Finally, this is intended to be a people centred system. Therefore, detailed intelligibility testing on human volunteers will be carried out.

5 Summary

This paper has presented an overview of three state of the art multimodal speech enhancement methods, along with a review of a new MATLAB based Toolbox

developed by Barbosa et al (2007) for processing audio-visual data. It is shown that this is not a fully featured speech processing solution, but rather a versatile base for research. Preliminary results have been presented which demonstrate the utility of the Toolbox, along with a summary of immediate and long-term future research developments.

Acknowledgements

This work was funded with the aid of an euCognition travel grant and a research studentship from the University of Stirling. Thanks are also due to Prof. Leslie Smith (University of Stirling) and Dr. Mohamed Chetouani (University Pierre Marie Curie, France) for their invaluable help and input.

References

1. Haykin, S., Chen, Z.: The Cocktail Party Problem. Neural Computation 17(9), 1875–1902 (2005)
2. Sumby, W.H., Pollack, I.: Visual Contribution to Speech Intelligibility in Noise. J. Acc. Soc. America 26(2), 212–215 (1954)
3. Schwartz, J.L., Berthommier, F., Savariaux, C.: Audio-visual scene analysis: evidence for a "very-early" integration process in audio-visual speech perception. In: ICSLP 2002, pp. 1937–1940 (2002)
4. Barker, J., Shao, X.: Audio-Visual Speech Fragment Decoding. In: AVSP 2007, paper L5-2 (accepted, 2007)
5. Almajai, I., Milner, B.: Maximising Audio-Visual Speech Correlation. In: AVSP 2007, paper P16 (accepted, 2007)
6. Barbosa, A.V., Yehia, H.C., Vatikiotis-Bateson, E.: MATLAB toolbox for audio-visual speech processing. In: AVSP 2007, paper P38 (accepted, 2007)
7. Rivet, B., Girin, L., Jutten, C.: Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures. IEEE Trans. on Audio, Speech, and Lang. Processing 15(1), 96–108 (2007)
8. Almajai, I., Milner, B., Darch, J., Vaseghi, S.: Visually-Derived Wiener Filters for Speech Enhancement. In: ICASSP 2007, vol. 4, p. IV-585–IV-588 (2007)
9. Scanlon, P., Reilly, R.: Feature analysis for automatic speechreading. Mult. Sig. Processing. In: 2001 IEEE Fourth Workshop on, pp. 625–630 (2001)
10. Hazen, J.T., Saenko, K., La, C.H., Glass, J.R.: A Segment Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments. In: ICMI 2004: Proceedings of the 6th international conference on Multimodal interfaces, pp. 235–242 (2004)
11. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent Advances in the Automatic Recognition of Audiovisual Speech. Proceedings - IEEE, part. 9, 91, 1306–1326 (2003)
12. Goecke, R.: Current Trends In Joint Audio-Video Signal Processing: A Review. In: Proceedings of the Eighth Int. Symposium on Signal Processing and Its Applications, pp. 70–73 (2005)
13. Potamianos, G., Neti, C., Deligne, S.: Joint Audio-Visual Speech Processing for Recognition and Enhancement. In: AVSP 2003, pp. 95–104 (2003)

14. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag (2008)
15. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T.: AVICAR: audio-visual speech corpus in a car environment. In: Interspeech 2004, pp. 2489–2492 (2004)
16. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. IEEE Trans. On Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)