Can Audio-Visual Speech Recognition outperform Acoustically Enhanced Speech Recognition in Automotive Environment?

Rajitha Navarathna*, Tristan Kleinschmidt*, David Dean*, Sridha Sridharan*, Patrick Lucey*†

* Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia.
†Disney Research Pittsburgh, USA.
{r.navarathna, t.kleinschmidt, d.dean, s.sridharan, p.lucey}@qut.edu.au

Abstract

The use of visual features in the form of lip movements to improve the performance of acoustic speech recognition has been shown to work well, particularly in noisy acoustic conditions. However, whether this technique can outperform speech recognition incorporating well-known acoustic enhancement techniques, such as spectral subtraction, or multi-channel beamforming is not known. This is an important question to be answered especially in an automotive environment, for the design of an efficient human-vehicle computer interface. We perform a variety of speech recognition experiments on a challenging automotive speech dataset and results show that synchronous HMM-based audio-visual fusion can outperform traditional single as well as multi-channel acoustic speech enhancement techniques. We also show that further improvement in recognition performance can be obtained by fusing speech-enhanced audio with the visual modality, demonstrating the complementary nature of the two robust speech recognition approaches.
Index Terms: Speech enhancement, robust speech recognition, audio-visual automatic speech recognition, synchronous HMM

1. Introduction

Automatic speech recognition (ASR) has matured into a technology which is becoming more common in our day-to-day activities. In noise-free environments, word recognition performance has been shown to approach 100% [1], however the performance of these systems degrades rapidly in noisy environments. There are a number of methods available for making ASR more robust in these conditions, including model compensation, robust feature extraction and recognition algorithms, as well as speech enhancement techniques. Speech enhancement is a popular approach for this purpose as it requires little-to-no prior knowledge of the environment to effectively reduce the levels of noise in the speech signal, and ultimately improve recognition accuracy.

All of these techniques aim to improve the quality of the ASR system by operating only on the acoustic channel. An alternative approach applied with some success is using visual features extracted from the visual movement of a speaker's mouth region in conjunction with the acoustic channel to improve noise robustness. This is known as audio-visual ASR (AVASR) and a significant amount of research has been conducted in this field [2]. It is of significant interest to compare these two disparate approaches (i.e. speech enhancement versus visual information fusion) to improve the noise robustness of speech recognition in adverse environments. However due to the lack of data which can facilitate such comparisons (i.e. a dataset with audio captured via a microphone-array and video captured synchronously on multiple streams), it has been very difficult to compare whether acoustic speech enhancement or visual fusion is superior or whether these approaches can be combined to further increase robustness in noisy environments.

AVASR studies to date have generally concentrated on improving the quality of the visual information [2], and inherently assume that adding visual information to the ASR system will improve its robustness under noise. Moreover the answer to the question as to whether AVASR will perform better than ASR incorporating well-known acoustic enhancement techniques, such as spectral subtraction, or multi-channel beamforming, is not known. One of the only examples where this comparison was made in [3] where an AVASR system was presented incorporating spectral subtraction [4]. This system showed significant benefits in combining speech enhanced audio and visual speech information through late integration.

Since this study, there have been a number of advances in both speech enhancement and AVASR. Late integration approaches have been superseded by middle integration techniques such as synchronous hidden Markov models (SHMMs) which are considered to provide better speech recognition performance [2], and considerable improvements have also been made in visual feature extraction [2]. Further, the use of multi-channel speech enhancement techniques such as beamforming [5] have become more viable. It is therefore worthwhile re-examining the current state of AVASR by comparing state-of-the-art audio-visual integration with comparably advanced speech enhancement techniques. The key contributions stemming from our work are summarised below:

- We provide a comparison of the recognition performance of single channel and multi-channel enhanced speech with the performance of audio-visual speech using data from a challenging automotive environment (AVICAR [6]), which introduces a number of visual challenges, including changes in illumination and speaker pose as well as severe audio impairment arising from car engine, wind and road noise.
- We show that, SHMM-based audio-visual fusion can outperform traditional single as well as multi-channel acoustic speech enhancement techniques in automotive environment.
- We extend this study to also demonstrate the complementary nature of visual information and enhanced audio observed in [3] still holds true when using multi-channel speech enhancement algorithms and state-of-the-art middle integration techniques (i.e SHMM) for audio-visual fusion. Experimental results show that the combination of acoustic speech enhancement and SHMM-based

AVASR can provide further gains in accuracies.

2. Audio-based Speech Enhancement

Enhancement techniques can be broadly classified by the number of microphones used. Single-channel techniques are well suited to a number of applications, for example where hardware costs are a key factor. Multi-channel techniques, whilst increasing hardware requirements, can reduce the distortion introduced by single-channel techniques through the use of spatial filtering [5], which consequently improves ASR performance. In this section, we describe two speech enhancement techniques commonly used when comparing the performance of novel algorithms, and which represent a similar level of sophistication to the audio-visual fusion technique described in Section 3.

2.1. Spectral subtraction

Spectral subtraction (first proposed by Boll [4]) aims to estimate the spectrum of the clean speech signal by subtracting an estimate of the noise spectrum from that of the noise-corrupted speech. Subtraction typically takes place in the magnitude or power spectrum assuming that the noise and speech signals are statistically independent [7] and can therefore be regarded as being added acoustically.

The generalised magnitude-domain spectral subtraction rule derived from [4, 7] is defined as:

$$|\hat{S}_t(f)| = |Y(f)| - \alpha(f)|\hat{D}(f)|$$

$$|\hat{S}(f)| = \begin{cases} |\hat{S}_t(f)| & |\hat{S}_t(f)| > \beta|Z(f)| \\ \beta|Z(f)| & \text{otherwise} \end{cases} \quad (1)$$

where $|\hat{D}(f)|$ is the estimate of the noise spectrum calculated using time-recursive averaging, $|Z(f)|$ is either the instantaneous noisy speech signal magnitude $|Y(f)|$ or the noise magnitude estimate. The frequency-dependent subtraction factors, $\alpha(f)$, compensate for over- or under-estimating the noise spectrum, and β is the noise floor factor which ensures the clean speech spectrum cannot become negative. A number of variations to this magnitude subtraction rule have been proposed in literature, including subtraction in the power spectral domain [7], and multi-band spectral subtraction (MBSS) [8].

2.2. Delay-sum beamforming

Multi-channel beamforming combines the acoustic signals from all microphones to perform spatial filtering which differentiates the signal of interest from the background noise based on propagation delays between the source and each microphone. Having compensated the delays, microphone channels are individually weighted and combined in order to reinforce the speech signal. This is referred to as filter-sum beamforming which is represented as:

$$S(f) = \frac{1}{N}\sum_{n=1}^{N} G_n(f)Y_n(f)\exp^{-j2\pi f\Delta_n} \quad (2)$$

where N is the number of microphones, $Y_n(f)$ is the signal received at the n^{th} microphone, $G_n(f)$ are the filter coefficients ($G_n(f) = 1$ for Delay-Sum Beamforming (DSB) [5]), and the exponential term is compensation for the delay Δ_n.

3. Video-based Speech Enhancement

Performance of acoustic speech recognition can be improved using visual information, which has previously been demonstrated with speech enhancement and late audio-visual integration [3]. Since this work in 1997, the state of the art in audio-visual integration has transitioned to a middle-integration SHMM approach using a cascading appearance-based feature extraction technique [2], which will be outlined in this section.

3.1. Visual speech features

Visual speech is best discriminated by the movement of the visual articulators. Cascading appearance-based features, devised by Potamianos et. al [2] has been established as the state-of-the-art for visual feature extraction for AVASR [2]. The visual features are extracted based on a combination of two-dimensional separable, discrete cosine transform (2D-DCT) and an inter-frame linear discriminant analysis (LDA) technique to maximise separation between speech events in the visual features.

3.2. Audio-visual fusion

Middle integration schemes have been developed to allow classifier scores to be combined in a weighted manner within classification. It has the advantage over feature-fusion due to its ability to reliably weight each modality (i.e. audio features and visual features) on an individual basis. The most widely used technique for middle intergration is a SHMM, which can be viewed as a typical acoustic-speech, left-to-right, hidden Markov model (HMM), but with two observation-emission density functions, one for audio and one for video, rather than the single one in a typical HMM.

Accordingly, the observation-emission score of an SHMM state can be written using the audio and visual observation vectors, $o_{a,t}$ and $o_{v,t}$ respectively, as follows:

$$P(o_{a,t} : o_{v,t}|u) = P(o_{a,t}|u)^{\alpha}P(o_{v,t}|u)^{1-\alpha} \quad (3)$$

where α and 1- α are the audio stream and visual stream weighting parameter respectively and $0 < \alpha < 1$.

3.3. Speech enhancement with audio-visual fusion

Prior to this paper, the only work of acoustic speech enhancement techniques in combination with visual fusion was performed by Cox et al. [3]. In their paper, the authors found that adding visual information improved the performance more than the gains obtained using spectral subtraction. Apart from this work, there has been no detailed study reporting AVASR with more modern speech enhancement techniques and audio-visual feature extraction and fusion techniques. Using this as our motivation, we present a comparison of both AVASR using SHMM and audio-only speech recognition with single and multi-microphone speech enhancement.

4. Experiments

4.1. Evaluation protocol

The experiments were conducted using the AVICAR automotive speech database [6], consisting of audio and video files for 100 speakers. Each recording session contains speech recorded under five different driving conditions: i.e. idling (**IDL**), driving at 35mph with windows down (**35D**) and up (**35U**), and driving at 55mph with windows down (**55D**) and up (**55U**). The speech data consists of isolated digits, isolated letters, phone numbers

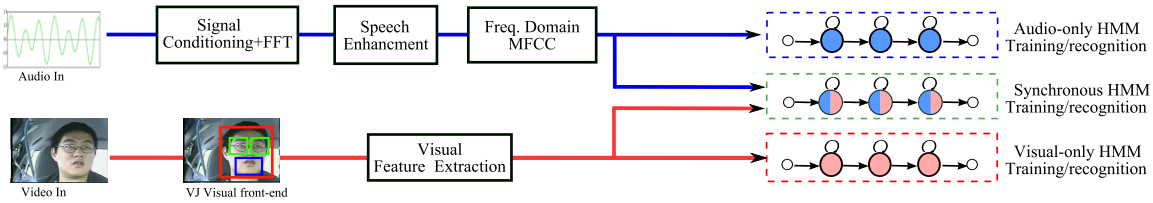


Figure 1: Block diagram of the AVASR system, which is a combination of audio-only and visual-only speech and audio-visual recognition systems.

and TIMIT sentences; in these experiments we use only the phone numbers task.

Speech recognition experiments have been performed using native-English speakers according to the protocol recently developed for this database [9], designed to ensure synchronisation between the audio and video streams. For each of the five evaluation folds, 3 groups of speakers (out of 5) were selected for training, the fourth for evaluation and system tuning, and the fifth for evaluation of the ASR system. All speech recognition results are collated by noise condition, and averaged across all folds and are reported in HTK-style word accuracies.

An overview of the experimental design is shown in Fig. 1, showing both the acoustic and visual speech recognition systems in combination with the audio-visual SHMM approach.

4.2. Audio-enhancement speech recognition

Experiments were conducted using several audio-based speech enhancement techniques. For the acoustic speech enhancement experiments, MFCC-based features were extracted using four speech enhancement techniques (and a baseline system without speech enhancement) operating in the frequency domain within a standard MFCC feature extraction process, as shown in Fig. 1. The five sets of 39-dimensional acoustic features used in this experiments were: (i) Baseline MFCC (13 MFCC including C_0, plus 13 delta and 13 acceleration coefficients) (ii) MFCC with spectral subtraction (SpecSub) (iii) MFCC with Kamath & Loizou's MBSS [8] (iv) MFCC with 2-channel delay-sum beamforming (2-ch DSB) (v) MFCC with 7-channel delay-sum beamforming (7-ch DSB).

Both spectral subtraction algorithms were optimised empirically across all data (i.e. not optimised for any particular noise condition) using the evaluation partitions of the evaluation protocol. The two microphones chosen for the 2-channel DSB system are symmetrically spaced around the center of the microphone array.

For each set of acoustic features, 9 state left-to-right HMM word models (i.e. zero, oh, one, ..., nine, sil) were trained to enable speaker-independent speech recognition. Each HMM state was represented using an 8 mixture Gaussian mixture model (GMM).

4.3. Visual speech recognition

In order to extract visual features for the visual-only and fusion experiments, a Viola-Jones-based visual front-end system, was used to track the mouth region [9]. Image mean normalisation, conceptually similar to cepstral mean subtraction in the audio domain, was used to remove any redundant information, such as illumination or speaker variances from the extracted mouth regions. A 2D-DCT was then applied to the mean-removed image, and the top 100 energy components were selected to capture the static visual features for each frame. In order to incorporate dynamic speech information, 3 neighbouring video feature vectors on either side of each static feature vector were concatenated, and projected via LDA (trained using the acoustic HMM states as classes) to yield a 40-dimensional feature vector. A separate word-model HMM was trained for video-only features, using 9 states and 8 mixture GMMs, similar to the acoustic HMM.

4.4. Audio-visual speech recognition

In order to investigate the audio-visual fusion performance, the baseline, spectral-subtraction and 7-channel DSB acoustic features were combined with the visual features using SHMM-based fusion. These acoustic features were chosen to represent no speech enhancement and the best-performing single and multi-channel approaches (see results in Section 5.1).

To ensure synchronisation of audio and visual features, the visual features were up-sampled from 30 Hz to 100 Hz using nearest-neighbour interpolation. For the AVASR experiments, the video and audio features were time-aligned and used to train left-to-right word-model SHMMs, with 9 states and 8 mixtures each for both the acoustic and visual streams.

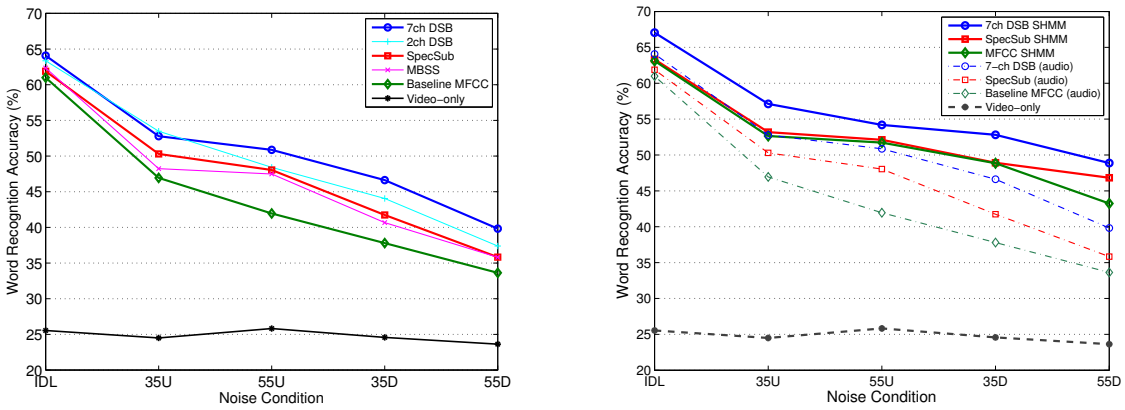
In order to demonstrate robustness over multiple noise conditions, the SHMM weighting parameter α was empirically chosen as 0.5, as this value was found to provide the best recognition performance when averaged across all noise conditions.

5. Results and Discussion

5.1. Audio-enhancement speech recognition

The audio-only ASR results with each of the speech enhancement algorithms are shown in Fig. 2(a). These results demonstrate a clear improvement over baseline MFCC performance by applying speech enhancement for all noise conditions. The 7-channel DSB technique outperforms all other speech enhancement techniques, with the dual-channel system providing the next best overall ASR performance. This result is consistent with the belief that microphone-array based speech enhancement is superior to single-channel techniques which typically distort the desired signal, and have access to less information about the audio signal [5]. It is also important to note that the spectral subtraction algorithm described by (1) outperformed Kamath & Loizou's multi-band spectral subtraction [8] when both algorithms were empirically optimised.

Based on these observations, we selected three of the speech enhancement techniques for fusion with the video features: baseline MFCC (i.e. a system without enhancement), spectral subtraction as the superior single-channel technique, and 7-channel DSB as the superior multi-channel enhancement technique.



(a) Audio, including speech enhancement, and video only

(b) Audio-visual SHMM fusion

Figure 2: Speech recognition performance over all noise conditions in the AVICAR database with (a) audio, including speech enhancement, and video only and (b) with SHMM-based audio-visual fusion. Only MFCC, spectral subtraction as per (1) and 7-channel DSB were chosen for comparison with audio-visual fusion.

5.2. Visual speech recognition

The visual speech recognition results are also shown in Fig. 2(a), and it can be seen that the accuracy using the AVICAR dataset differ marginally according to the noise condition, due to changes in lighting and head movement between driving conditions.

5.3. Audio-visual speech recognition

Fig. 2(b) reports the results of the AVASR experiments for the chosen acoustic speech enhancement techniques. The results show that the fused audio-visual system results increased or similar word accuracy compared with the ASR results in every noise condition, regardless of speech enhancement method chosen. For the case of the 7-channel DSB SHMM, this results in an average 10.2% relative word accuracy improvement over the 7-channel DSB acoustic features alone. This observation confirms that audio speech enhancement and visual speech information are still complementary when applied with state-of-the-art fusion techniques. Importantly, the addition of visual features to baseline audio features (i.e. MFCC SHMM in Fig. 2(b)) performs better than all speech enhancement techniques (except for 7-channel DSB enhancement in the IDL condition, where it performs similarly), confirming that incorporating visual speech with un-enhanced acoustic features is competitive with speech enhancement alone.

6. Conclusion

The fusion of visual features to improve the performance of acoustic speech recognition is known to work well, particularly in noisy conditions. However, limited studies have been conducted on the comparative performance of audio-visual speech recognition and audio enhanced speech recognition and in particular, no comparison has been made between the recognition performance with multi-channel speech enhancement algorithms and state-of-the-art middle integration techniques for audio-visual fusion.

In this paper, we perform a variety of synchronous HMM-based AVASR experiments and acoustic speech enhancement experiments and our results show that a simple MFCC-based AVASR system can outperform recognition based in the best of the acoustic speech enhancement systems using a challenging automotive audio-visual database. Experimental results also confirm that the combination of acoustic speech enhancement and SHMM based audio-visual speech recognition can provide further gains in recognition accuracies. This study will be used as a basis for our future investigations on further improvements to both audio and visual feature extraction in automotive environment, including the use of model adaptation in both acoustic and visual domains.

7. Acknowledgements

This work was supported through the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

8. References

[1] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA:, 2001.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *in Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[3] S. Cox, I. Matthews, and J. Bangham, "Combining noise compensation with visual information in speech recognition," *Proc. ESCA Workshop on AVSP*, pp. 53–56, Rhodes,1997.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Proc. of ICASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Berlin: Springer-Verlag, 2008.

[6] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: An audiovisual speech corpus in a car environment," *In Proc. Interspeech 2004*, pp. 2489–2492, Jeju Island, Korea.

[7] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of ICASSP*, Washington, DC, USA, 1979, pp. 208–211.

[8] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. of ICASSP*, Orlando, FL, USA, 2002, pp. 4160–4163.

[9] R. Navarathna, D. Dean, P. Lucey, C. Fookes, and S. Sridharan, "Recognising audio-visual speech in vehicles using the AVICAR database," *Int'l. Conf. on Speech Science & Technology*, pp. 110–113, Australia, 2010.