ANALYSIS-BY-SYNTHESIS IN AUDITORY-VISUAL SPEECH PERCEPTION: MULTI-SENSORY MOTOR INTERFACING

Virginie van Wassenhove

Division of Biology, California Institute of Technology, Pasadena, California, 91125, USA
vvw@caltech.edu

ABSTRACT

In conversational settings, one *sees* as much as one hears the interlocutor. Compelling demonstrations of auditory-visual integration in speech perception are the classic McGurk effects [1]: in McGurk "fusion," an auditory [p] dubbed onto a face articulating [k] is perceived as a single fused percept [t], but in McGurk "combination," an auditory [k] dubbed onto a visual [p] is heard as multiple combinations of [k] and [p]. The natural spatio-temporal co-occurrence of auditory-visual (AV) speech signals is thus a likely feature used by the brain. AV speech integration offers interesting challenges for neuroscience and speech science alike. How, when, where, and in what format do auditory and visual speech signals integrate? A set of empirical studies are described, whose results suggest that multisensory speech integration relies on a dynamic set of predictive computations involving a large-scale cortical network (including sensory and motor systems). In building on the classical 'analysis-by-synthesis' framework [2], it is suggested that speech perception entails a predictive brain network, which computationally operates on abstract speech units.

Keywords: McGurk, EEG, fMRI, psychophysics, predictive coding.

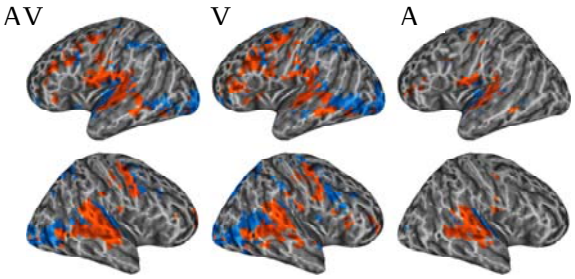
1. INTRODUCTION

Watching an interlocutor's face when engaged in a conversation provides information not only about identity or emotion, but also about facial kinematics. Speech kinematics provide a complex and dynamic structuring of visual information that can influence auditory speech processing. Multisensory integration in speech can occur at different levels of perceptual and neural processing (i.e. with more or less specificity with respect to the phonological categorization of the speech signal). For instance, the common source of AV (speech) signals is likely to influence auditory source localization [3] or enhance the detection of auditory speech [4]. In this paper, we focus on the neural mechanisms underlying the phonological categorization of AV speech signals.

The McGurk effect demonstrates that (i) the modality of input and (ii) the content of information received by each sensory modality lead to specific constraints on potential perceptual outcomes. The rate of McGurk fusion is taken as a means to evaluate the degree of AV speech integration. Although AV speech perception is an evolutionary and ecologically valid occurrence of speech signals, classic models of speech processing have focused on the *auditory* source of speech information and, as such, mostly use acoustic inputs as raw material. The evidence for the integration of visual (V) speech information with auditory (A) speech thus raises major issues for classical models of speech processing, particularly with respect to the interfacing of different sensory modalities and their underlying physiological pathways supporting these multisensory interactions.

The empirical findings described here support the view that visual speech plays a crucial role in shaping an internal prediction of the auditory speech signals that follow. It is proposed that the dynamics of the facial articulators perceived by the listener enable the brain to narrow down the set of potential auditory speech inputs. This visual-based prediction is shown to engage a large-scale brain network that includes the motor system. The latter is hypothesized to operate on the distinctive features of speech [2, 5] hence abstract (i.e., amodal) speech representations. The empirical results are consistent with an 'analysis-by-synthesis' (AbS) framework (e.g., Halle and Stevens [2, 5, 6]). The amodal AbS model provides a biologically plausible theoretical framework for speech perception reminiscent of the forward models described in studies of the motor system [7, 8].

Figure 1 (adapted from [9]): Cortical regions involved in auditory-visual, visual and auditory speech perception (conjunction analysis). The red areas indicate the regions of overlap between the production and the perception of identical speech syllables. Ventral Premotor (PMv) cortices are mostly apparent for AV and V speech perception. The blue areas are those regions uniquely active during the perception of AV, V and A speech and that do not overlap with speech production areas.



2. ANATOMICAL LOCI OF AUDITORY-VISUAL SPEECH INTEGRATION

Functional Magnetic Resonance Imaging (fMRI) is a useful imaging tool for anatomical study of brain function. In the current study, fMRI was used to visualize the function of brain areas involved in the integration of AV speech [9]. Specifically, we sought to contrast the contribution of the motor cortex in speech production with its role in perceiving A, V and AV syllables.

In a series of experiments, participants actively produced syllables or passively perceived A, V and AV syllables in the fMRI scanner. The AV stimuli consisted of both congruent AV [k], [p], and [t] and McGurk fusion stimuli (audio [p] dubbed onto a face articulating [k]). All consonants were presented in a pre-vocalic context (i.e., followed by the vowel [ae]) and all stimuli were natural speech.

The first finding shows that the pattern of cortical activation during the perception of V and AV speech greatly overlaps with that observed in speech production; this was not the case for auditory-alone speech perception (Figure 1). In particular, the areas showing greater than 50% of overlap in speech production and perception were bilateral the anterior and posterior Superior Temporal lobes (STa and STp, respectively), as well as the ventral premotor cortex (PMv).

Second, McGurk perceptual fusion elicited patterns of activation that correlated differently across cortical areas with the perception of a congruent AV [p] signal (i.e., the auditory component is identical to the auditory McGurk fusion stimulus), AV [k] (i.e., the visual component is identical to the visual McGurk fusion stimulus) or AV [t] (i.e., the perceived illusory [t] elicited by the McGurk fusion stimulus). The activations observed in the frontal motor areas and in the auditory and somatosensory cortices during McGurk presentation correlated more with the perceived syllable (AV [t]) than with the actual syllables presented in either sensory modality (A [p], V [k]). In the visual cortex, activation correlated most with the presentation of a congruent AV [k] (Figure 2). Note that the areas activated in the perception of AV speech converge with those areas hypothesized in recent models of speech perception [10, 11] and more generally, with areas involved in interfacing auditory and motor representations [10, 12].

Additional classification analysis highlighted differences in activation depending on the perceptual outcome of the McGurk fusion. In some instances, participants reported perceiving “ka” and this was accompanied by differential activation in the middle and inferior frontal gyri. When a participant reported the percept as being “ta,” activation in the nearby left somatosensory areas, PMv and primary motor cortices was differentially observed.

Additional consideration of the time course of activation across different cortical areas indicated a distinct and sequential pattern, sensory cortices were first observed to correlate with their congruent counterparts during McGurk perception. Motor, auditory and somatosensory cortices followed, which in turn correlated with the actual perception or phonological categorization of the listener. All together, these results strongly suggest an active participation of motor cortices in the phonetic interpretation of V and AV speech signals and hence in the perceptual categorization of AV speech. Specifically, we interpret these results as evidence that visual speech inputs provide the motor system with relevant information to synthesize a hypothesis with respect to the intended speech category emitted by the speaker. The information based on the configuration and dynamics of the facial articulators provided to the listener thus interface with auditory and somatosensory cortices via the motor system. An efferent copy mechanism could account for the

Figure 2 (adapted from [9]): Correlation analysis of auditory-visual speech stimuli perception across three cortical regions of interest. Black, gray and white bars show the correlation between activation obtained during the perception of a McGurk fusion stimulus and that obtained during the perception of a congruent AV[p], a congruent AV[k] and a congruent AV[t] stimulus, respectively.

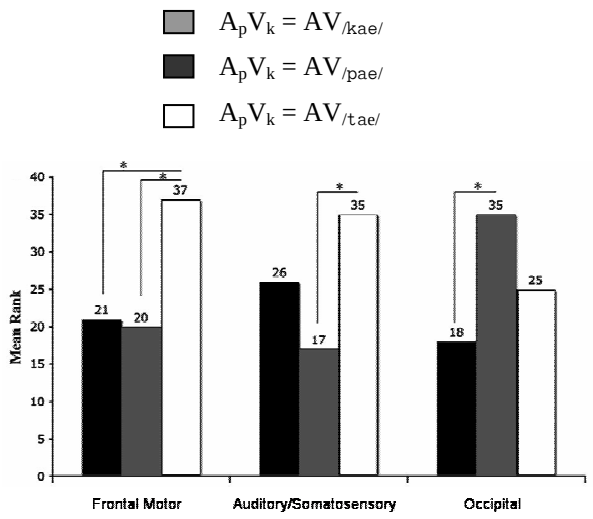
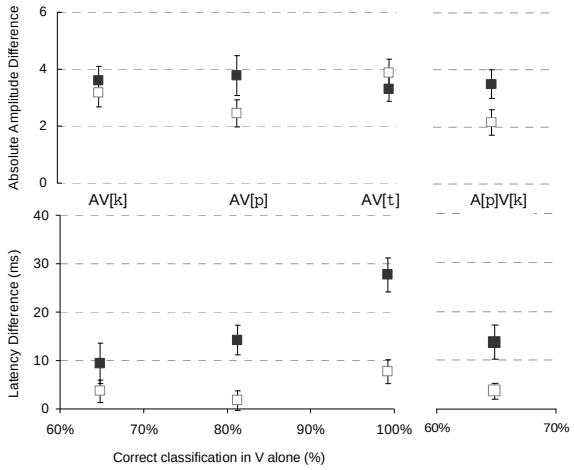


Figure 3 (adapted from [13]): Absolute amplitude (top panels) and latency (bottom panels) differences of the auditory evoked-related responses (N1 are open symbols, P2 are filled symbols) as a function of correct identification of visual speech. Top panel: a similar amplitude decrease for N1 (less negative) and P2 (less positive) is observed for all congruent and incongruent AV presentations as compared to audio alone presentations. Bottom panel: the better the identification of visual speech alone, the earlier the N1/P2 complex occurs



hypothesis testing of the internal predictor and the actual sensory 'feedback' (i.e., here, AV speech inputs). The time course of fMRI activation suggests that sensory cortices may serve as the loci of a hypothesis-testing mechanism in which incoming sensory speech inputs are matched against the internal (category) predictor. Note that within this framework, the *stronger* the prediction of the speech category, the less information is needed to confirm it.

3. TEMPORAL LOCI OF AUDITORY-VISUAL SPEECH INTEGRATION

In order to address the temporal dynamics and the *when* of AV speech integration, we used electroencephalography (EEG) [13]. The aforementioned fMRI study indicates that motor cortices may process facial articulatory gestures to build predictions of the categorization of incoming auditory speech. It is noteworthy that in AV speech the movements of the visible articulators during production often precede the acoustic utterance. Thus, if the precedence of visual speech is sufficient to build an internal hypothesis of incoming auditory speech inputs as suggested in the fMRI results, auditory event-related potentials may reflect this hypothesis-testing mechanism and provide more precise information on the dynamics of the hypothesized analysis-by-synthesis mechanism. Identical A, V and AV stimuli (including McGurk fusion) as in the fMRI experiment were presented to the participants while they were recorded under EEG. This time, participants reported what they "heard," "saw" or "heard while watching the face," respectively in a 3-AFC paradigm.

EEG recordings showed two major early modulations of auditory-specific ERPs in AV speech as compared to A-alone speech. First, the auditory ERPs occurring at ~100 to 300 ms post-auditory-stimulus onset showed a *decreased* amplitude to the presentation of AV speech as compared to A speech. Second, the *latency* of the auditory-specific ERPs was shifted significantly earlier as a function of the information provided by visual speech. For instance, a visual bilabial [p] was more easily categorized than a velar [k] in visual alone conditions. This is in agreement with the general classification of visemes [14]. Accordingly, we observed that the auditory-specific ERPs evoked by the presentation of AV

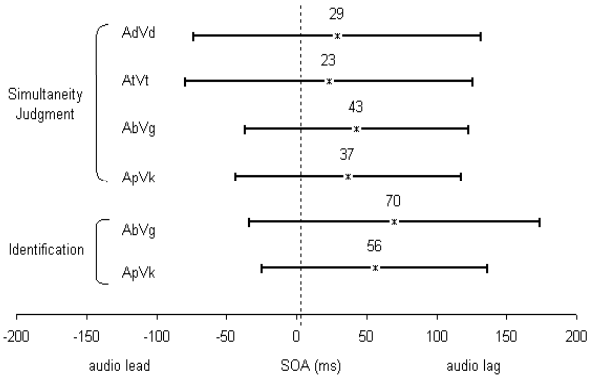
[p] occurred earlier than for A [p], whereas the presentation of AV [k] did not influence the latency of the auditory-specific ERPs (Figure 3).

These results suggest that the precedence of visual speech is in fact sufficient for the speech system to predict and reduce the set of possible incoming auditory speech inputs. Within the context of an 'analysis-by-synthesis' model of speech processing the auditory ERPs are interpreted as residual errors of the hypothesis-testing mechanism taking place between incoming internal predictions elicited by prior visual speech inputs and the incoming auditory syllables [11,13]. Additionally, the ERP data converge with the pattern of activation observed in the fMRI study and further suggest that the computations taking place in motor cortices allow the synthesis of a very specific predictor based on the visual inputs. Hence, the more salient the visual speech signals are (e.g., bilabials), the stronger the internal predictor and, in turn, the stronger the internal predictor is, the less information is used to confirm (or refute) it. This mechanism is reflected in the dependency of auditory ERPs latency on visual speech inputs coupled with a systematic amplitude reduction of the same auditory ERPs. Note that an amplitude reduction of auditory ERPs has also been reported for self-produced speech [15], suggesting that sensorimotor mechanisms in speech production and perception may generally affect sensory processing.

4. TEMPORAL WINDOW OF AUDITORY-VISUAL SPEECH INTEGRATION AND THE SYLLABLE

Given the importance of dynamics in AV speech perception, psychophysical studies were designed to estimate the temporal constraints of AV speech integration [16]. Two sets of AV speech stimuli were used (voiced and voiceless auditory bilabials dubbed onto visual velars). For these experiments, the auditory and visual speech signals were desynchronized in steps of 33ms and the stimulus sets ranged from the auditory signals leading the visual signals by ~500ms to the auditory signals lagging the visual signals by ~500ms. All stimuli were tested using (i) a speech *identification* task (3-AFC, similar to that used in the EEG experiment described above) and (ii) a temporal synchrony judgment task (i.e., a *detection* task). Overall, results showed that both AV speech detection and identification tolerated ~250 ms of asynchrony

Figure 4 (adapted from [14]): Temporal window of integration in AV speech. Identification and simultaneity judgment of AV syllables lead to ~250 ms window of desynchrony tolerance in the perception of AV speech.



between the auditory and the visual speech stimuli. Both McGurk fusion and congruent syllables followed this pattern (Figure 4). Additionally, AV speech integration evaluated by detection and identification of both congruent and incongruent tokens, tolerated visual leads better than auditory leads, leading to an asymmetrical temporal window of integration, consistent with the natural occurrence of the auditory-visual signals. The duration of the temporal window of integration approximates the average syllabic duration across languages [17], suggesting that syllables could be an important unit of computation in AV speech perception – as well as in speech perception in general [18, 19].

5. ANALYSIS-BY-SYNTHESIS IN SPEECH PERCEPTION

Analysis-by-synthesis was described by Stevens and Halle [2, 5, 6] as a model that uses the underlying motor representations of speech production for the analysis of speech perception. As in the "Motor Theory" of speech perception [20, 21], AbS posits a set of abstract representations that can be used in the production and in the perception of speech. It is the distinctive features of speech that constitute the set of possible representations [6].

Although visual speech was not directly addressed in the original description of AbS, I have described in this paper a set of empirical studies suggesting that the processing of visual and auditory-visual speech may be based on such a mechanism. In particular, the involvement of pre-

motor cortices suggest that the computational rules used in the production/perception of speech may be analogous to those described in forward-inverse models of motor production [7, 8]. In these models, the motor system builds a set of internal predictions (or estimates) of the sensory feedback expected to occur after motor production. Here however, fMRI and EEG results suggest that in AV speech perception, it is the categorization of the speech signals which is being processed. In forward-inverse models, the forward part consists in building a set of predictions as to the expected outcome of a movement; the inverse part consists in testing the prediction in order to control and adjust 'online' the movement. In such models, the efference copy consists in the predicted sensory feedback (based on the motor production) with the actual incoming sensory feedback. Any discrepancy between the predictors and the feedback enables the motor system to correct online the trajectory of a planned movement. How then do forward-inverse models relate to AbS and AV speech perception?

In AV speech perception, it is proposed that the internal predictions are triggered by the perception of visual speech, hence, no explicit articulatory movements or motor plans of the listener are taking place in the forward component of the model. Rather, as suggested by the fMRI data, the motor system may recover these motor plans by processing the articulatory movements of the speaker. The resolution of the prediction elaborated on the basis of visual speech alone will be constrained by that of visemes, which provides extremely valuable information for place-of-articulation. The strength of the predictor will be determined by such factors as the set of possible interpretations contained within a visemic category, the confusability of the visemic category with others and, of course, the initial signal-to-noise ratio of the original visual speech input. In this framework, the predictors are amodal in that they are fed back into the sensory cortices and may constrain the analysis of auditory, visual and somatosensory signals. Each of these predictions may carry the prediction of what 'the sensory feedback would be like' in each sensory modality if the listener was articulating the gestures of the speaker. At these levels, the efference copy is compared with incoming auditory speech, the residual of which ultimately adjust the perceptual outcome. Note that fMRI results also suggest that such comparison may be occurring also in somatosensory and visual cortices. For instance, STp may be a particularly relevant candidate for the interfacing of auditory and motor speech [10]. In this hypothesis, the residual error of an AV [p] is smaller than that of an AV [k] – leading to a faster categorization of the bilabial in comparison to a velar as observed in the EEG study. It is as yet unclear whether auditory speech processing would entail such mechanism when the auditory speech signals are unambiguous.

6. CONCLUSIONS

Natural AV speech provides a novel and ecologically valid means to study speech processing beyond its specification from the sensory modality of input. Advances in the understanding of how auditory, visual, and somesthetic signals combine in the brain during speech perception/production will not only benefit speech theories but also open new clinical applications in which synergistic multisensory interactions can be taken into account. For instance, cochlear implants could be optimized to boost auditory speech information not available visually.

The AbS is a model in which each processing step involves an *active* comparison of internal predictions and actual speech inputs (at least for visual or auditory-visual speech signals). Such a mechanism challenges classical neuroscientific views of speech processing in which information is analyzed in successive steps (i.e. in hierarchically organized pathways). Rather, it is suggested that when sufficient information is present, the speech system compares the internal predictions established in a preceding time window with incoming information. This may occur at each processing step. Although the global hierarchical organization remains, local feedback loops permit systematic and active re-adjustments of internal predictions based on the evidence of incoming sensory inputs. Such a computational design provides an efficient and rapid means of processing complex information based on prior knowledge in the system. The AbS model described in this paper suggests that visual speech is processed via a dorsal brain pathway involving motor cortices. The specificity to phonological processing is reflected in (i) the fMRI study where activations correlated with the phonetic interpretation of AV speech and

(ii) the auditory residual errors reported in the EEG study which correlated with the strength of the (visual) prediction, and which is presumably fed back from the motor cortices [24].

A necessary step to build upon these results is to determine what an auditory-visual speech *feature* may be for the brain. One promising avenue is that of the co-modulation between auditory and visual speech signals. The temporal window of integration observed in AV speech perception suggests a tolerance to temporal misalignments which is distinctive (i.e., a temporal resolution which is not observed in non-speech auditory-visual stimuli). However, this is not to say that AV co-modulation would be specific enough to extract the kind of information sufficient for speech categorization. The co-modulation patterns may be processed by specific pathways that integrate auditory and visual inputs as a single perceptual entity. The fact that articulatory gestures are also a specific case of 'biological motion' is particularly relevant. Evidence points out to a brain pathway located in the superior temporal cortex specializing in face motion processing and that is particularly sensitive to gaze and mouth movements (e.g., [23]). The superior temporal cortex also hosts multisensory neurons and is the site of intricate connections across different sensory areas [24]. Future studies may reveal that this pathway is relevant for the analysis of supra-segmental cues such as prosody.

In sum, I have argued that AbS offers a biologically plausible implementation of AV speech processing in the brain and that predictive mechanisms offer a novel approach to old theoretical issues.

7. REFERENCES

[1] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746-748.

[2] Stevens, K.N., Halle, M. 1967. Remarks on analysis by synthesis and distinctive features. In: Wathem-Dunn, W. (ed), *Models for the Perception of Speech and Visual Form. Cambridge, MA: MIT Press.*

[3] Macaluso, E., George, R., Dolan, N., Spence, C., Driver, J. 2004. Spatial and temporal factors during processing of audiovisual speech: A PET study. *NeuroImage* 21, 725-732.

[4] Grant, K.W., Seitz, P.-F. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197-1208.

[5] Halle, M. 2002. *From Memory to Speech and Back: Papers on Phonetics and Phonology, 1954-2002*. Berlin: Walter de Gruyter.

[6] Stevens, K.N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872-1891.

[7] Wolpert, D.M., Flanagan, J. R. 2001. Motor prediction. *Cur. Biol.* 11, 729-732

[8] Wolpert, D.M., Ghahramani, Z., Jordan, M.I. 1995. An internal model for sensorimotor integration. *Science* 269, 1880-1882.

[9] Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L. 2007. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex.* doi:10.1093/cercor/bhl147, Advance access published online on January 11, 2007.

[10] Hickok, G., Poeppel, D. 2007. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393-402, 2007.

[11] Poeppel, D., Idsardi, W. J., van Wassenhove, V. 2007. Speech perception at the interface of neurobiology and linguistics. *Phil. Trans. R. Soc. Lond. B.,* in press.

[12] Warren, J., Wise, R. J. S., Warren, J. 2005. Sounds do-able: auditory-motor transformations and the posterior temporal plane. *Trends in Neurosciences* 28, 636-643.

[13] van Wassenhove, V., Grant, K. W., Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci.* 102, 1181-1186.

[14] C. G. Fisher. 1968. Confusions among visually perceived consonants**.** *Journal of Speech & Hearing* 11, 796-804**.**

[15] Houde, J.F., Nagarajan, S.S., Sekihara, K., Merzenich, M.M. 2002. Modulation of auditory cortex during speech: an MEG study. *J. Cognitive Neurosci.* 14, 1125-38.

[16] van Wassenhove, V., Grant, K.W., Poeppel, D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598-607.

[17] Arai, T., Greenberg, S. 1997. The temporal properties of spoken Japanese are similar to those of English. *Proc. Eurospeech* 1011-1114.

[18] Greenberg,S., Carvey, H., Hitchcock, L., Chang S. 2003. Temporal properties of spontaneous speech – A syllable-centric perspective. *J. Phonetics* 31, 465-485.

[19] Greenberg, S. 2005. A multi-tier theoretical framework for understanding spoken language. *In: Greenberg, S. Ainsworth, W. A. (eds) Listening to Speech: An Auditory Perspective: Mahwah, NJ: Lawrence Erlbaum Associates*, 411-433.

[20] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. 1967. Perception of the speech code. *Psych. Rev.* 74, 431-61.

[21] Liberman, A.M., Mattingly, I.G. 1985. The motor theory of speech perception revised. *Cognition* 21, 1-36.

[22] Kiebel, S. J., van Wassenhove, V., Friston, Karl J., von Kriegstein, K. 2007. Predictive coding in speech perception. *Abstract. 13th Annual Meeting of the Organization for Human Brain Mapping, Chicago, USA.*

[23] Puce, A., Allison, T., Bentin, S., Gore, J.C., McCarthy, G. 1998. Temporal Cortex Activation in Humans Viewing Eye and Mouth Movements. *J. Neurosci.* 18, 2188-2199.

[24] Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A. 2004. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* 7, 1190-1192.