THE STIMULUS AS BASIS FOR AUDIOVISUAL INTEGRATION

Eric Vatikiotis-Bateson, Harold Hill, Miyuki Kamachi*, Karen Lander† Kevin G. Munhall**

* ATR Human Information Science Laboratories, Dept. 2, Kyoto, Japan
† Manchester University, Dept. of Psychology, Manchester, UK
‡ Queen’s University, Depts. of Psychology & Otolaryngology, Kingston, Ont., Canada

ABSTRACT

We argue here that examination of the stimulus source is a prerequisite to understanding how audiovisual (AV) stimuli are processed perceptually. This is based on mounting evidence that the act of speech production generates multimodal events whose audible and visible components are highly correlated with each other and the vocal tract source. How this multimodal structuring is exploited perceptually, or not, needs to be demonstrated by conducting studies that take the properties of the stimulus source into account.

1. INTRODUCTION

We all know that being able to see a speaker’s face enhances speech intelligibility under poor acoustic conditions. It was demonstrated experimentally almost 50 years ago [1], and there have been myriad studies since then fleshing out various performance factors associated with visual speech perception (speechreading) and AV speech perception. However, until recently very little attention has been paid to the stimulus source and what effect the production system has on shaping the information to be processed perceptually. Specifically, we are convinced that in order to understand what role AV integration and other key processes in multimodal perception play, it is first necessary to know what production of the multimodal stimulus brings to perception.

Another way to put this is that we cannot imagine that speech production and perception are independent of each other. Perception-action links in biological behavior, especially among con-species, are more likely the rule rather than the exception. This has received a great deal of notoriety recently with the interest in mirror neurons [2], and brain function (fMRI) studies in our lab have demonstrated something similar for the processing of AV speech [3].

Long before mirror-neurons were an issue, people were making strong claims about the linkage between the production and perception of acoustic speech. The most notable example is the Motor Theory of Speech Perception championed and refined for the past 40 years by Alvin Liberman and colleagues at Haskins Labs and elsewhere. In this view, knowledge of the stimulus source is essential to every act of speech perception.

In its earliest form, Motor Theory made the connection between perception of stop consonants and the behavior of the vocal tract [4]. During their pioneering work with the Pattern Playback, one of the earliest modern speech synthesizers, Liberman and Frank Cooper discovering that perceptual distinctions between ba, da, ga were easily stimulated by simple manipulation of the time-varying patterns of the formants following stop release. Over the years, the theory grew to accommodate modularity [5] and elements of action theory [6], which was heavily influenced by J.J. Gibson’s theory of direct perception [7]. In its later form, Motor Theory necessarily became more concerned with representation, culminating in the claim that the acoustic signal is a transparent map of the vocal tract gestures responsible for shaping the output acoustics [8].

It is not our intention in this paper to take a stand on Motor Theory in its current, highly-evolved state. Knowledge of the stimulus source may be essential to every act of perception, but whether or not this is done directly or mediated by representation (e.g., vocal tract gestures) remains an open question. We are content for now to insist only on the scientific necessity of approaching perception via a close examination of production (and vice versa). To do so we have examined physical and statistical linkages between acoustic and visual events that either were not previously known or were ignored.

There is considerable evidence, reviewed briefly in Section 2 (S: 2), that the act of speech production generates multimodal events whose audible and visible components are highly correlated with each other and the vocal tract source. The correspondence between audible and visible events suggest to us that the perceptual system processes AV stimuli rich in information about the stimulus source including the criteria for perceptual integration (among other things).

Two types of research in our lab are aimed at examin-
ing how these richly structured source signals are handled perceptually. The first is a long-term commitment to developing a rigorous procedure for using measured, multimodal production data to generate talking head animations, which are then used as stimuli in studies of multimodal perception. Since this process and its use in perception and brain function (fMRI) studies have been described elsewhere [9, 10, 3, 11], it is not discussed further in this paper. Sample animations can be seen following the links at http://www.his.atr.co.jp/kuratate. The second type of research attempts to set questions for AV speech perception that reconcile the findings for AV production with research in face perception. A study of this type is described in S: 3.

2. AUDIO-VISUAL SPEECH PRODUCTION

Our basic experimental paradigm consists of recording the speech acoustics and the motions of the head and face, while speakers produce scripted sentence materials and longer (> 1 min.) stretches of spontaneous speech. Such data have been analyzed for speakers of English, Japanese, French, German, and two dialects of Chinese (Mandarin and Cantonese). In addition, we have also collected corresponding vocal tract data for one speaker of Japanese and another of English. EMG activity of 7-9 orofacial muscles has been collected for these and other subjects. Fig. 1 depicts the configurations and dimensionality of measures within each domain and the two axes of analysis across measurement domains (for details, see [12, 13]).

The data are analyzed within and across measurement domains. Domain-specific analyses use principal component analysis (PCA) to determine the number of orthogonal dimensions required to describe the data within each domain. As shown in Fig. 2, only 7-9 components are needed to recover 99% of the variance, regardless of the number of measured dimensions; typically 51 (17 3D position sensors) for face motion, 14 (7 2D transducers) for vocal tract motion, and 10 spectral coefficients (10th order LPC) for the acoustics.

Being able to describe the data of each domain with small numbers of components is convenient for the cross-domain analyses, because matrices can be square and well-formed (i.e., no one-to-many mappings). Our initial cross-domain analyses used singular value decomposition (SVD) to estimate the multilinear correspondence between one domain and another. While the dimensionality of this mapping is also small, the recovered variance is not as high as for the within-domain analyses (see Fig. 3). The figure also shows that face motion is better estimated from vocal tract motion than the reverse, suggesting that vocal tract measures are somewhat richer in information than those of the face.

Linear techniques were also used to compute the correspondences between muscle EMG activity and motions of the vocal tract and face and between the face and the speech acoustics. The results indicated strong cross-domain correspondence [13] sufficient, for example, to synthesize intelligible speech acoustics from measured face motion alone. However, nonlinear estimation techniques were needed to estimate accurate face motion from muscle EMG activity or from the spectral and amplitude components of the acoustics [14].
Using these basic techniques we have also confirmed that:

– Vocal tract and acoustic correlates are distributed everywhere on the face below the eyes, not just at the mouth [15]; and different locations make specific contributions – e.g., small motions of the cheeks are highly correlated with tongue tip position [13].

– The jaw is the primary component of face (and vocal tract) motion followed by lip shaping – unless the speaker is French in which case the order is often reversed [10, 13].

– Visual events occur primarily at the same low rate as vocal tract events (< 10Hz) [13]. As can be easily inferred by calculating power spectra (FFT) for motions of the face and head during speech, there is little activity above 10Hz (-40 dB, -60dB at 20Hz).

– The correspondence between visible and audible events is continuous, if not uniformly high, and applies to all sounds rather than only those few that obviously require posturing of the lips and/or jaw [13].

– One of the simplest and most robust findings is that head motion is continuously and highly correlated with fundamental frequency (F0) during speech [16]. Thus, head motion may inform perceivers not only about prosodic events, but any events associated with pitch contours. The linkage between head motion and F0 may be due to a loose anatomical coupling involving laryngeal structures such as the cricothyroid cartilage and adjacent strap muscles associated with postural control of the head. When the head changes position, the strap muscles might impinge just enough on the larynx to entrain its changes of orientation to coincide with the direction of motion of the strap muscles. Although not yet demonstrated, it also stands to reason that head motion and acoustic amplitude are correlated, as it is impossible to speak emphatically without moving the head. In this case, motion of the vocal tract might be sufficient to entrain postural behavior of the head [17].

In summary, our studies have revealed high degrees of correspondence between the two modalities of visible events on the face and head and audible events in the acoustic signal. These events are primarily the direct result of configuring the vocal tract which simultaneously shapes the acoustics and deforms the face. Correspondence analyses, by definition, identify redundancies in the event structure; therefore, unrelated modality-specific events are outside the scope of these analyses, so we do not claim to have fully characterized the structure of the multimodal stimulus. Of course, we believe that perceivers capitalize on these intermodal redundancies, but this must be ascertained using stimuli whose structures can be manipulated systematically, hence our efforts to develop a talking head animation system with fully controllable and known production parameters.

3. MULTIMODAL TALKER IDENTIFICATION

In this section, we discuss a perception study that relies on and perhaps extends the scope of the structural linkage between the visible and audible characteristics of speech stimuli to include idiosyncratic aspects of behavior useful in talker identification.

<table>
<thead>
<tr>
<th>Learning</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 trials</td>
<td>A V } same sentence different speaker</td>
</tr>
<tr>
<td>10 trials</td>
<td>V A } same sentence different speaker</td>
</tr>
</tbody>
</table>

Fig. 4. Experiment scheme for testing talker identification across sensory modalities (A – audio only, V – video only). Trials were blocked by training modality.

There is a large body of research examining our abilities to identify static faces [18]. While the results are compelling, there is a growing awareness of the importance of motion for face processing tasks. The pilot study reported
here tested whether or not time-varying speaker characteristics are specified across sensory modalities. Speech is a natural context for face motion and often coincides with talker identification. The cross-modal task is depicted in Fig. 4, and consisted of modality specific learning phase followed by an identification test in the other modality. Subjects were introduced to a talker by either hearing or seeing a sentence production. They then had to identify the known speaker from two speakers’ productions of a sentence slightly different from the one used in the learning phase.

Stimuli were made from sentence productions of 40 speakers (20 male, 20 female), organized into 20 training-test trials in which one speaker was a target and the other the distractor. Trials were presented to 27 subjects in two blocks: 10 audio and 10 visual targets.

One-tailed t-tests showed that subjects identified the correct talker reliably better than chance (5 correct) for both target conditions (voice: \( t_{26} = 3.7, p < .05 \); face: \( t_{26} = 4.8, p < .05 \)). There was no difference between target conditions \( t_{26} = 0.3, \) n.s.\)

These pilot results suggest that the act of producing multimodal speech conveys information about talker identity that perceivers can access across modalities and despite differences introduced by slight differences in phonetic content and speaking episode. The effectiveness of one prior experience in the other modality shows that long-term familiarity with the speaker is not required. Finally, one modality does not appear to be more informative of the other.

4. REFERENCES


