1102

Reprinted from

SPEECH COMMUNICATION

Speech Communication 26 (1998) 65–73

Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances

Robert E. Remez [a,*], Jennifer M. Fellowes [b], David B. Pisoni [c], Winston D. Goh [c], Philip E. Rubin [d]

[a] *Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598, USA*
[b] *Columbia University College of Physicians and Surgeons, 630 West 168th Street, New York, NY 10032, USA*
[c] *Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA*
[d] *Haskins Laboratories and Department of Surgery, Yale University School of Medicine, 270 Crown Street, New Haven, CT 06511, USA*

Received 26 January 1998; received in revised form 12 June 1998; accepted 14 July 1998

ELSEVIER

SPEECH COMMUNICATION

Scope. Speech Communication is an interdisciplinary journal for the development and dissemination of all basic and applied aspects of speech communication processes. Speech Communication features original research work, tutorial and review articles dealing with the theoretical, empirical and practical aspects of this scientific field.

Editorial Policy. The journal's primary objectives are:

- to present a forum for the advancement of human and human-machine speech communication science;
- to stimulate cross-fertilization between different fields of this domain;
- to contribute towards the rapid and wide diffusion of scientifically sound contributions in this domain.

Speech Communication is an interdisciplinary journal whose primary objective is to fulfil the need for the rapid dissemination and thorough discussion of basic and applied research results. In order to establish frameworks to inter-relate results from the various areas of the field, emphasis will be placed on viewpoints and topics of a transdisciplinary nature. The editorial policy and the technical content of the Journal are the responsibility of the Editors and the Institutional Representatives. The Institutional Representatives assist the Editors in the definition and the control of editorial policy as well as in maintaining connections with scientific associations, international congresses and regional events. The Editorial Board contributes towards the gathering of material for publication and assists the Editors in the editorial process.

Subject coverage. Subject areas covered in this journal include:

- Basics of oral communication and dialogue: modelling of production and perception processes; phonetics and phonology; syntax; semantics and pragmatics of speech communication; cognitive aspects.
- Models and tools for language learning: functional organisation and developmental models of human language capabilities; acquisition and rehabilitation of spoken language; speech & hearing defects and aids.
- Speech signal processing: analysis, coding, transmission, enhancement, robustness to noise.
- Models for automatic speech communication: speech recognition; language identification; speaker recognition; speech synthesis; oral dialogue.
- Development and evaluation tools: monolingual and multilingual databases; assessment methodologies; specialised hardware and software packages; field experiments; market development.
- Multimodal human computer interface: using speech I/O in combination with other modalities, e.g., gesture and handwriting.

ELSEVIER

Speech Communication 26 (1998) 65–73

SPEECH COMMUNICATION

Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances

Robert E. Remez [a,*], Jennifer M. Fellowes [b], David B. Pisoni [c], Winston D. Goh [c], Philip E. Rubin [d]

[a] *Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598, USA*
[b] *Columbia University College of Physicians and Surgeons, 630 West 168th Street, New York, NY 10032, USA*
[c] *Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA*
[d] *Haskins Laboratories and Department of Surgery, Yale University School of Medicine, 270 Crown Street, New Haven, CT 06511, USA*

Received 26 January 1998; received in revised form 12 June 1998; accepted 14 July 1998

Abstract

Theoretical and practical motives alike have prompted recent investigations of multimodal speech perception. Theoretically, multimodal studies have extended the conceptualization of perceptual organization beyond the familiar modality-bound accounts deriving from Gestalt psychology. Practically, such investigations have been driven by a need to understand the proficiency of multimodal speech perception using an electrocochlear prosthesis for hearing. In each domain, studies have shown that perceptual organization of speech can occur even when the perceiver's auditory experience departs from natural speech qualities. Accordingly, our research examined auditor-visual multimodal integration of videotaped faces and selected acoustic constituents of speech signals, each realized as a single sinewave tone accompanying a video image of an articulating face. The single tone reproduced the frequency and amplitude of the phonatory cycle or of one of the lower three oral formants. Our results showed a distinct advantage for the condition pairing the video image of the face with a sinewave replicating the second formant, despite its unnatural timbre and its presentation in acoustic isolation from the rest of the speech signal. Perceptual coherence of multimodal speech in these circumstances is established when the two modalities concurrently specify the same underlying phonetic attributes.

Keywords: Multimodal speech perception; Perceptual organization; Auditory-visual speech perception; Intersensory integration; Speechreading; Sinewave speech

1. On the multimodal perceptual organization of speech

How does the perceiver find the speech signal amid an uninterrupted flux of sensory activity? A traditional 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 proposals aim to explain the special circumstances of perceptual organization when the listener also looks at the talker. There is hardly any doubt that multimodal perceptual organization does actually occur, and a

* Corresponding author. Tel.: +1 212 854 4247; e-mail: remez@paradise.barnard.columbia.edu.

0167-6393/98/$ – see front matter
PII: S0167-6393(98)00050-8

small but sturdy literature describes the perceptual phenomena which preclude an explanation relying on elaborate post-perceptual intersensory convergence (for instance, (Green and Miller, 1985)).

Research on sinewave replicas of speech has been singular in promoting an alternative account of phonetic perceptual organization (although, note the reservation expressed by Julesz and Hirsh (1972)). In spite of the fact that a tonal analog of a speech signal is intelligible (Remez et al., 1981), it lacks the typical acoustic manifestations of natural vocal sound production. This fact arguably demonstrates the limitations of Gestalt-derived auditory accounts of perceptual organization, which describe the integration of sensory constituents based on their similarity (Bregman, 1990), and probabilistic accounts of speech perception, which describe the integration of signal elements based on likelihood (Massaro, 1994). Instead, findings with sinewave replicas of speech appear to warrant an account of perceptual organization based on perceptual susceptibility to the unique spectro-temporal characteristics of a phonologically modulated source of sound (Remez et al., 1994). Perceptual organization of speech signals, in this view, occurs by virtue of a perceiver's sensitivity to patterned spectra despite the momentary dissimilarity among the constituents of the pattern, in contrast to the piecemeal assessment of elemental details of the acoustic stream which is central in prior accounts. The perceptual phenomena pertinent to elaborating this alternative view of organization exhibit a common feature, namely, the coherence of sensory elements despite detailed dissimilarity in their physical and psychoacoustic properties.

1.1. Multimodal perceptual organization of speech

Studies of sinewave replicas of speech demonstrate the organizational principle of coherence despite detailed sensory dissimilarity. For example, the resonances of a speech signal change in frequency and amplitude asynchronously, to different extents, and at different rates, and these physical conditions oppose grouping by Gestalt principles. The coherence of aperiodic bursts, aspiration and friction likewise defy the Gestalt grouping principles, due to their physical and auditory dissimilarity from the periodic portions of speech signals (see (Remez et al., 1994), for a detailed exposition). From this perspective, the perceptual organization of speech requires the establishment of coherence among dissimilar sensory elements whether sensation is unimodal and auditory or multimodal and auditor-visual.

Behavioral evidence plainly indicates that visual and auditory inflow in a bimodal case of speech perception are conjoined preliminary to perceptual analysis. A clear case of this phenomenon is seen in a report by Green and 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 as evidence of a kind of context effect. In the bimodal case, though, no perceptual function is readily available to explain the lability of phonetic analysis to a combination of visual and auditory stimulation. The finding of Green and 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 detailed.

Although Welch and Warren (1980) proposed 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 and Ladefoged, 1957; Remez et al., 1994), and the findings of a direct investigation of multimodal integration of spatially disparate visual and auditory sources of phonetic information (Bertelson et al., 1997). In phonetic perceptual organization, there does not appear to be a single simple factor that determines the formation of intermodal coherence.

1.2. The problem of the second formant

Two studies set the question of the present tests of multimodal organization directly. In one, by Breeuwer and Plomp (1985), speechreading was supplemented with pure tones modulated at the frequencies of the first and the second formant.

Subjects transcribed the auditor-visual multimodal conditions relatively poorly, as if the auditory effects of the tone analogs of the formants were barely fused with the visual impressions 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 auditor-visual multimodal presentation.

One clue about the cause of the different results is the different method used in each study to analyze the formant pattern. Breeuwer and 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 used linear prediction analyses with minimal correction to estimate formant values for the voiced portions of the speech signal 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 components were presented in potentially misleading fragments.

In contrast, Bernstein et al. applied the labels F1 and F2 to 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 also 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 et al., 1996). We used single tones from sinewave utterance replicas in combination with a display of an articulating 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 pattern of the phonatory frequency (Fø), and a noise band modulated in amplitude according to the overall energy in the signal. The most effective auditory-visual multimodal presentation combined the second formant analog and the face, occurring here 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 the project of Breeuwer and Plomp (1985), which used uncorrected linear-prediction estimates (compare with (Bosman and Smoorenburg, 1997)).

However, the performance levels in our study were surprisingly 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 can approach performance levels twice as high. The cause, we suspected, was the talker whose speech we sampled, which proved unpredictably difficult for our listeners, a possibility which we verified in the present study by using a new talker.

To conduct a better test of multimodal coherence, we based our audio-visual presentation on the speech of a demonstrably intelligible talker (see (Bradlow et al., 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 visual and

auditory signals were 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.

1.3. A note about procedure

The aim of this research is to determine the multimodal conditions under which the perceptual organization of speech occurs. Accordingly, a direct test of perceptual organization ought to ask a participant in the procedure simply to report instances in which a visual display of a talker and a concurrent acoustic signal seem to cohere. However, recent findings indicate that this method is not entirely suitable for determining the principles governing phonetic integration of multimodal sensory sources. In conditions of temporal discrepancy (Munhall et al., 1996) or spatial discrepancy (Bertelson et al., 1997) of auditory and visual contributions to phonetic perception, perceivers exhibited different standards of perceptual coherence for registering the nature of the event and the phonetic stream within it. This was evident in the performance of subjects who integrated visual and auditory sensory effects in perceiving speech despite the fact that temporal or spatial discrepancies were readily detected. Such findings suggest that the principles governing organization in speech perception diverge from those of the perception of other events, and that the appropriate measure to use in a test of multimodal speech perception must therefore reflect the organizational standards of phonetic perception.

Reliance on phonetic measures of the perceptual organization of speech is a point of general principle, and applies to the extreme case of sinewave replication used here. When a perceiver hears a sinewave sentence, the nonspeech qualities of the sinewave vehicles persist despite the impression that the tones compose a voice articulating an intelligible sentence. In a multimodal setting, the unspeechlike quality of a sinewave signal combined with the visual impression of a real articulating face makes this event irresistibly discrepant despite the phonetic coherence, or so we would claim based on our precedent (Saldaña et al., 1996). In other words, sinewave speech is unnatural albeit intelligible, and no natural talker can be made to appear to the eye to be a source of sinewave speech. Seemingly discrepant with a visual impression of a talking head, these tonal carriers of linguistic properties provide a test of phonetic perceptual coherence despite the incoherence of sound and sight as a perceptual event. The findings that we report here express the functions of phonetic perceptual organization in extreme conditions of stimulation which are nonetheless sufficient to evoke phonetic perception.

2. Method

Test materials. The test materials consisted of unmodified video samples of the shadowless frontal view of the face of a natural talker presented concurrently with sinewave signals derived from the natural speech of that talker. An adult female whose natural speech had been verified as acoustically intelligible produced utterances that were sampled for video and audio reproduction. Ten sentences were selected from the set of Bradlow et al. (1996) and were spoken from a list (see Appendix A).

Video samples were recorded on analog tape and converted to digital records for testing. Tone analogs of formant frequency and the fundamental frequency of phonation were derived from the speech samples by an interactive method described by Remez et al. (1994). Acoustic samples were digitized and the sampled data were analyzed to estimate formant center frequencies and amplitudes. To designate formant centers, an expert phonetician compared discrete Fourier spectra and linear prediction estimates, deriving the synthesis parameters by tracing the formant patterns. The frequency of phonation was estimated similarly for each sentence from a narrow-band Fourier representation of the natural spectra. Frequency and amplitude values taken at 10 ms intervals for Fø, F1, F2, F3 and fricative formants were converted to time-varying sinusoids using a software synthesizer (Rubin, 1980). A complete replica of each of the 10 sentences was composed of tone analogs of the three

oral formants and a fourth tone reproducing fricative formants when these were present and discontinuous with the oral formants. The tone analog of Fø was not used in composing a multitone sentence replica (see (Remez and Rubin, 1984)).

The computed sinusoidal waveforms were combined and synchronized with the video samples using digital editing software, and were presented via microcomputer-controlled video monitors and calibrated headsets to participants in individual testing carrels.

Procedure. A participant in this study was assigned randomly to one of eleven test conditions. The primary tests of the integration of the visual display of the articulating face with single-tone analogs comprised 8 conditions: 4 single-tone presentations that were auditory – ToneFø, ToneF1, ToneF2 and ToneF3 – and 4 audio-visual presentations consisting of a single tone synchronized with the video – Video + ToneFø, Video + ToneF1, Video + ToneF2 and Video + ToneF3. These conditions permitted us to estimate the phonetic effects of multimodal presentations in comparison to the information available in the single tones alone. Three control conditions were also tested. In one – Video – the video displays of the face speaking the test sentences were presented without concurrent acoustic signals. In another – TonesF1F2 F3Ffric – the complete sinewave replica of each sentence was presented for transcription without accompanying video display of the articulating face. In the last – Video + TonesF1F2F3Ffric – the video samples of the talker were presented with the complete sinewave replicas of the test sentences. These three control conditions aimed to assess the information available in the video presentations alone, a control for the single tone multimodal conditions; to assess the information available in the complete sinewave sentence replica, a control to estimate the level of intelligibility of the acoustic materials relative to our experience with sinewave replication; and, last, to determine whether a multimodal presentation of the sinewave sentence replicas improves performance relative to the acoustic presentation of the complete replica.

Each test session consisted of three parts, (1) a familiarization sequence, (2) a test condition of the experiment, and (3) a verification sequence. The familiarization sequence aimed to acclimate the subjects to the unspeechlike timbre of the sinewave sentences. It consisted of a sequence of eight multitone sinewave sentences derived from the speech of one of the authors, and did not duplicate any of the test sentences used in the multimodal conditions. The first three sentences in this test were transcribed for the subjects by the experimenters. Each sentence was presented five times with 10 s between iterations and 20 s between sentence blocks. Subjects transcribed while they listened.

Following the familiarization sequence, a test of one of the eight conditions of multimodal perceptual organization or one of the three control conditions began. Each acoustic or audiovisual sentence was repeated five times with 10 s between successive repetitions and 20 s between sentence blocks. After each repetition, the subject was cued to write a faithful rendition of the message in a specially prepared test booklet. A warning tone 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 condition, the initial set of eight sinewave sentences was presented acoustically once again in a verification test. 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 apparent immunity to the phonetic properties of sinewave signals of a substantial subset of volunteer subjects. Five of the participants in this study were excluded on such grounds (see (Remez et al., 1994)).

Subjects. One hundred and thirty-eight adults drawn from the Indiana University community participated in these tests. Participants reported normal hearing and vision, and no familiarity with tonal analogs of speech signals. Each was a paid volunteer or was a student in Introductory Psychology who received course credit in exchange for participating.

3. Results

Twelve subjects contributed transcription performance for 10 sentences to the eight main

conditions of this experiment (ToneFø, ToneF1, ToneF2, ToneF3 and the four conditions in which these tones were accompanied by the visual display). Of the three remaining conditions, there were 14 subjects in one (Video), 12 in the second (TonesF1F2F3Ffric), and 11 in the third (Video + TonesF1F2F3Ffric). 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 (Remez et al., 1981).

The results of the test conditions in which a single sinewave was used were analyzed in a two-way analysis of variance, with the factors Tone (ToneFø, ToneF1, ToneF2 and ToneF3) and Presentation (acoustic or audiovisual). The finding of a main effect of the factor Presentation reflects the consistency with which the multimodal performance level exceeded the acoustic conditions ($F(1,88) = 386.3$, $p < 0.0005$). A significant interaction of Tone and Presentation ($F(3,88) = 6.31$, $p < 0.001$) was also observed in the analysis, indicating differential effects of the combination of visual and auditory sources. A post-hoc means test (Tukey, $\alpha = 0.05$) revealed that among the audiovisual conditions, the performance level for the condition Video + ToneF2 exceeded the performance level of the other combinations of a single tone analog and the visual presentation of the face.

The group performance in the eight principal test conditions is shown in Fig. 1; error bars portray the confidence interval for the post-hoc means test. 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 for the three other tones.

To analyze the three control conditions, a one-way analysis of variance was performed, finding significant differences in performance level among the silent video, full sentence replicas and the multimodal presentation of the full replicas ($F(2,34) = 56.7$, $p < 0.0005$). Pairwise contrasts were significant, as revealed by a post-hoc means test (Tukey, $\alpha = 0.05$). Table 1 presents the mean performance levels for each of the three groups contributing to this analysis.

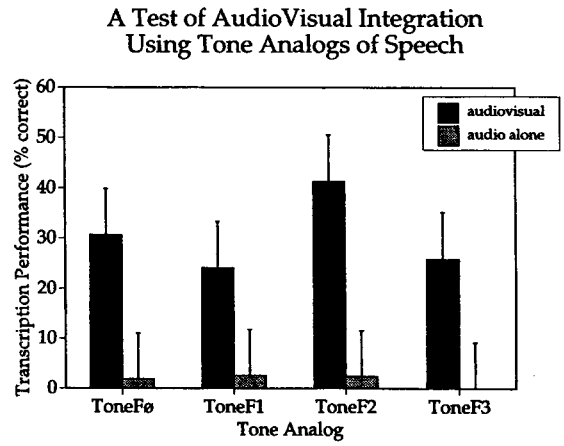

Fig. 1. Results of a test of auditory-visual multimodal speech perception with tone analogs of speech. Each bar shows the group performance with a different signal component. Error bars represent the confidence region for a post-hoc means test (Tukey, $\alpha = 0.05$).

4. 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, the pattern of performance levels here replicated our earlier observation (Saldaña et al., 1996). A tone exhibiting the pattern of F2 was a more effective acoustic accompaniment to the video samples than the tone analogs of the other formants or the fundamental, and our findings show that there was no second best; performance in the three conditions with 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 et al., 1981) to combine readily

Table 1
Results of control tests of auditory-visual multimodal speech perception with tone analogs of speech

Condition	Mean performance	*n*
Video	26.2	14
TonesF1F2F3Ffric	55.0	12
Video + TonesF1F2F3Ffric	84.0	11

with the video samples in evoking an impression of the linguistic message. Why did the results of our test depart from the precedents? Differences in the linguistic test materials are important to consider, because the sentences that we used here may have inadvertently suppressed the differences in effectiveness of tones other than ToneF2. It is also likely that differences in the techniques for estimating the acoustic properties of speech are responsible for the different pattern of results we saw here, in which the second formant alone proved useful in audiovisual presentation. Nonetheless, for multimodal perceptual organization in which the auditory component lacks the timbre of natural speech, it is reasonable now to conclude that the unique effectiveness of the analog of the second formant is established more solidly by these results. Indeed, our results would conflict with the suggestion that any acoustic supplement to vision improves speech perception.

Further, the unique effectiveness of the combination of ToneF2 and the visual impression of the face provides a bit of evidence of the organizational principle at work here. Specifically, it is tempting to speculate that the coherence of visual and auditory components depends on circumstances in which each sensory stream supplements the other. By this premise, multisensory combination produces more complete phonetic information than either component does alone. This explanation would apply had our findings shown that the tone analog of the fundamental frequency combined most effectively with the visual impression of the face, because there would be little in the visual pattern indicating voicing or intonation, and little in the acoustic pattern indicating articulatory place, or voiced stop holds, for instance (compare with (Massaro, 1998)). However, the findings of our study favor an alternative description of the basis for perceptual coherence, specifically, that visual and auditory sources of phonetic information are combined when they manifest concurrent and consistent specification of the phonetic segments. Both the frequency variation of the second formant and the visual appearance of the face provide perceptual information about articulatory place, for instance. Agreement between seen and heard speech promotes fusion, according to this alternative principle of organization.

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 a sentence replica. Transcription performance for the concurrent presentation of all tone analogs 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 evaluated 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 for a listener with normal hearing. Yet, in analogy to the dichotic case, Fig. 1 shows that the concurrent presentation of the articulating face and the tone analog of the second formant allowed listeners to organize a multimodal pattern and to transcribe almost 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 fusion of ToneF2 and the visual display of the face. A clue to perceptual organization may reside in this similarity, though additional tests are required to confirm this. If a synthetic second formant exhibiting natural timbre proves to be more effective multimodally than a tone analog of F2, this would indicate that some organizational functions are contingent on the auditory qualities of the short-term spectrum. Alternatively, sustained exposure to sinewave signals may acclimate subjects to the anomalous timbre of the sinewave voice, and such a procedure may be seen as an improvement in performance due mainly to perceptual tuning.

Third, it is perhaps surprising to note that transcription performance of the multimodal presentation of the full sinewave sentence replica exceeded performance relative to the unimodal acoustic presentation. The integration of visual and auditory modalities provides phonetic information which neither modality apparently does

alone. On the precedent of studies of practiced speechreaders, it is evident that the visual presentation of an articulating face contains more phonetic information than normal hearing listeners typically exploit. It is tempting to speculate that the sinusoidal variation made this visual information more salient to the normal listeners in our study, although it is also likely that the visual impression of the articulating face directed the perceiver's attention to phonetic attributes of tone variation which otherwise were ambiguous or inaccessible. This is a topic deserving more technical and theoretical scrutiny.

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 explain the multimodal case of speech, the organizational principle satisfied by the auditory and visual confluence must be still more abstract. By such means the perceiver treats the sensory pattern as information about a phonetic event distributed across multiple modalities: auditory, visual, vibrotactile, haptic orosensory, and motor. The 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.

Acknowledgements

A preliminary version of this study was described at the Workshop on Audio-Visual Speech Perception in Rhodes, Greece, 27 September 1997. The authors thank: Christian Benoît for encouraging us to undertake this project; Luis Hernandez for applying his ingenuity to our videographic and electroacoustic challenges; Darla Sallee for producing speech clear to the eye and ear; and Dalia Shoretz for lending scholarly and technical assistance to the project. This research was supported by grants from NIDCD (DC00308 to Barnard College, and DC00111 to Indiana University) and NICHD (HD01994 to Haskins Laboratories).

Appendix A

Ten sentences were used in multimodal tests.
Always close the barn door tight.
This is a grand season for hikes on the road.
He ran halfway to the hardware store.
Kick the ball straight and follow through.
The term ended in late June that year.
Use a pencil to write the first draft.
Cut the pie into large parts.
The boy was there when the sun rose.
A cup of sugar makes sweet fudge.
What joy there is in living.

References

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

Bertelson, P., Vroomen, J., de Gelder, B., 1997. Auditory-visual interaction in voice localization and in bimodal speech recognition: The effects of desynchronization. In: Benoît, C., Campbell, R. (Eds.), Proc. ESCA Workshop on Audio-Visual Speech Processing: Cognitive and Computational Approaches. Rhodes, ESCA, pp. 97–100.

Bosman, A.J., Smoorenburg, G.F., 1997. Speechreading supplemented with auditorily presented speech elements in the profoundly hearing impaired. Audiology 36, 29–45.

Bradlow, A.R., 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. J. Acoust. Soc. Amer. 77, 314–317.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge.

Broadbent, D.E., Ladefoged, P., 1957. On the fusion of sounds reaching divergent sense organs. J. Acoust. Soc. Amer. 29, 708–710.

Green, K.P., Miller, J.L., 1985. On the role of visual rate information in phonetic perception. Percept. Psychophys. 38, 269–276.

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

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

Massaro, D.W., 1998. Perceiving Talking Faces. MIT Press, Cambridge, MA.

Munhall, K.G., Gribble, P., Sacco, L., Ward, M., 1996. Temporal constraints on the McGurk effect. Percept. Psychophys. 58, 981–993.

Remez, R.E., Rubin, P.E., 1984. Perception of intonation in sinusoidal sentences. Percept. Psychophys. 35, 429–440.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M., 1994. On the perceptual organization of speech. Psych. Rev. 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: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Man and Machines: Models, Systems and Applications. Springer, Berlin, pp. 145–151.

Welch, R.B., Warren, D.H., 1980. Immediate perceptual response to intersensory discrepancy. Psych. Bull. 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: Ellis, W.D. (Ed.), 1938. A Sourcebook of Gestalt Psychology. Routledge and Kegan Paul, London, pp. 71–88).

Subscription information. *Speech Communication* (ISSN 0167-6393) is published in three volumes (twelve issues) a year. For 1998 Volumes 24–26 are scheduled for publication. Subscription prices are available upon request from the publishers. Subscriptions are accepted on a prepaid basis only and are entered on a calender year basis. Issues are sent by surface mail except to the following countries where air delivery (S.A.L. – Surface Air Lifted) is ensured: Argentina, Australia, Brazil, Canada, China, Hong Kong, India, Israel, Japan, Malaysia, Mexico, New Zealand, Pakistan, Singapore, South Africa, South Korea, Taiwan, Thailand, USA. For the rest of the world, airmail and S.A.L. charges are available upon request. Claims for missing issues will be honoured free of charge within six months after the publication data of the issues.

For full membership information of the Associations, possibly combined with a subscription at a reduced rate, please contact: EURASIP, P.O. Box 134, CH-1000 Lausanne 13, Switzerland; ESCA, ICP-Université Stendhal BP 25X, 38400 Grenoble Cedex, France.

Orders, claims, and product enquiries: please contact the Customer Support Department at the Regional Sales Office nearest you:

New York: Elsevier Science, PO Box 945, New York, NY 10159-0945, USA; phone (+1) (212) 633 3730 [toll free number for North American customers: 1-888-4ES-INFO (437-4636)]; fax: (+1) (212) 633 3680; e-mail: usinfo-f@elsevier.com

Amsterdam: Elsevier Science, PO Box 211, 1000 AE Amsterdam, The Netherlands; phone (+31) 20 4853757; fax: (+31) 20 4853432; e-mail: nlinfo-f@elsevier.nl

Tokyo: Elsevier Science, 9-15 Higashi-Azabu 1-chome, Minato-ku, Tokyo 106, Japan; phone: (+81) (3) 5561 5033; fax: (+81) (3) 5561 5047; e-mail: info@elsevier.co.jp

Singapore: Elsevier Science, No. 1 Temasek Avenue, #17-01 Millenia Tower, Singapore 039192; phone: (+65) 434 3727; fax: (+65) 337 2230; e-mail: asiainfo@elsevier.com.sg

Rio de Janeiro: Elsevier Science, Rua Sete de Setembro 111/16 Andar, 20050-002 Centro, Rio de Janeiro – RJ, Brazil; phone: (+55) (21) 509 5340; fax: (+55) (21) 507 1991; e-mail: elsevier@campus.com.br [Note (Latin America): for orders, claims and help desk information, please contact the Regional Sales Office in New York as listed above]

Advertising information. Advertising orders and enquiries can be sent to: **Europe and ROW:** Rachel Gresle-Farthing, Elsevier Science Ltd., Advertising Department, The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK; phone: (+44) (1865) 843565; fax: (+44) (1865) 843976; e-mail: r.gresle-farthing@elsevier.co.uk. **USA and Canada:** Elsevier Science Inc., Mr Tino DeCarlo, 655 Avenue of the Americas, New York, NY 10010-5107, USA; phone (+1) (212) 633 3815; fax: (+1) (212) 633 3820; e-mail: t.decarlo@elsevier.com. **Japan:** Elsevier Science Japan, Advertising Department, 9-15 Higashi-Azabu 1-chome, Minato-ku, Tokyo 106, Japan; phone: (+81) (3) 5561-5033; fax: (+81) (3) 5561 5047.

US mailing notice – *Speech Communication* (ISSN 0167-6393) is published monthly except in January, February and March by Elsevier Science B.V., Molenwerf 1, Postbus 211, 1000 AE Amsterdam, The Netherlands. Annual subscription price in USA US $784 (subject to change), including air speed delivery. Application to mail at periodicals postage rate is pending at Jamaica, NY 11431.
USA POSTMASTERS: Send address changes to Speech Communication, Publications Expediting, Inc., 200 Meacham Avenue, Elmont, NY 11003. Airfreight and mailing in the USA by Publication Expediting.

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper)

Published twelve times a year

Printed in Netherlands

ELSEVIER SCIENCE

prefers the submission of electronic manuscripts

Electronic manuscripts have the advantage that there is no need for the rekeying of text, thereby avoiding the possibility of introducing errors and resulting in reliable and fast delivery of proofs.

The preferred storage medium is a 5.25 or 3.5 inch disk in MS-DOS format, although other systems are welcome, e.g. Macintosh.

After **final acceptance,** your disk plus one final, printed and exactly matching version (as a printout) should be submitted together to the accepting editor. **It is important that the file on disk and the printout are identical.** Both will then be forwarded by the editor to Elsevier.

Please follow the general instructions on style/arrangement and, in particular, the reference style of this journal as given in "Instructions to Authors."

Please label the disk with your name, the software & hardware used and the name of the file to be processed.

SPEECH COMMUNICATION

☐ Please send me a free sample copy

☐ Please send me subscription information

☐ Please send me Instructions to Authors

Name _____

Address _____

_____

_____

ELSEVIER SCIENCE B.V.

Send this coupon or a photocopy to:

ELSEVIER SCIENCE B.V
Attn: Engineering and Technology Department
P.O. Box 1991, 100 BZ Amsterdam, The Netherlands

c:\sheila\repcovs.94