Final Report: GR/T04823/01
Audio-visual speech recognition in the presence of non-stationary noise

Jon Barker and Xu Shao
University of Sheffield, Department of Computer Science, Sheffield S1 4DP

1 Summary

This project has concerned the development of novel techniques for exploiting visual speech information (e.g. lip and face movements) in the design of automatic speech recognition systems. The approaches explored are motivated by the desire for reliable speech recognition in the presence of highly non-stationary noise sources, such as background speech. The basis of the project is a recent approach to robust automatic speech recognition that couples the problems of source separation and speech recognition by 'piecing together' spectro-temporal *fragments* of speech recovered from regions of a time-frequency representation in which the signal *locally* dominates the noise. The project has extended this approach into the audio-visual domain. The audio-visual system exploits the correlation that exists between audio and visual aspects of speech to resolve ambiguities in the acoustic fragment labelling that occur when attempting to recognise speech in the presence of noises with speech-like characteristics.

As well as providing a new audio-visual speech recognition framework, this research has built on and integrated work from earlier projects dealing with the development of Auditory Scene Analysis algorithms (EPSRC grant GR/H53174/01), the recognition of occluded speech (EPSRC grant GR/K18962/01, 'RESPITE' - EC ESPRIT LTR project) and the development of the speech fragment decoding architecture (EPSRC grant GR/R47400/01). The principal contributions of the project have been,

- The collection of an audio-visual speech corpus suitable for both measuring speech intelligibility and for testing robust audio-visual automatic speech recognition (AV-ASR) systems.
- An improved set of algorithms for analysing acoustic mixtures and generating spectro-temporal sound source fragments.
- A complete software implementation of a fully functional audio-visual fragment-based recognition system.
- A demonstration of the advantages of the fragment-based AV-ASR technique in comparison with conventional 'multistream' approaches when applied to a simultaneous-speaker recognition task.
- The development and evaluation of versions of the AV fragment-based techniques that accommodate the *asynchrony* that can occur between corresponding phoneme and viseme boundaries.

2 Background

Humans are incredibly adept at understanding speech in noisy conditions (e.g. Deshmukh et al., 1996). In face-to-face conversation, the robustness of our speech processing depends partly on the fact that speech is not just an acoustic signal, but also has an information-rich visual component. As background noise levels increase listeners subconsciously make increasing use of visual speech information – i.e. they attend more carefully to lip, jaw and subtle facial movements. As long ago as 1954, studies demonstrated that this visual component of the speech signal can increase intelligibility to the same extent as increasing the SNR by as much as 15 dB (Sumby and Pollack, 1954). Since then it has become increasingly clear that the role of vision is central to the perception of speech. For a review see Summerfield (1992).

In recent years there has been much research activity directed towards exploiting the visual component of the speech signal in robust ASR systems. This research has mainly focused on two question: i) How best to parameterise visual features? ii) How best to integrate the audio and visual feature data? Many different systems have been proposed using various combinations of visual feature and integration mechanism (for a review see Potamianos et al., 2003). Most existing AV-ASR systems derive advantage from the visual stream by exploiting its phonetic content. However, an additional role for the visual stream can be imagined: the visual stream may aid recognition by helping the ear to *separate* the speech from the noise background. Recent audio-visual perception studies provide evidence that visual features are employed in this manner by humans. Schwartz et al. (2004) demonstrate that visual speech can improve intelligibility even in situations where it carries no useful phonetic information. In simultaneous speech conditions, Helfer and Freyman (2005) and Wightman et al. (2006) have shown that visual cues can provide benefit to intelligibility by helping the auditory system to selectively attend to one speaker.

This project has aimed to build AV-ASR systems that model the dual role of the visual speech information that these recent perceptual studies suggest – that of supplementing the energetically masked phonetic content of the audio signal, and that of providing information that can aid the separation of the speech and the background. As a starting point the work has considered the *auditory scene analysis* (ASA) account of auditory perception that describes how the acoustic mixture arriving at the ear can be effectively 'unmixed' so that the individual sound sources may be perceived separately (Bregman, 1990; Darwin and Carlyon, 1995). This account can be broadly characterised by two processing stages: primitive grouping processes identify spectro-temporal fragments whose energy is dominated by an individual sound source, then expectation-driven processes group these fragments to form the best match to models of individual sounds sources, e.g. speech.

A robust speech recognition technique known as *speech fragment decoding* (SFD) has been previously proposed that models both the fragment generation and grouping stages of ASA. Previously, this model has only been evaluated in audio-only speech recognition tasks. The project has aimed to demonstrate that this model also presents a valid framework for integrating audio and visual information to produce an AV-ASR system that exploits visual information in the ways discussed above.

The research has been organised into five work packages: (i) speech data collection and preparation – collection and annotation of an audio-visual speech corpus plus generation of simultaneous speech data for evaluation; (2) primitive auditory grouping – further development of techniques for sound source fragment generation and application of these techniques to the simultaneous speech task; (3) synchronous AV speech fragment decoding – extension of the speech fragment decoding technique to handle audio-visual data; (4) asynchronous AV speech fragment decoding – extension of (3) to allow the modelling of AV asynchrony; (5) evaluation – development of baseline results using conventional AV-ASR techniques, and comparison with the performance of AV-SFD. Progress in each of these areas is summarised in the section to follow.

The project has had the involvement of two personnel (one academic - Barker; one full-time RA, Shao).

3 Key Advances and Supporting Methodology

3.1 WP1: Speech data collection and preparation

Due to the lack of an existing audio-visual speech corpus suitable for the planned research, the project has been committed to the collection of new data. The proposal described the collection of a total of 10 hours of digit utterances involving 20 speakers. It was planned to use reflective markers to capture lip dynamics. In response to reviewers' comments the scope of the planned corpus has been increased. With the additional support of a University of Sheffield Research Facilitation grant, a larger corpus totalling 34 speakers and 20 hours of speech has been collected. The digit task has been replaced by a more phonetically diverse alpha-digit recognition task, in which alpha-digit 'grid references' are embedded in simple command sentences. Also in response to reviewers' comments the plan to use reflective markers was replaced. Instead, *semi-automatic* lip-tracking techniques have been employed to ensure that uniformly reliable and high-quality lip features can be extracted (Shao and Barker, 2007).

The corpus of AV data, know as the 'Grid corpus', has been documented in a letter to the *Journal of the Acoustical Society of America* (Cooke et al., 2006). It is designed, not just for audio-visual speech recognition experiments, but also for the study of human speech intelligibility. It is therefore ideal for direct human speech recognition (HSR) versus ASR comparisons. It has already lead to two significant publications in this area (Barker and Cooke, 2007; Cooke et al., conditionally accepted). It has also been employed for a recent audio-only simultaneous speech recognition competition known as the Pascal Speech Separation Challenge (PSSC) [1] in which nearly a dozen international competitors have participated. A special issue of the journal *Speech Communication* is presently in preparation as a result of this challenge. The PSSC has meant that although the Grid corpus is new, many state-of-the-art audio-only ASR external baselines are already available. The audio-visual research in this project, has been evaluated using an audio-visual simultaneous speech recognition task that employs the same set of speech utterance mixtures as the PSSC so that our (audio-only) results can be viewed in the context of these baselines.

The audio and visual data from the Grid corpus has been made freely available for download from the web.[2] High quality versions of the video data are being made available on request. The data will also be distributed on DVD to interested participants at this years Audio-Visual Speech Processing workshop, where the results of the current project are due to be presented. Currently only the audio and visual signal are available, however, a future release is planned which will also include the extracted visual *features*, hence lowering the overhead for institutes that wish to participate in AV-ASR research.

[1] see http://www.dcs.shef.ac.uk/∼martin/SpeechSeparationChallenge.htm

[2] http://www.dcs.shef.ac.uk/spandh/gridcorpus/

3.2 WP2: Primitive auditory grouping (sound source fragment generation)

The fragment decoding technique is dependent on the quality of the spectro-temporal fragments that are generated by the front-end signal processing. It is important that the fragments should each be dominated by the energy of a single source. Also, in order for the grouping process to constrain the search space of possible foreground/background segmentations considered by the second stage of the system, it is desirable that the fragments have a large area. WP2 has concentrated on improving the fragment generation algorithms to meet these goals, and on applying the algorithms to the simultaneous speaker data generated in WP1.

Multi-pitch detection and tracking techniques can be used as the basis for generating reliable spectro-temporal fragments: filterbank channels can be grouped across frequency if they are excited by a common fundamental frequency; grouped channels can be integrated through time if their corresponding pitches form valid pitch trajectories. The project has made important contributions to both these stages.

Ma et al. (in press) have developed new robust algorithms for estimating the pitches of simultaneous sources exploiting an autocorrelogram representation of the mixed signal. In this technique the signal is passed through a gammatone filterbank with filters spaced to mimic the non-linear frequency resolution of the ear. At a given time instant, an autocorrelation is performed on the signals in each channel. The resulting two-dimensional representation (with axes of frequency and autocorrelation delay) contains competing tree-like structures whose stems are located on the delays corresponding to the pitches of the sources present. Ma et al. (in press) show how the stem positions can be robustly estimated by using a template that matches the local details of the tree-like structure.

Coy and Barker (2007) have developed a novel multi-pitch tracking algorithm that uses an HMM to model the change of voicing-state of a speech source, and a simple model of pitch dynamics in voiced segments. Independent HMMs are used to model each speech source, and a separate noise-process is used to model the spurious pitch estimates generated by the pitch detection algorithm. A Viterbi decoding is then able to form the most likely description of the data in terms of a number of potentially overlapping pitch track segments.

Additionally, the speech fragment decoding architecture has been generalised to accept a 'confidence mask' input which encodes the degree of certainty with which each spectral-temporal point has been allocated to a fragment (Coy and Barker, 2007). This mask effectively lends a probabilistic interpretation to the otherwise categorical foreground/background segmentations that the decoder generates (the categorical nature of the fragment labelling was a weakness noted by the project proposal reviewers). Softening the fragment interpretation is similar to the use of 'soft masks' in missing data systems (Barker et al., 2000; Coy and Barker, 2005b). Our evaluations have shown that it increases recognition performance to a similar degree (Ma et al., in press).

3.3 WP3+WP4: Synchronous and asynchronous audio-visual speech fragment decoding

WP3 and WP4 have concerned the development of the synchronous and asynchronous implementations of the AV speech fragment decoder respectively. These systems can be built without need for extension of the underlying speech fragment decoder machinery. For the synchronous version it is only required that the HMM's acoustic model is replaced with an AV model trained on AV features vectors that are formed from simple concatenation of the A and V feature vector components. Each HMM state is represented using a GMM with a number of diagonal-covariance components. These can be constructed by treating A and V as independent given the state – the implementation first trains a multistream AV HMM on clean speech and then converts this into a standard HMM by 'squaring out' the GMMs contained in the A and V streams. Alternatively, the AV HMM can be trained directly on the concatenated AV data. In either case, during recognition, the AV-SFD treats the visual features like one large fragment that is known to belong to the target speech source, and the decoder runs exactly the same algorithm and probability calculations that are employed for the audio-only case.

The asynchronous AV-SFD is based on the product-HMM implementation introduced by Neti et al. (2000). The goal here has been to validate that the performance advantage afforded by asynchrony in multistream approaches carries through to the SFD system. A state-synchronous multistream AV HMM is first trained. Then it is assumed that within each word-model the A and V streams can be desynchronised by a maximum of N states (usually 1 or 2). Composite AV states are constructed by squaring out the A and V GMM components of states that can correspond when allowing for the possibility of N states worth of asynchrony. An HMM is constructed by adding the transitions between composite states obtained by combining the transitions that would occur in independent A and V HMMs (e.g. for non-skip HMMs it is possible to either stay in the same state for both A and V, advance only in A, advance only in V, or advance in both A and V). The reconstructed AV HMM is then compatible with the standard fragment decoding software.

Barker and Shao (2007) presents implementation details and evaluation of the synchronous decoder; the asynchronous

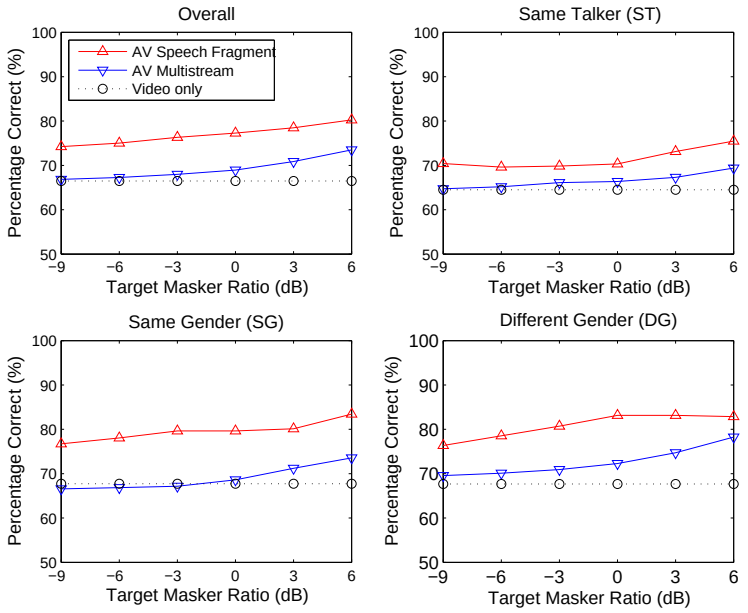

Figure 1: *Comparison of AV-SFD and the optimal multistream system when evaluated on the simultaneous speaker recognition task (see text).*

version can provide a small additional relative performance improvement ($\sim 5\%$ at most SNRs). A paper providing full details of both the synchronous and asychronous decoders is currently in preparation (Barker and Shao, in preparation).

3.4 WP5: Evaluation

Much of the work in WP5 has consisted of developing credible baseline results against which to compare the results of the systems developed in WP3 and WP4. State-of-the-art AV-ASR systems are typically constructed around a multistream architecture in which audio and visual stream weights are optimised to minimise the effects of acoustic noise. In order for these systems to work effectively in non-stationary noise conditions, the stream weights need to adapt over time in accordance with a measure of the local SNR. The project has developed multistream baseline systems using both fixed stream weights and adaptive stream weights. It has also compared systems that have set the stream weight using either a measure of the true SNR (obtained using knowledge of the clean speech prior to mixing with noise) or an estimated SNR derived from the noisy AV data. Novel techniques have been developed to estimate SNR using audio and visual HMM state-likelihood data. It has been demonstrated that visual information aids SNR estimation in situations where the noise background is confusable with the speech signal (e.g. simultaneous speech). The development of these baseline systems is reported in Shao and Barker (2006) and Shao and Barker (2007).

Figure 1 shows a comparison of the performance of the synchronous AV-SFD technique and the optimal multisource AV-ASR baseline system (time varying stream weight optimised by assuming knowledge of local SNR). The evaluation employs the AV version of the Pascal Speech Separation Challenge in which a target and masker talker are mixed at a variety of SNRs. Note how the fragment decoder is able to exploit the acoustic signal at very low SNRs. Details of these experiments, which also compare audio-only and audio-visual fragment decoding systems, appear in (Barker and Shao, 2007). An extended version of this paper, including results of both the synchronous and asynchronous systems is in preparation for submission to the journal *Speech Communication* (Barker and Shao, in preparation).

References

J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

*J. P. Barker. Tracking facial markers with an adaptive marker collocation model. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP-2005)*, pages 665–669, Philadelphia, PA, 2005.

*J. P. Barker and M. Cooke. Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417, 2007.

J. P. Barker, M. P. Cooke, and D. P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005.

* J. P. Barker, A. Coy, N. Ma, and M. Cooke. Recent advances in speech fragment decoding techniques. In *Proceedings of Interspeech 2006*, pages 85–88, Pittsburgh, PA, 2006.

* J. P. Barker and X. Shao. Audio-visual speech fragment decoding. In *Proceedings of AVSP 2007*, Hilvarenbeek, The Netherlands, 2007.

* J. P. Barker and X. Shao. Audio visual speech fragment decoding: synchronous versus asynchronous models. *Journal of the Acoustical Society of America*, in preparation.

A. S. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, MA, 1990.

* M. Cooke, J. P. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

* M. Cooke, M. L. Garcia Lecumberri, and J. P. Barker. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America*, conditionally accepted.

A. Coy and J. P. Barker. Soft harmonic masks for recognising speech in the presence of a competing speaker. In *Proceedings of Interspeech 2005*, pages 2641–2644, Lisbon, Portugal, 2005b.

* A. Coy and J. P. Barker. A multipitch tracker for monaural speech segmentation. In *Proceedings of Interspeech 2006*, pages 1678–1681, Pittsburgh, PA, 2006.

* A. Coy and J. P. Barker. An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Communication*, 49(7):384–401, 2007.

C. J. Darwin and R. P. Carlyon. Auditory grouping. In B. C. J. Moore, editor, *The handbook of perception and cognition, Volume 6, Hearing*, pages 387–424. Academic Press, Inc., 1995.

N. Deshmukh, R. J. Duncan, A. Ganapathiraju, and J. Picone. Benchmarking human performance for continuous speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, 1996.

K. S. Helfer and R. L. Freyman. The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America*, 117(2):842–849, 2005.

* N. Ma, P. Green, J. P. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, in press.

C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. Large-vocabulary audio-visual speech recognition: A summary of the johns hopkins summer 2000 workshop. In *In Proceedings of the Workshop on Multimedia Signal Processing*, pages 619–624, Cannes, France, 2000.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

J. L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93:B69–B78, 2004.

* X. Shao and J. P. Barker. Audio-visual speech recognition in the presence of a competing speaker. In *Proceedings of Interspeech 2006*, pages 1292–1295, Pittsburgh, PA, 2006.

* X. Shao and J. P. Barker. Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. *Speech Communication*, 2007. submitted.

W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in n oise. *Journal of the Acoustical Society of America*, 26:212–215, 1954.

Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 1273(335):71–78, 1992.

F. Wightman, D. Kistler, and D Brungart. Informational masking of speech in children: Auditory-visual integration. *Journal of the Acoustical Society of America*, 119(6):3940–3949, 2006.

(Publications marked with an asterisk have involved the support of the project.)