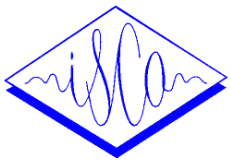

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

AUDIO-VISUAL SPEECH PERCEPTION WITHOUT TRADITIONAL SPEECH CUES: A SECOND REPORT

Robert E. Remez
Jennifer M. Fellowes
Barnard College
New York, New York

David B. Pisoni
Winston D. Goh
Indiana University
Bloomington, Indiana

Philip E. Rubin
Haskins Laboratories
New Haven, Connecticut

ABSTRACT

Theoretical and practical motives alike have prompted investigations of multimodal speech perception. Theoretically, such studies lead the explanation of perceptual organization beyond the familiar modality-bound accounts deriving from Gestalt psychology. Practically, existing perceptual accounts fail to explain the proficiency of multimodal speech perception using an electrocochlear prosthesis for hearing. Accordingly, our research sought improved measures of audiovisual integration of videotaped faces and selected acoustic constituents of speech signals with an acoustic signal that departs from the natural spectral properties of speech. A single sinewave tone accompanied a video image of an articulating face; the frequency and amplitude of the phonatory cycle or of one of the lower three oral formants supplied the pattern for a sinewave signal. Our results showed a distinct advantage for the condition pairing the video with a sinewave replicating the second formant, despite its unnatural timbre and its presentation in acoustic isolation from the balance of the speech signal.

INTRODUCTION

How does the perceiver find the speech signal amid an uninterrupted flux of sensory activity? The customary answer to this question discusses the principles of perceptual organization intrinsic to each of the sensory modalities, following Wertheimer (1923). In essence, two classes of general principle, visual and auditory, are available to apply to speech, and few clear proposals aim to explain perceptual organization when the listener also looks at the talker. There is hardly any doubt that multimodal perceptual organization does actually occur, and a small but sturdy literature describes perceptual phenomena that falsify the description of post-perceptual integration (for instance, Green & Miller, 1985).

Research on sinewave replicas of speech has been singular in promoting a different approach to the problem of organization than is customary (Julesz & Hirsh, 1972). Although a tonal analog of a speech signal is intelligible (Remez, Rubin, Pisoni & Carrell, 1981), it lacks the typical acoustic manifestations of vocal sound production. This fact arguably demonstrates the limitations of general auditory accounts of perceptual organization (Bregman, 1990) and probabilistic accounts of speech perception that rest on likely correspondence of signal element and phonetic segment (Massaro, 1994). Instead, findings with sinewave replicas of speech provide evidence of an alternative account of perceptual organization based on susceptibility to the unique spectrotemporal characteristics of a phonologically modulated source of sound (Remez, Rubin, Berns, Pardo & Lang, 1994). Perceptual organization, in this view, exploits a perceiver's sensitivity to patterned spectra, in contrast to the piecemeal assessment of elemental details of the acoustic stream warranted by prior accounts. These findings belong to an emerging class of reports about speech which note the integration of sensory elements despite detailed dissimilarity in their physical properties.

Multimodal Perceptual Organization of Speech

Although the perceptual organization of speech in an auditory system appears well characterized from a consideration of sinewave replicas, it also seems that multimodal speech perception exhibits some of the main characteristics identified by this line of research. Principally, the organization of visual and auditory inflow in a bimodal case of speech perception appears to conjoin stimulation from the modalities preliminary to an analysis of the patterned unimodal sensations. A clear case of this phenomenon is seen in a report by Green & Miller (1985) who observed that the identification of syllables in an auditory voicing series was a function of silent visual information about the rate of articulation. Had the rate information been specified acoustically, the outcome of the tests would have been explained agreeably as evidence of a kind of context for analysis of the spectrotemporal acoustic pattern that varied to evoke an impression of voiced and voiceless consonants. In the bimodal case, though, there is no perceptual function readily available to explain the lability of phonetic analysis to a combination of visual and auditory stimulation. Although Welch & Warren (1980) held that multimodal integration might depend on a common spatial locus for sound and sight, this premise falsely predicts failure of dichotic fusion of speech

(Broadbent & Ladefoged, 1957; Remez et al., 1994). The finding of Green & Miller (1985) is especially provocative considering that their subjects perceived a phonetic contrast that depends on fine resolution of sequential patterning, indicating that sensory streams are combined in a manner that is temporally veridical.

Two studies set the question of multimodal organization directly. In one, by Breeuwer & Plomp (1985), speechreading was supplemented with pure tones modulated at the frequencies of the first and the second formant. Subjects transcribed the audiovisual conditions relatively poorly, as if the tone analogs of the formants were barely fused with the visual impression of the articulating face. In contrast, Bernstein et al. (1992) used an acoustic or tactile presentation of the frequency band of the first or the second formant, and observed great benefit to speechreading of either F1 or F2 in a concurrent auditory signal, and an enhancement of speechreading with a tactile vocoder driven by the variation in the frequency region of F2. Clearly, a tone reproducing the frequency variation of the second formant cannot both be effective and ineffective in audiovisual presentation.

The Problem of the Second Formant

One clue about the cause of the different effects is the method used in each study to analyze the formant pattern. Breeuwer & Plomp argued that accurate assessment of formant frequency cannot be accomplished in real time. Their goal of assessing the prospects of an instrumental aid to perception required them to use existing signal processing technology, and they adopted linear prediction with minimal correction to determine formant values for voiced speech only. Although we can be confident that the temporal alignment of the resulting frequency modulated tones was accurate, the unvoiced formant values were simply missing, and other samples were unquestionably erroneous due to interpolation when the LPC analysis simply failed. This was not a completely satisfactory test of the perceptual organization of time-varying auditory and visual stimulation during speechreading because the auditory values were probably misleading.

The group led by Bernstein used the labels F1 and F2 to describe the patterns produced by their vocoders, but in actuality they used the output of stationary filter banks that approximated the range over which the first or second formant frequency excursions occurred. For F1, this was 75-900 Hz; for F2, it was 975-2625 Hz. It is likely, therefore, that the nominal F2 often included the third formant, and it is possible that the nominal F1 contained the second formant for some back vowels and labial consonants. This method fell short of an exact test of the perceiver's disposition to organize visual displays of the face and individual formant bands in speech perception.

Our own recent attempt to provide a clear resolution to this multimodal problem of integrating the second formant and the visual impression of a talker was less than successful (Saldaña, Pisoni, Fellowes & Remez, 1996). We used single tones from sinewave utterance replicas in combination with a video display of the face, and found that the greatest benefit to normal-hearing subjects occurred when the moving image of the face was combined with the tone analog of the second formant. Other multimodal conditions included tone analogs of the first formant, of the Fø pattern, and a noise band modulated in amplitude according to the overall energy in the signal. The finding of greatest benefit attending the audiovisual combination of the second formant analog and the face occurred without natural timbre, of course. This result is consistent with prior findings by Bernstein et al. (1992), and suggests that accurate estimates of the frequency of the second formant produce benefits in the multimodal case, contrary to Breeuwer & Plomp (1992) who used uncorrected linear-prediction estimates.

However, the performance levels in our earlier study were low (Saldaña et al., 1996). In a control condition using complete tonal replicas based on the utterances of this talker, average performance did not exceed 35% of syllables correct, whereas more typical performance on sinewave sentences approaches 80% correct. The cause, we suspect, was the talker, whose speech was unpredictably difficult for our listeners, a possibility which we verify in the present study.

THE PRESENT TEST OF MULTIMODAL INTEGRATION

To conduct a fairer test of multimodal coherence, we based our audio-visual presentation on the speech of a demonstrably intelligible talker (Bradlow, Torretta & Pisoni, 1996) to attempt to bring test performance off the floor, thereby resolving any differential effects of the single tones in combination with the video presentation. On the basis of the performance in this dataset, we recruited an individual to read a sentence list while video and audio signals are sampled. The natural speech was converted to sinewave replicas, and multimodal coherence was assessed in transcription tests combining the visual presentation with the tonal analog of the first, second or third formant; and with a tone replicating the pattern of the fundamental frequency of phonation.

Test materials. An adult female whose natural speech was verified as highly intelligible produced utterances that were sampled for video and audio reproduction. Ten sentences were selected from the dataset of Bradlow et al. (1996) and were spoken from a list.

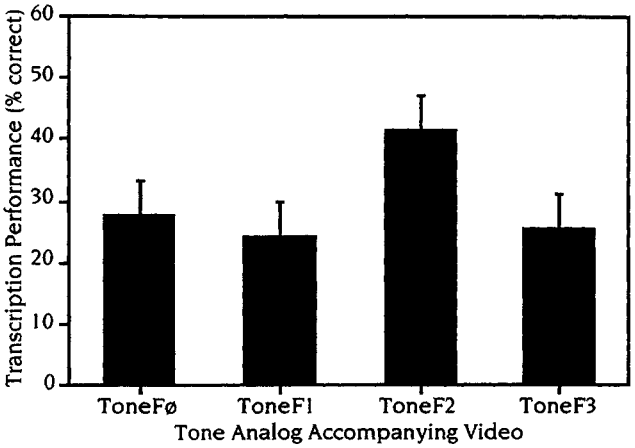

Figure 1. Results of a test of audiovisual speech perception with tone analogs of speech. Each bar shows the group performance with a different signal component. Error bars represent the confidence region for a post hoc means test (Tukey, $\alpha = .05$).

Formant center frequency and amplitude were estimated interactively by comparing discrete Fourier spectra and linear prediction estimates; frequency of phonation was estimated from a narrow-band Fourier representation of the spectrum. Frequency and amplitude values for Fø, F1, F2, and F3 were converted to four single-tone time-varying sinusoids using a software synthesizer (Rubin, 1980). The computed sinusoidal waveforms were combined and synchronized with the video, and presented on-line to listeners in individual testing carrels.

Procedure. Thirty-eight test takers were assigned randomly to one of four audiovisual conditions, Video+ToneFø, Video+ToneF1, Video+ToneF2, and Video+ToneF3. Each session began with a sequence of eight four-tone sinewave sentences, presented to give the subjects a brief chance to adjust to the odd timbre of sinewave signals. These test items were based on the speech of one of the authors, and did not duplicate any of the test sentences used in the multimodal conditions.

Following the familiarization sequence, an audiovisual condition began. Each sentence was repeated five times, after which the subject was cued to write a faithful rendition of the message in a specially prepared test booklet. A warning tone also occurred before the start of a new sentence block, to alert the subject to finish writing and to look at the video monitor.

At the conclusion of the audiovisual test conditions, a repetition of the initial set of eight audio sinewave sentences occurred. This served as a check on the absolute ability of subjects to derive phonetic impressions from sinewave signals. This sort of assessment has been necessary due to the immunity to the phonetic properties of sinewave signals of a substantial subset of volunteer subjects, though none of the thirty-eight participants in this sample was excluded on such grounds (see Remez et al., 1994).

Results. A transcription provided by a participant in an audiovisual test was scored by tallying the percent of the syllables in each sentence that had been transcribed correctly, following established procedure (Remez et al., 1981). Each subject contributed ten values, one for each of the test sentences, to a one-way analysis of variance of the effect on transcription performance of the four single tones combined with the video sample. Despite the small size of the groups (10 subjects in the first and second groups, 9 subjects in the third and fourth groups), the effect of the tone manipulation was found to affect performance [$F(3, 34) = 5.99$, $p < .002$]. The group performance in the four test conditions is shown in Figure 1. It is plain to see that the tone analog of the second formant, in combination with the video samples, produced performance that was significantly better than that which we observed in the three other tones.

DISCUSSION

The pattern of results, in which the tone analog of the second formant combined more effectively with the video samples than the other single tones that we tested, suggests an interpretation of the three studies that had set the specific empirical problem for us. First, although we derived test materials from visual and acoustic samples of the speech of an intelligible talker, the performance levels here replicated the pattern of our earlier observation (Saldaña et al., 1996). A tone exhibiting the pattern of F2 made a more effective acoustic accompaniment to the video samples than did the tone analogs of the other formants or the fundamental, and our findings show that there is no second best; performance in three conditions with the other tones was equal.

On the reports of prior research, we might have expected the analog of the first formant (Bernstein et al., 1992) or of the fundamental frequency of phonation (Rosen, Fourcin & Moore, 1981) to combine readily with the video samples in evoking an impression of the linguistic message. Differences in the linguistic test materials are important to consider, because the unforgiving sentences that we used here may have inadvertently suppressed the differences in effectiveness of tones other than ToneF2. Nonetheless, for multimodal perceptual organization in which the auditory component lacks the timbre of natural speech, it is safe to conclude that the unique effectiveness of the analog of the second formant is established more solidly by these results.

Second, a comparison is also appropriate of this multimodal circumstance to the effects of dichotic presentation of sinusoidal sentence components (Remez et al., 1994). In that study, one ear received an isolated tone analog of the second formant, the other ear received the

balance of the tones composing the sentence replica. Transcription performance for the concurrent presentation well exceeded the performance predicted by assaying the intelligibility of the components separately. The same kind of concurrent benefit is likely to have obtained in the multimodal case presented here. Neither the video samples of the talker's face nor the impressions evoked by the analog of the second formant are known to elicit accurate or definite impressions of the phonetic properties of a message. Yet, in analogy to the dichotic case, the concurrent presentation allowed listeners to organize the multimodal inflow and to transcribe about half of the syllables correctly in a difficult set of sentences.

Coincidentally, the performance levels are roughly the same for dichotic sinewave sentences and multimodal sinewave sentences. A clue to perceptual organization may reside in this similarity. If a synthetic second formant exhibiting natural timbre is more effective multimodally than a tone analog of F2, this would indicate that organizational functions may be contingent on short-term spectrum in some instances. Alternatively, prolonged exposure to sinewave signals may acclimate subjects to the anomalous timbre of the sinewave voice, and such a procedure may be seen as improvements in performance due solely to perceptual tuning.

Last, the principle that we proposed to explain the dichotic combination of acoustic information was based on susceptibility to the spectrotemporal patterns of an acoustic signal independent of its superficial properties. Specifically, in the case of speech the principle is evidently matched, albeit abstractly, to the physical structure of vocal resonators and the functional organization of phonologically governed articulation. Because a sinewave differed physically from the acoustic signal elements it replicated in coarse grain, no perceptual evaluation of elementary "speech cues" alone would accommodate the finding.

To accommodate the multimodal case of speech, the organizational principle satisfied by the auditory and the visual inflow must be still more abstract. By such means the perceiver treats the sensory inflow as information about a unitary event distributed across multiple modalities: auditory, visual, vibrotactile, haptic orosensory, and motoric. Our search for a description of this system of linguistic contrasts and multiple sensory projections may eventually explain why the frequency excursions of the second formant combine so readily with the visual presentation of the articulating face.

REFERENCES

Bernstein, L. E., Coulter, D. C., O'Connell, M. P., Eberhardt, S. P., & Demorest, M. E. (1992). Vibrotactile and haptic speech codes. Lecture presented at the Second International Conference on Tactile Aids, Hearing Aids, & Cochlear Implants. Royal Institute of Technology, Stockholm, Sweden, June 9-11, 1992.

Bradlow, A. B., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20, 255-272.

Breeuwer, M., & Plomp, R. (1985). Speechreading supplemented with formant-frequency information from voiced speech. *Journal of the Acoustical Society of America* 77, 314-317.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching divergent sense organs. *Journal of the Acoustical Society of America*, 29, 708-710.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, 38, 269-276.

Julesz, B., & Hirsh, I. J. (1972). Visual and auditory perception: An essay of comparison. In E. E. Denes and P. B. Denes (Eds.), *Human Communication: A Unified View* (pp. 283-340). New York: McGraw-Hill.

Massaro, D. W. (1994). Psychological aspects of speech perception: Implications for research and theory. In M. A. Gernsbacher (Ed.)., *Handbook of Psycholinguistics* (pp. 219-263). New York: Academic Press.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, 101, 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell T. D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lip-reading. *Nature*, 291, 150-152.

Rubin, P. E. (1980). Sinewave synthesis. Internal memorandum, Haskins Laboratories, New Haven, Connecticut.

Saldaña, H. M., Fellowes J. M., Remez, R. E., & Pisoni, D. B. (1996) Audio-visual speech perception without speech cues: A first report. In D. G. Stork and M. E. Hennecke (Eds.), *Speechreading by Man and Machines: Models, Systems and Applications* (pp. 145-151). Berlin: Springer-Verlag.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88, 638-667.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. *Psycholigische Forschung*, 4, 301-350. [Reprinted in translation as "Laws of organization in perceptual forms," in W. D. Ellis (Ed.), *A Sourcebook of Gestalt Psychology* (pp. 71-88). London: Routledge & Kegan Paul, 1938.]