Special Session: Audio-Visual Spoken Language Processing

Jeesun Kim¹, Chris Davis¹ and Jinyoung Kim²

¹Department of Psychology, The University of Melbourne, Australia
²Electronics Computer and Information Engineering, Chonnam University, Korea

Abstract
The multimodal character of speech processing has attracted research endeavors that range from engineers working on automatic speech recognition through to psychologists interested in language processing. This paper serves as a broad introduction to the special Audio-Visual Spoken Language (AVSP) session. The paper focuses on recent developments in the area and touches on theory as well as application. We consider an important tool of the conception of AVSP has broadened: how issues concerning the developmental aspects of AVSP have been tackled; how measures of AV processing have become more sophisticated and how new applications incorporating the visual aspects of speech have been devised. The main aim of this limited review is to highlight the range of the issues and the innovative nature of research programs that fall within this area.

1. Introduction
Audio-visual speech processing is a research domain that exemplifies the cognitive science enterprise. The varieties of topics that contribute to this subject make it a challenging and stimulating study. The topic provides a forum in which specialists from different areas can learn from each other and experience anew the excitement (and discomfort) of the unfamiliar. In what follows we present a selection of themes that capture some aspects of developments in the field.

2. AVSP: A broad conception
If speech processing is primarily viewed in terms of auditory processing, then the concept of audio-visual speech processing appears somewhat quizzical. It might then come as a surprise to realize that the benefits to speech intelligibility of seeing the head and face of the talker (visual speech) have been established for at least fifty years [1]. An archetypal demonstration of AV speech processing is the McGurk effect [2] in which an auditory speech syllable [ba] dubbed onto the lip movements for [ga] results in the perception of “da” or “tha”. As Sekiyama and Burnham [this session] point out, this demonstration shows definitively that visual information is used in speech perception. Indeed, this phenomenon forms the basis of an important tool in investigating AV speech processing [3]. The investigation of the properties of AV speech, however, extends beyond this phenomenon and uses many other tools and methods (briefly outlined in Section 4). This diversity reflects the relevance of the multimodal perspective for a broad assortment of speech related disciplines. In general terms, AV speech phenomena can be characterized by the observation that although sensation might be controlled by the energy array, perception is more concerned with information. Information is modality neutral in the sense that the focus of communication is person-centered. That is, the embodied talker provides a “common fate” for the perceptual analysis of the signals they transmit.

Research falling under this overarching perspective ranges from that focused on an empirical examination of the linkage between the visual kinematic correlates of speech and speech acoustics [4, 5] to that concerned with language as a means of social communication in which an assortment of signals (including visual speech) act as regulators [6, 7; see Krahmer & Swertz, this session]. Indeed, it is interesting to note that recent research has provided links between the kinematic-acoustic studies and those studies investigating aspects of speech perception. Visual speech not only acts to provide robustness in transmission but also appears to assist in developing appropriate constraints on the analysis of the auditory input more broadly [8, 9].

The highlighting of visual signals in communication should not be taken as indicting that these are always crucial. It is obvious that speaking gives rise to energy patterns that can be relayed in many ways and such signals can be manipulated to stimulate only a single sense (e.g., audition via telephony). However, it is clear that seeing people speak affords so much more information. For example, we can gain tacit understanding by observing a conversation even in a language we do not comprehend. Indeed, compare the richness of face-to-face interaction (even with a stranger) to that available via email or the telephone (studies have shown that spoken dysfluencies occur far more often in two-person phone calls compared to face-to-face dialogues [10]). Further, at a more basic level, the study of speech perception and production from a multimodal perspective acts to remind researchers that speech is after all a dynamic physically-based process; both auditory and visual speech arise from movement.

3. The development of AVSP
3.1 Issues of time and scale
Given that speech information is conveyed by both auditory and visual signals then two basic questions arise: 1) How are these signals integrated to form a coherent percept? 2) What features of the stimuli control this process? The article by Schwartz and Catibari [this session] demonstrates that issues of architecture and control are still very much current. In what follows, we touch on some recent results that offer an interesting perspective on the issue of the way in which the audio and visual signals interact.

The study of AV speech is shaped by the properties of the underlying phenomena. Within both modalities, information exists over a range of time scales in which fine-grained rapidly changing details appear to be specified in the context of slightly slower changing properties (a nice way of capturing this was expressed in the paper by Vatikiotis-Bateson and colleagues [this session] who likened the movements of the
head to the source and movements of the face to the filter). Reflecting the richness of AV Speech phenomena, studies of AV speech processing have examined an assortment of research measures and techniques and a picture is emerging that visual and auditory speech signals might be integrated at a number of different grain-sizes and at various time scales. For instance, Munhall and colleagues [9] have recently demonstrated that observing the head movements associated with uttering a sentence enables the recovery of a degraded auditory signal. Davis and Kim [this session] have shown that participants can make reliable visual-visual and audio-visual matches about a spoken sentence purely on the basis of the movements of the upper part of the head and face. These studies suggest that the information provided by the upper part of the face (plausibly reflecting suprasegmental information such as prosody and amplitude) provide a constraint on lexical processing. This work has demonstrated that audio-visual integration might occur due to relatively long-term correlation between visual and auditory configurations [see also 8].

At the other end of the temporal scale, there are what Schwartz and colleagues [11, this session] call “very early” interaction effects. Such effects have been shown to occur over a time-scale of a few hundred milliseconds [see also, 12]. This work used a novel paradigm in which a specially constructed visual stimulus (consisting of a fixed sequence of lip gestures based on previously analyzed lip profiles) was used to demonstrate that the visual speech signal acted as a temporal cue to assist in auditory processing. That is, seeing the movement of the talker’s lips enabled the listener to better hear useful acoustic details. The development of novel techniques and perspectives such as those mentioned above have provided new viewpoints on established debates and provided fresh evidence for the ubiquity of visual speech in speech processing.

3.2. Modeling the process: Connection and control

As the previous section has indicated, the practice of experimentation and theorizing about AV speech processing can be innovative and is still in the process of setting out broad taxonomies within which to capture the different theoretical approaches. Schwartz and Cathiard [this session] examine the different types of models of AV speech processing using as a key the features of “when” and “how” the streams of information have been proposed to combine. Although this issue of has been a perennial one, new elements have been added to the debate. One is, as mentioned above, that new experimental work has indicated that the different streams of information might interact at numerous loci. Another new element has been in model evaluation, with different types of analyses brought to bear (e.g., the likelihood of the model in addition to the fit of the model to the data, see [13]). Perhaps the most intriguing development in terms of constraints upon models of AV speech processing comes from recent findings that show that AV interactions differ across individuals (see Schwartz and Cathiard [this session] and cultures (see Sekiyama and Burnham [this session, and below]). The emergence of these developments makes it clear how the issue of modeling AV speech has continued to evolve.

3.3. Cross-language comparisons

As outlined above, a basic question concerning AV speech processing is when and how auditory and visual speech information combines. However, posing the question this way has the implication that the process is relatively stationary, yet as Sekiyama and Burnham [this session] review, not only does AV speech processing vary across the course of an individual’s life [see 2, 14] but also across different cultures/languages. Cross language comparison of adults has produced behavioural evidence that the way that Japanese and Chinese people process visual speech is different from the use of English speakers [e.g., 15, 16]. The paper by Sekiyama and Burnham [this session] helps to pinpoint at what stage this difference occurs. That is, they report on data that indicates there is an increase in AV speech processing between 6 and 8 years for English but not Japanese children. Investigating the reason for this difference has the potential to reveal much about the development of AV speech processing and provide information about its function.

3.4. Adaptation and recalibration

Another issue that needs to be considered in developing models of AV speech processing concerns the response of the language processing system over time or multiple exposures. That is, up until now most studies have been concerning with capturing a single snapshot of AV speech. Vroomen and colleagues [this session] go beyond this one-off approach and have investigated effects that arise with repeated exposure. This work is based on the finding that visual speech can recalibrate auditory speech identification when the auditory and visual information conflict [17]. The study reported in this session explores the pattern of two types of after-effects over time and has examined how they build up. The results provide strong support for there being several different mechanisms involved in recalibration and adaptation.

4. Measurement techniques

Having briefly outlined some of the ways that issues about the interpretation of AV speech processing have been developed, in this section we review some recent measurement techniques. The first part will briefly describe some novel behavioral techniques and the second part will consider techniques for measurement of changes that occur simply as a consequence of producing or perceiving AV speech.

4.1. Experimental approaches and response measures

Recently new techniques have been introduced in the area that allow for sensitive measures of behavioral response and for efficient manipulation of visual stimuli. With regards to measures of behavioral response, it is interesting to note that the processing of visual speech has mostly been demonstrated by its effect on the perception of auditory speech, i.e. audio–visual interaction effects. The extent of the linguistic processing of visual speech by itself has typically been assessed by participants’ overt identification of words (so called “speechreading” see [18]). However, speechreading ability is variable and inaccurate (particularly with normal hearers, e.g., [18]). It might be that visual speech does not actually convey detailed linguistic information, or it might be that overt identification simply provides a conservative processing estimate. It is therefore important to develop a range of measures to index visual speech processing. The development new measures will help promote the study of visual speech per se and the integration of results into general theories of speech perception. One novel method for indexing
visual speech processing was recently developed by Kim and colleagues [19]. This method involved determining whether visual speech can activate (prime) other linguistic representations. This was done by showing a talker silently speaking before heard or seen target words. Observers were asked to either read the seen target word out aloud (naming) or classify it (lexical decision). Observers performed the tasks significantly faster (were primed) when the previously presented silent visual speech matched the target word (response priming did not occur when the visual speech only partially matched). This suggests that the information provided by visual speech was reasonably detailed. That priming was found in all tasks is consistent with the view that similar processes operate in both visual and auditory speech processing.

Another recent development in measuring the perception of AV speech has been to use 3D animation techniques [9]. These techniques can produce realistic heads that can be driven by captured head and face movements and have been used in studies of speaker and emotion identification [20, 21]. Such techniques allow for extensive stimulus control and also permit the systematic adjustment of visual speech properties. That is, particular movement types can be independently displayed or systematically altered via exaggeration or dampening. For example, Munhall et al [9] used this technique for the examination of the effect of the movements in identification of speech in noise. In the study, the head movement was eliminated or exaggerated while other parts of visual speech (i.e., non-rigid facial movements) were preserved.

4.2. Naturalistic measures of behavior

4.2.1 Measures from the outside. A number of innovative measurement techniques have been employed to determine the degree to which speech acoustics are linked to movements in the head, face and vocal tract [4, 5]. Vatikiotis-Bateson and colleagues [4, this session] have continued to develop sophisticated monitoring techniques (e.g., markerless tracking, [22]) with the aim of identifying key aspects of communicative realism. In conjunction with studies of the talker, there are an increasing number of studies that track the eye movements of observers (see [23]).

4.2.1 Measures on the inside. Modern brain imaging techniques enable the study of neural sites that underlie crossmodal processing. Sams [this session] provides an excellent review of the recent imaging literature. Such techniques have enormous potential for locating the sites involved in multimodal interactions well as the timing of such interactions. Different brain activity measures signal change over various time scales and so the literature presents a mosaic of temporal snapshots. For example, MEG and EEG (see Sams et al [this session] are able to register interactions over a few hundred milliseconds; however fMRI and PET provide less temporal sensitivity. An interesting new technique, TMS ([24 - Watkins, 2003]) can actively interfere with processing and is useful to provide confirmation of the function of particular regions. As is clear from this brief survey, an assortment of different imaging techniques has been used. As mentioned, these techniques have different spatial and temporal resolution. Furthermore, different behavioral tasks are often used by different researchers. Although the diversity of the AV speech processing committee is one of its strengths, the spread of research techniques and methods might also be a potential weakness.

5. Applications

5.1. Audio-Visual Automatic Speech Recognition

Research in AV speech is not only about human processing since visual speech information helps improve performance of automatic speech and person identification systems. Research on methods to improve both machine and human processing forms a major component of the practical application of the use of multimodal inputs. Her e we provide a brief overview.

5.1.1 AV databases. Miller and colleagues [this session] has raised an important issue for the Audio-visual speech community of the design and implementation of AV speech corpora. The current situation with AV corpora is reminiscent of that which faced auditory speech recognition community prior to 1986, i.e., the lack of widely accepted common data bases for system training and comparison. This situation is changing with more corpora being made available and data base design issues being discussed.

5.1.2 AVASR. The development of Automatic Speech Recognition (ASR) has proven to be a complex project that is subject to problems of robustness in uncontrolled environments. This has acted to inhibit the uptake of speech as a user interface method. A promising avenue in producing robust ASR is to include visual speech (as this is not affected by acoustic noise, i.e., AVASR). However, issues related to the visual front end still remain daunting (see [25] for a recent review). A visual front end would appear to require robust head and face feature and shape tracking (reliable across individual variation, pose, lighting, etc). It is clear from experimental work and measurement of the relationship between visual speech and speech-acoustics (see section 4) that in addition to tracking the mouth region it is important to track other dynamics, such as the jaw, cheek, upper face and head movements. One promising development in this area has been an approach that might incorporate joint shape and appearance models for 3D face modeling ([25]; also [26] for a similar idea about lip tracking). Apart from the visual front-end another essential feature of an AVS system concerns how the visual and auditory information are to be integrated. Once again, comparisons to the human system might be helpful, for as outline previously, it appears that there are many points of contact between the processing of audio and visual information over many different time-scales. Thus hybrid systems might be worth investigating, ones that integrate information at both feature and classification score level and are jointly trained with all properly tied parameters (e.g., see [25]).

5.2. Visual Speech synthesis

A clear application for AV speech is to incorporate visual speech into telecommunication applications, Berthommier [27, this session] has developed a new method of video synthesis of lip movements from speech sounds only and a model of low-level audio-visual interaction has been proposed. This idea is similar to that of the audio dubbing of a movie, but, here, this is video dubbing. This new concept is extremely well suited for particularly as it avoids heavy capture and transmission of video data.
5.3. AV training of speech contrasts
AV speech is potentially important for communication between people having different languages. The issue is whether language learning and interlanguage can be improved. Hazan et al. (this session) presents a number of results in this regard that indicate that the issue is complex and distinctions need to be made between what contrasts are to be made, perception and production and type of training.

6. Conclusions
The study of AV speech processing is a multifaceted research program that caters for diverse interest in speech science and technology. With progress in low-cost computer power and storage and more affordable imagining techniques, more researchers will be able to conduct sophisticated AV speech processing experiments. Indeed, as Bernstein and colleagues [28] have noted this research will progressively “invade” and energize more traditional speech research domains. Given this expansion, it will be important for researchers to keep in mind the importance of interaction and integration.

7. Acknowledgements
The first author wishes to acknowledge support from Australia Research Council (Grant number DP0209664).

8. References