PERCEPTUAL INTEGRATION OF VISUAL EVIDENCE OF THE AIRSTREAM FROM ASPIRATED STOPS

Connor Mayer
Department of Computer Science, University of British Columbia, Vancouver, British Columbia

Bryan Gick
Department of Linguistics, University of British Columbia, Vancouver, British Columbia
Haskins Laboratories, New Haven, Connecticut

Tamra Weigel
School of Audiology and Speech Sciences, University of British Columbia, Vancouver, British Columbia

D. H. Whalen
Haskins Laboratories, New Haven, Connecticut
Program in Speech-Language-Hearing Sciences, City University of New York, New York, New York

ABSTRACT

This study investigates whether indirect visual evidence of aspiration can influence speech perception as previously found for tactile information. Participants were shown video of a speaker producing the sequence "pom" and "bomb" in a noisy setting. In some tokens, a candle was visibly perturbed by aspiration. All participants were more likely to correctly identify “pom” and incorrectly identify “bomb” in the presence of visible perturbation, indicating that perceptual integration was taking place. This effect was stronger for participants who reported being consciously aware of the candle as a predictor. This indicates that ambient information can be incorporated in speech perception even when presented via an indirect modality, and that active attention can amplify this effect.

RÉSUMÉ

Cette étude observe si une preuve d’aspiration visuelle et non directe peut influencer la perception de la parole comme cela a été démontré dans le cas d’une information tactile. Les participants ont visionné des extraits vidéo dans lesquelles un locuteur produisait des séquences “pom” et “bomb” dans un environnement bruyant. Dans certains extraits, la flamme d’une bougie était visiblement perturbée par l’aspiration. En présence de l’indication visuelle de perturbation, les participants étaient plus susceptibles d’identifier correctement “pom” et de moins bien reconnaître les séquences “bomb.” Cet effet était d’autant plus fort, lorsque les participants étaient conscients du facteur prédictif de la bougie. Ainsi, une information ambiante peut être incorporée à la perception de la parole, même présentée sous la forme d’une modalité indirecte; cet effet peut être amplifié par une attention active.

1 INTRODUCTION

Perceivers of speech integrate visual and acoustic information from articulator movements, resulting in both interference (e.g., McGurk and MacDonald 1976) and enhancement (e.g., Sumby and Pollack 1954) of auditory perception. Only a few studies have investigated the role of other types of information in speech perception. Fowler and Dekle (1991) and Gick et al. (2008) observed that tactile feedback from the “Tadoma” method of speechreading was integrated even by those who had just learned the system. Gick and Derrick (2009) found that during auditory speech perception, perceivers integrated tactile information in the form of light air puffs. These puffs, delivered cutaneously on the hand or neck, were designed to resemble speech aspiration (Derrick, Anderson, Gick, and Green 2009). When puffs were present, aspirated stops were more often correctly identified as being aspirated, and unaspirated stops were more often misidentified as aspirated, showing that listeners integrate tactile information in auditory perception in much the same way as visual information. Light taps in the same location, without direct relevance to speech, produced no effect.

The goal of the present study was to examine the influence of a related form of information on speech perception: indirect visual evidence of speech aspiration. This type of information is novel in several important respects: while previous studies have found perceptual integration of direct results of articulation (e.g., visible or palpable articulator movements, audible fluctuations in air pressure), the information studied here relies on the influence of speech production on an entity other than the speaker (e.g., aspiration moving a candle, hair, fabric, etc.). In addition to this greater degree of remove from the information source, speakers have likely had less experience with this type of information, which may make it less likely to be integrated. It is worth mentioning, however, that Derrick and Gick (2013) found integration for puffs of air received on the ankle, a situation that perceivers presumably encounter even less frequently than on the neck or hand. Finally, there are potential issues related to timing: the strength of integration increases as stimuli become more synchronous, as shown for both audio-visual (Munhall et al. 1996) and audio-tactile (Gick et al. 2010) integration. The processing of visual information is relatively slow compared to acoustic information because of the time required for the photochemical processes in the rods and cones of the eye (Welch and Warren 1986) and the greater amount of neural processing required for vision (Levine and Shefner 2000: 347). Thus, the latency in the visual modality coupled with the delay introduced between the production of aspiration and the motion of the candle flame could prove too long for the indirect information to affect the percept.

We considered three possibilities: perceivers could exhibit similarly automatic integration to that shown in previous studies, they could show strategic incorporation which relies on actively attending to the indirect information and incorporating it in post-perceptual judgements, or they could show no use of indirect information at all.

2 METHODS

2.1 Stimuli

Stimuli were produced by a 23-year-old female native speaker of west coast Canadian English saying the words "pom" (short for "pomegranate") and "bomb", and recorded using a Sony Mini-DV Handicam and a Sennheiser MK66 short shotgun microphone. There were a total of nine conditions in the experiment, based on the presence or absence of a candle, the definiteness of the acoustic information (clear or ambiguous) and matching of audio and video speech information (matched or mismatched). The conditions were separated into three different groups for analysis. Conditions *no-candle-pom-ambiguous* and *no-candle-bomb-ambiguous* used the video from conditions *no-candle-pom-matched* and *no-candle-bomb-matched*, described below, but with ambiguous audio between “pom” and “bomb” created by morphing audio of randomly selected pairs of the two words from conditions *no-candle-pom-matched* and *no-candle-bomb-matched* using the program STRAIGHT, with equal weighting on each word (Kawahara 2003). Because morphing resulted in half the original sound files, both the “pom” videos and “bomb” videos in these conditions used the same audio. This condition was intended to factor out the unlikely possibility of facial cues disambiguating the sounds (e.g., Owen and Blazek, 1985). The previous two conditions make up the first group: a one-way design. Conditions *candle-pom-matched* and *candle-bomb-matched* had a candle placed approximately 18 cm in front of the speaker: in *candle-pom-matched*, the speaker said “pom”, visibly perturbing the candle by the aspiration of the /p/, while in *candle-bomb-matched* the speaker said “bomb”, and the candle was not perturbed because of the lack of aspiration of /b/. Conditions *candle-pom-mismatched* and *candle-bomb-mismatched* used the same video as conditions *candle-pom-matched* and *candle-bomb-matched*, but with mismatched audio: in condition *candle-pom-mismatched*, perceivers saw a video “bomb” accompanied by an auditory “pom”, while in condition *candle-bomb-mismatched* they saw the opposite. The above four conditions make up the second group: a 2 x 2 x 2 factorial design. Conditions *no-candle-pom-matched* and *no-candle-bomb-matched* were identical to *candle-pom-matched* and *candle-bomb-matched* except that the candle was placed to the side of the speaker, and thus was not perturbed. The previous two conditions make up the final group: a 2 x 2 x 2 factorial design. Condition *training* featured the candle to the side as in conditions *no-candle-pom-matched* and *no-candle-bomb-matched*, but with perturbation of the candle flame occurring at times not corresponding to the effects of the airstream. This condition was designed primarily for training purposes: perceivers were shown 10 tokens of it at the beginning of the experiment to downplay the significance of the flickering candle, decreasing the likelihood of a strategic response. Additional efforts were made to distract attention from the candle, such as placing a variety of props on the bar (chips, beer, etc.) and actors in the background. Aside from condition *training*, all conditions had 20 repetitions, resulting in a total of 170 tokens. Each token was approximately one second in length.

2.2 Participants

A total of 39 native North American English listeners participated. No participants had any training in linguistics nor any reported language or hearing problems.

2.3 Procedure

Participants were seated in a soundproof room and shown short video clips of the speaker producing the sequence “pom” and “bomb” in a noisy bar setting with multi-talker babble. The babble was mixed into the video signal and set to such a volume that correct auditory-only identification of the sounds was about 70% (based on a pilot study of ten listeners). This signal-to-noise ratio was kept constant across participants. Participants listened through a pair of headphones.

Participants were told to assume the role of the bartender and that the speaker was ordering a drink. They were given a forced-choice task to identify whether they heard “pom” or “bomb” in each video clip by pressing the left and right arrows on a keyboard. Aside from the initial presentation of condition *training* for training purposes, stimuli were presented in random order including all conditions. Half the participants pushed left for “pom”, the other half pushed right. Stimuli were presented and input recorded using Psyscope B53 on an iMac. When the experiment was completed, participants were asked if there were any aspects of the video that helped inform their responses. If they responded negatively, they were then asked whether they had been consciously aware of the candle flickering and whether they had used it in any conscious strategy to disambiguate the sounds. Although several participants who did not mention the candle in their initial response reported being aware of the candle after being prompted by the experimenters, all of them claimed not to have used it as a conscious decision strategy, and so were included in the negative response group. A total of 13 participants claimed to have incorporated the candle in their decision-making process while 26 did not. Data from the training condition were not included in the analysis.

3 RESULTS

Participants showed an overall bias towards “bomb” responses in all conditions (see fig. 1 and table 1). A paired t-test showed no difference in response between conditions *no-candle-pom-ambiguous* (67% “pom”) and *no-candle-bomb-ambiguous* (66% “pom”) across all participants [t(38) = 0.0336; p = 0.74]. This indicates that facial information alone was not sufficient for participants to distinguish between the productions. Tokens with ambiguous audio were therefore excluded from further analysis. The bias towards “pom,” which contrasts with the general trend in the data, may indicate that the ambiguous audio was more similar to acoustic “pom” than “bomb.”

Looking only at data where the candle was in the airstream, a 2 (“pom” vs. “bomb” audio; within factor) x 2 (“pom” vs. “bomb” video; within factor) x 2 (noticed vs. not noticed candle; between factor) repeated measures ANOVA on response across conditions *candle-pom-matched, candle-bomb-matched, candle-pom-mismatched* and *candle-bomb-mismatched* showed a significant effect of audio [F (1, 37) = 33.744; p < 0.001]; “bomb” was more accurately identified than “pom.” There were significant interactions between audio and video [F (1, 37) = 26.9392; p < 0.001], and between audio, video, and whether the participant noticed the candle [F (1, 37) = 8.047; p < 0.01]. The percentages correct by listener group for all conditions with the candle in the airstream are shown in table 1. This latter interaction indicates that having seen the candle as a useful perceptual cue affected participants' responses, suggesting that strategic responding may have occurred in participants who noticed the candle: we thus conducted separate analyses on participants who noticed the candle and participants who did not.

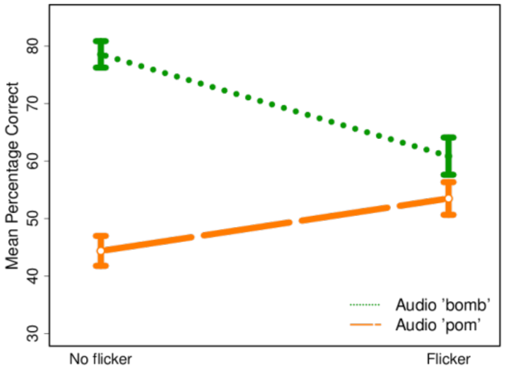


Figure 1: Interaction graphs with standard error bars across all participants in conditions with the candle present. Participants were more likely to respond correctly if the audio and video matched. Conditions *candle-pom-matched, candle-pom-mismatched, candle-bomb-matched* and *candle-bomb-mismatched*).

For participants who did not notice the candle, a 2 (“pom” vs. “bomb” audio) x 2 (“pom” vs. “bomb” video) repeated measures ANOVA showed significant effects for audio [F (1, 25) = 30.96; p < 0.001] and a significant interaction between video and audio [F (1, 25) = 14.1; p < 0.001], but no effect for video [F (1, 25) = 2.056; p = 0.164].

Audio	Video	Noticed candle	Did not notice candle
ba	ba	78%	79%
ba	pa	51%	66%
pa	ba	41%	46%
pa	pa	60%	50%

Table 1: Percentage of tokens correctly identified for conditions where the candle was in the airstream.

For participants who did notice the candle, a 2 x 2 repeated measures ANOVA showed a near-significant effect for audio [F (1,12) = 4.535; p = 0.0546] and a significant interaction between audio and video [F (1, 12) = 13.54; p < 0.01] but no effect of video [F (1, 12) = 0.429; p = 0.525].

For both groups of participants, the flickering candle induced more “pom” responses (see fig. 1). The “pom” visual signal both increased correct responses for audio “pom” (*candle-pom-matched*) and reduced correct responses for audio “bomb” (*candle-bomb-mismatched).* This effect was larger, however, for participants who were aware of the candle, explaining the interaction between noticing the candle, audio and video seen in this group.

A 2 (“pom” production vs. “bomb” production) x 2 (noticed vs. not noticed candle) ANOVA on conditions *no-candle-pom-matched* and *no-candle-bomb-matched* across all participants showed a significant effect for word being produced [F (1, 37) = 92.099; p < 0.001], with “bomb” being correctly identified (76%) more often than “pom” (44%). There was no significant effect for whether the participant noticed the candle [F (1, 37) = 0.747; p = 0.393] nor any interaction between the production and whether the candle was noticed [F (1, 37) = 0.025; p = 0.876]. This indicates that people were indeed responding to the candle and not facial cues.

4 DISCUSSION

Participants showed a bias towards “bomb” responses: indeed, the responses to “pom” audio are close to chance (see Table 1). This may be due to the Ganong effect (Ganong 1980): when presented with a stimulus that is ambiguous between a word and a non-word, listeners are more likely to choose the classification that results in a word. While “bomb” is a common word in English, “pom” is much rarer. This question could be studied in more detail by reproducing this experiment but having participants choose between “palm” (for those speakers who do not pronounce the /l/) and “bomb” instead.

All participants showed an increase in “pom” responses in the presence of a flickering candle. Depending on whether participants reported being consciously aware of it, however, the presence or absence of the candle in the airstream created by stop aspiration had different effects on their responses. Participants who reported being aware of the candle showed stronger integration and interference effects: although the increase in “pom” responses in the presence of a flickering candle held across all participants, those who reported being aware of it showed a higher rate of correct identifications of “pom” and incorrect identifications of “bomb.” This suggests that this kind of indirect evidence is still close to enough to the source to be unconsciously integrated in perception, but is also removed enough to be used as a strategic cue if listeners are consciously aware of it.

It is also noteworthy that the difference in correct classification between matched and mismatched video is more pronounced for the conditions with audio “pom” than those with audio “bomb” for all participants (see Table 1). This might indicate a difference in the use of positive and negative evidence: a flickering candle is stronger evidence for an aspirated stop than a steady flame is for an unaspirated one.

Despite no participants having linguistic training, the direction of the influence shows the correct association between a candle flicker and aspiration. This indicates that some implicit awareness of speech aerodynamics influenced perceivers’ interpretation of what a flickering candle should entail, regardless of whether they were consciously aware of its significance. Indeed, no participants who reported being aware of the candle were able to provide reasons for why aspiration and the flickering candle were associated, but only that they were. Neither group showed a difference between visual “pom” and “bomb” coupled with identical ambiguous audio, suggesting that participants were not able to use facial cues in differentiation; this accords with a lack of perceptual use of differences in face posture for distinguishing /p/ and /b/ (though /p/ and /m/ were distinguished) (Abel et al. 2011).

Previous studies have shown that both direct and indirect consequences of articulation, whether auditory, visual, or tactile, can influence perception. The present study supports and expands upon these results, showing that integration can be caused not only by primary sensory input but also by the secondary effects of speech on an external entity. Further research is needed to determine more clearly the limits of unconscious integration, the role of attention in multimodal speech perception, the differing roles of positive and negative evidence, and the extent of perceivers’ implicit understanding of those physical systems – factors that inform strategic incorporation of useful environmental information in speech perception.

ACKNOWLEDGEMENTS

Thanks to Donald Derrick, Mark Scott, Molly Babel, Naomi Francis, Denise Tom, Bénédicte Schoepflin, and the members of the UBC Interdisciplinary Speech Research Lab. This project was funded by an NSERC Discovery Grant to the second author and by NIH Grant DC-002717 to Haskins Laboratories.

REFERENCES

Abel, J., Barbosa, A.V., Black, A., Mayer, C., and Vatikiotis-Bateson, E. (**2011**). “The Labial Viseme Reconsidered: Evidence from Production and Perception,” *Proceedings of the 9th International Seminar on Speech Production (ISSP 2011)*. Montreal, Quebec, Canada. http://www.cogsys.ubc.ca/401/files/2011/05/issp2011-labials_final.pdf

Derrick, D., Anderson, P., Gick, B., and Green, S. (**2009**). “Characteristics of air puffs produced in English `pa': Experiments and simulations,” J. Acoust. Soc. Am. 125(4), 2272-2281.

Derrick, D., and Gick, B. (**2013**). “Full body aero-tactile integration in speech perception,” Multisensory Research. In press.

Fowler, C.A., and Dekle, D.J. (**1991**). “Listening with eye and hand: Cross-modal contributions to speech perception,” Journal of Experimental Psychology: Human Perception and Performance 17, 816-828. http://www.haskins.yale.edu/sr/SR107/SR107_04.pdf

Ganong, W. F. (**1980**). “Phonetic categorization in auditory perception,” Journal of Experimental Psychology: Human Perception and Performance 6, 110–125.

Gick, B., and Derrick, D. (**2009**). “Aero-tactile integration in speech perception,” Nature 462, 502-504.

Gick, B., Ikegami, Y. and Derrick, D. (**2010**). “The temporal window of audio-tactile integration in speech perception,” J. Acoust. Soc. Am. 128(5), EL342-EL346.

Gick, B., Jóhannsdóttir, K., Gibraiel, D., and Muehlbauer, J. (**2010**) “Tactile enhancement of auditory and visual speech perception in untrained perceivers, J. Acoust. Soc. Am. 123(4), EL72-76.

Kawahara, H., and Matsui, H. (**2003**). “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” ICASSP V.1 Hong Kong, 256–259. http://www.wakayama-u.ac.jp/~kawahara/PSSws/kwhrv2.pdf

Levine, M. W. (**2000**). *Levine and Shefner's Fundamentals of Sensation and Perception* (Oxford University Press, New York).

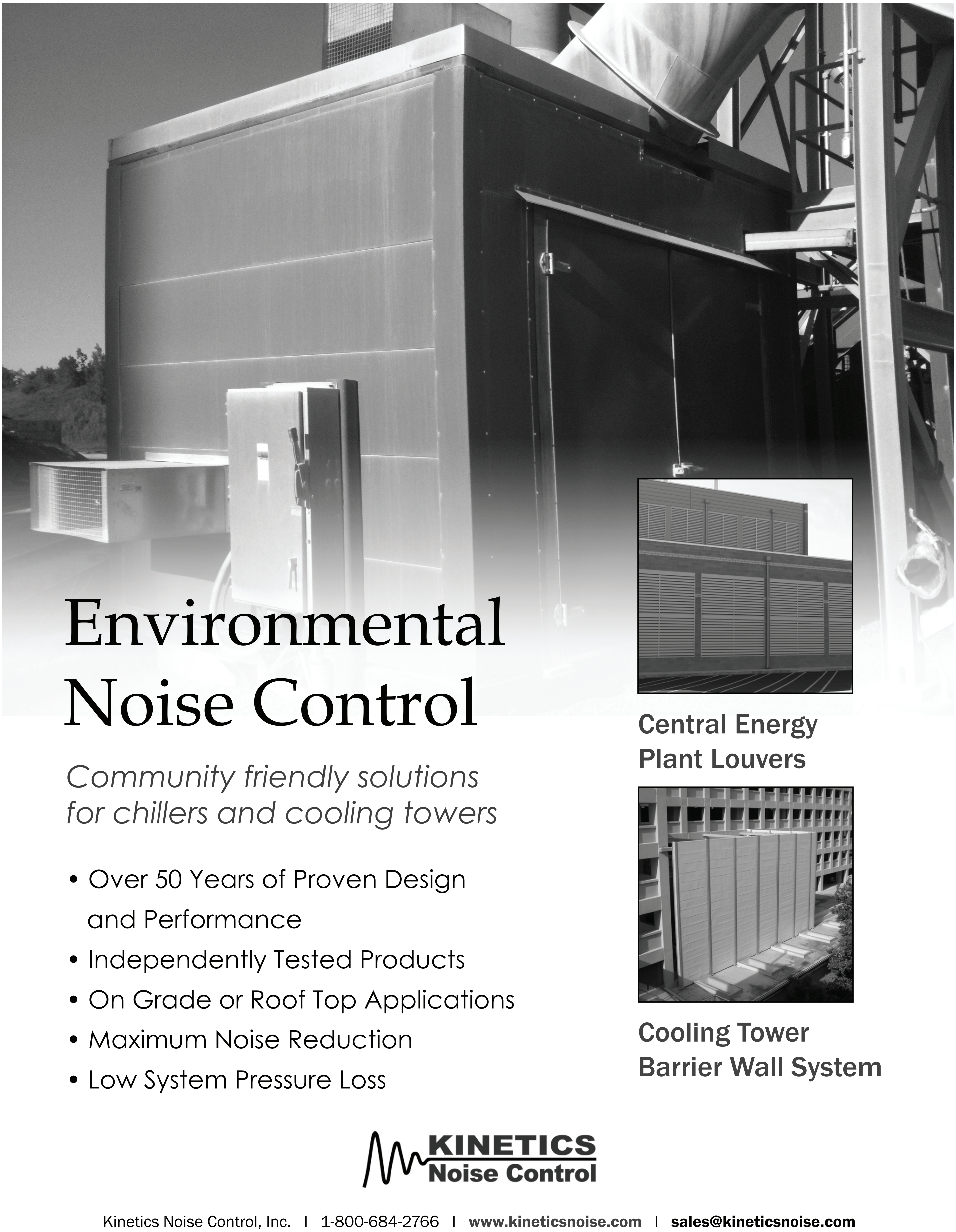
McGurk, H., and MacDonald, J.W. (**1976**). “Hearing lips and seeing voices,” Nature 264, 746-748.

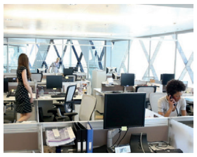
Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. (**1996**). “Temporal constraints on the McGurk effect,” Perception and Psychophysics 58(3), 351-362. http://gribblelab.org/publications/1996_PercepPsychophys_munhall.pdf

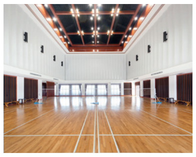
Owens, E., and Blazek, B. (**1985**). “Visemes observed by hearing-impaired and normal-hearing adult viewers,” Journal of Speech and Hearing Research 28, 381-393.

Sumby, W. H., and Pollack, I. (**1954**). “Visual contribution to speech intelligibility in noise,” J. Acoust. Soc. Am. 26, 212-215.

Welch, R. B., and Warren, D.H. (**1986**). “Intersensory interactions,” in *Handbook of Perception and Human Performance*, edited by K. Boff, L. Kaufman, and J. Thomas (Wiley, New York).







Out Now!
Cadna R®
Prediction of
Noise Levels inside Rooms
Software & Manual
Software & Manual
DataKustik
DataKustik