AUDIO-VISUAL SPEECH PERCEPTION WITHOUT SPEECH CUES

Helena M. Saldaña & David B. Pisoni

Speech Research Laboratory
Indiana University
Bloomington IN 47405

Jennifer M. Fellowes & Robert E. Remez

Psychology Department
Barnard College
New York. NY 10027

ABSTRACT

A series of experiments was conducted in which listeners were presented with audio-visual sentences in a transcription task. The visual components of the stimuli consisted of a male talker's face. The acoustic components consisted of : (1) natural speech (2) envelope-shaped noise which preserved the duration and amplitude of the original speech waveform and (3) various types of sinewave speech signals that followed the formant frequencies of a natural utterance. Sinewave speech is a skeletonized version of a natural utterance which contains frequency and amplitude variation of the formants, but lacks any fine-grained acoustic structure of speech. Intelligibility of the present set of sinewave sentences was relatively low in contrast to previous findings (Remez, Rubin, Pisoni, & Carrell 1981). However, intelligibility was greatly increased when visual information from a talkers face was presented along with the auditory stimuli. Further experiments demonstrated that the intelligibility of single tones increased differentially depending on which formant analog was presented. It was predicted that the increase in intelligibility for the sinewave speech with an added video display would be greater than the gain observed with envelope-shaped noise. This prediction is based on the assumption that the information-bearing phonetic properties of spoken utterances are preserved in the audio+visual sine-wave conditions. This prediction was borne out for the tonal analog of the second formant (T2), but not the tonal analog of the first formant (T1) or third formant (T3), suggesting that the information contained in the T2 analog is relevant for audio-visual integration.

1. INTRODUCTION

1.1. Sinewave Speech

Previous research has demonstrated that listeners can perceive linguistic content in non-speech stimuli consisting of three time-varying sinusoidal signals (Remez, Rubin, Pisoni, & Carrell, 1981) . The sinusoids in these stimuli tracked the center frequencies of the first three formants of a naturally produced sentence. Although the quality of these signals was very unnatural, listeners were able to correctly transcribe the sentence with a high degree of accuracy (Remez et al. 1981). This seminal study shifted the emphasis of theories of speech perception from the momentary spectral attributes that were believed to underlie phonetic perception (the so-called 'speech cues'), to the time-varying or spectro-temporal attributes of speech and it's perceptual organization. The sinewave replicas used in this study retained spectral variation in the absence of short-term spectra typical of vocal sound production. Remez et al argued that listeners were relying on the global time-varying properties of these non-speech patterns to make fine-grained phonetic distinctions.

In the original study, Remez et al presented listeners with every permutation of the sinewave stimuli (tones 1, 2, and 3 individually and in combinations), and demonstrated that individual tones were unintelligible (at less than 5%). Remez et al also observed high intelligibility for the combination of tones 1 and 2, but significantly less intelligibility for the combination tones 1 and 3 and combination tones 2 and 3. They argued that information retained in the sinewave replicas specifies the talkers vocal tract transfer function and how these attributes change over an utterance. The differential findings for the various conditions demonstrates that some portions of the speech signal are more informative than others about conveying the dynamic operations of the vocal tract.

In the years following the publication of the report by Remez et al, numerous studies have been carried out replicating the basic finding that speech perception can take place when traditional speech cues are eliminated from the signal (see Remez, Rubin, Berns, Pardo & Lang, 1994; Remez & Rubin, 1984). The technique of sinusoidal speech synthesis has proven to be extremely useful in generating non-speech patterns that preserve important phonetic properties of speech but don't produce speech like qualities The fact that listeners are able to perceive the phonetic content of the original utterance from these highly impoverished signals provides an "existence proof" that important phonetic information is available in these dynamic nonspeech patterns that are following the changes in the talker's vocal tract. Recent studies have further demonstrated that these patterns not only retain phonetic information but also reliably preserve detailed information about the identity of the specific talker (Remez, Fellowes, & Rubin in press).

2. MULTIMODAL INTEGRATION

It is well understood that the primary modality for spoken communication is audition. However, in noisy environments listeners utilize information from other sensory systems to aid in the recognition and comprehension of speech (Sumby & Pollack, 1954). This finding, in and of itself, is interesting because it suggests to us that the perceptual system takes in all relevant information for a given event and combines it to form one unified percept. As a result of experimental evidence for multimodal integration in speech (McGurk & MacDonald 1976), numerous theories have been proposed for the process of multi-modal integration (see Summerfield, 1981 for a review). However, a close look at these theories reveals that little attention has been paid to defining the type of information that is integrated by the system (Massaro, 1987; MacDonald & McGurk, 1978; however see Grant, Braida, & Renn, 1994).

In the present study, we utilized sinewave synthesis techniques to ask several questions about the nature of the information involved in multimodal integration speech. We believe that the performance of our subjects provides important new clues about the time-varying attributes of auditory signals that are important in perceiving multimodal displays of speech. Empirically, the project was a straightforward attempt to determine the relative effectiveness of several different kinds of acoustic speech signals when combined with a visual display of an articulating face. The video display in this study was conventional: A live subject was videotaped producing a list of English sentences. Although we presumed that some of the morphological and dynamic attributes of a talking face provide information about the linguistic attributes underlying the articulation, we did not manipulate this source of information in this project, we attempted here simply to control those factors in order to focus exclusively on a related auditory question: What kind of auditory attributes permit the perceiver to make use of the visual attributes? In other words: Does the perception of speech (multi- or unimodally) require specific auditory qualities?

The present investigation examined single-tone sinewave replicas (T1, T2, & T3), the pair-wise combination tones 1, 2, & 3, bit-flipped noise, and natural utterances. Previous research has shown intelligibility of single tone sinewave sentences in an auditory-alone condition is below 5% (Remez et al, 1981). We expected that visual information would integrate with these single sinewave replicas leading to better intelligibility than video or audio alone conditions. We also expected to observe different levels of intelligibility depending on which tone was combined with the visual display. This prediction was derived from previous findings which demonstrated different levels of intelligibility for different combinations of tones. If it is the case that some portions of the time-varying signal are better at specifying the dynamic attributes of the vocal tract over an utterance, we would expect to find different levels of integration with a visual display of an articulating face. We also proposed that the increase in intelligibility for the sinewave signals would be higher than any increase found for the bit-flipped noise conditions. This latter hypothesis is based on the assumption that time-varying spectral properties of the acoustic signal are necessary for audio-visual integration, and that these time-varying properties are obliterated in the bit-flipped noise signals, which only preserve the amplitude and duration of the speech envelope.

3. CROSS MODAL INTEGRATION WITH SINEWAVE SPEECH

3.1. Method

Participants. Two-hundred and ninety-six normal-hearing listeners, with no prior experience in speechreading served as participants in the experiment. They were each paid $5.00 for their service. All had normal or corrected vision and reported no history of a speech of hearing impairment at the time of testing.

Stimulus Materials. The speech materials consisted of the following ten sentences :

1. The swan dive was far short of perfect.
2. Where were you a year ago?
3. My dog bingo ran around the wall.
4. A large size in stockings is hard to sell.
5. Kick the ball straight and follow through.
6. The beauty of the view stunned the young boy.
7. Cut the meat into small chunks.
8. Rice is often served in round bowls.
9. The boy was there when the sun rose.
10. My TV. has a twelve inch screen.

The sentences were recorded by a male talker who was instructed to speak in a conversational style. The video image consisted of the talker’s head and part of his neck. The talker wore a black turtleneck and was recorded against a black background. The taped sentences were then digitized on a Macintosh Quadra 950 at a rate of 30 frames per second. The audio track was sampled at 22 kHz using 16 bit-resolution.

Sinewave Synthesis. The auditory portions of the video tape were first low-pass filtered at 4.5 kHz and then sampled at 10 kHz with 12-bit amplitude resolution and stored on a VAX-based computer system. The method of linear predictive coding (LPC) was used to estimate spectra at 5 ms intervals (Markel & Gray 1976). The output was hand checked for erroneous values and corrected when necessary. The formant estimates were then used to drive the output of a formant synthesizer (Rubin, 1980) which calculates the waveforms of signals generated by adding multiple independent audio-frequency oscillators (see Remez et al 1994).

Bit-flipped noise. The digital files of the audio tracks of the sentences were also subjected to a random bit flipping algorithm. The algorithm randomly flipped the sign-bit of 50% of the digital samples in each auditory file. This manipulation resulted in a signal that retained the amplitude envelope and duration of the original waveform but eliminated the fine grain structure of the speech signal.

Audio-Visual Sentences. The audio files were then combined with the corresponding video files using a digitally-controlled video editing package. The synchronization was checked by visually comparing the sinewave audio track with the natural audio track.

A final presentation tape was prepared for each condition. Each sentence was presented five times with a ten second ISI. Following the fifth presentation of the sentence, a prompt on the screen read “Please write your response now”. The prompt remained on the screen for 20 seconds. A timer at the bottom of the screen counted down the time remaining. When five seconds remained two brief tones were sounded.

Experimental conditions. The study materials were presented under twelve test conditions, which included a Video Alone control condition as well as the following:

Audio Alone (AA)	Audio Visual (A+V)
Tone1	Tone1
Tone2	Tone2
Tone3	Tone3
BFN	BFN
T123	T123
Natural	

Procedure. Subjects were run in groups of 1 to 9. They were seated in small experimental classroom with a 31 inch color Phillips 31P460-C402 monitor. The acoustic stimuli were presented using a loudspeaker at a comfortable listening level of approximately 75 dB SPL.

Prior to each test session, subjects were given a set of transcription practice sentences to familiarize them with the stimuli and the task (see Remez et al, 1994). The practice portion consisted of 8 sentences made up of T1+T2+T3. Each sentence was played five times. The first three sentence transcriptions were given to the subjects. The subjects had to try to transcribe the remaining five sentences without feedback. Following the test session, the practice items were again presented to listeners. The listeners were not told that they would be presented with the same 8 sentences again. Listeners were only included in the final data analysis if they could recognize the first three sentences as the first three of the practice session. Two hundred and sixty-two subjects met this selection criterion and were used for the final analysis.

3.2. Results

Figure 1 shows the percentage of correct syllables in the transcription task for the auditory-alone combination tone 1+2+3, audio-visual combination tone 1+2+3, and the visual alone condition. An overall analysis of variance for these three conditions revealed a highly significant effect of stimulus presentation F(2,81)=182.077, p<.001. Post-hoc contrasts revealed that the performance in the audio-visual condition was significantly greater than either the T 1+2+3 auditory alone condition or the video alone control condition F(1,81)=362.45, p<.001.

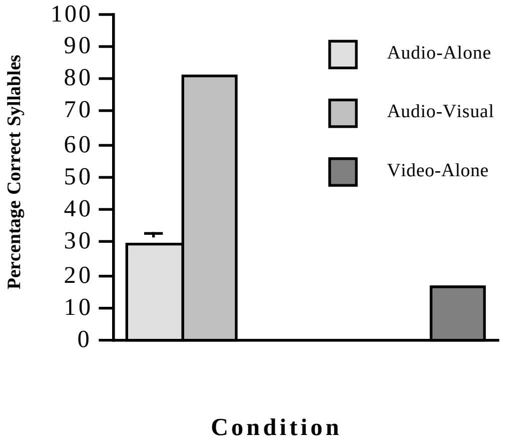

Figure 1. The percentage of correct syllables for the audio-alone and audio-visual combination tones and video-alone condition

Figure 2 shows the percentage of correct syllables transcribed for the auditory alone single tones and bit-flipped noise, the audio-visual single tones and bit-flipped noise, and the visual alone condition. The results from the audio natural condition were excluded from the overall ANOVA because subjects were at ceiling (100%). An overall ANOVA on the audio-alone (A-A) conditions revealed a significant difference in performance F(3,77)=16.370, p<.001. Post- hoc comparisons showed that intelligibility performance for tone 2 was significantly greater than tone 1, tone 3, or BFN F(1,77)=41.885, p<.001.

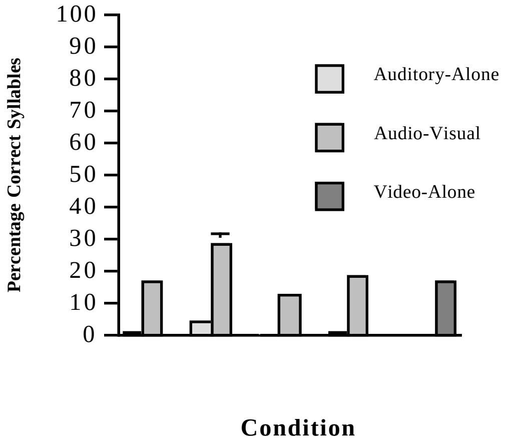

Figure 2 The percentage of correct syllables for the audio-alone and audio-visual single tones, BFN and video-alone condition

An overall analysis on the audio+visual conditions as well as the visual only condition also revealed an overall effect

F(4,113)=7.378, p<.001. Post-hoc contrasts showed that the percentage correct for the audio-visual tone 2 condition was significantly higher than the A-V condition for tones 1, 3, BFN, and video alone F(1, 113) 23.178, p<.001. The contrasts also showed a significant benefit for the tone 2 + video condition over the visual-alone condition F(1, 113) =13.218, p<.001. There were no differences between the visual-alone scores and the audio+visual scores for Tone 1, 2, or bit-flipped noise.

3.3. Discussion

The results from the present investigation suggest that the process of audio-visual integration is super-additive in nature. Intelligibility for the combination tone increased by over 150% when the visual signal was provided to the subject. This is well over the increase we would predict given the participants' performance in the video alone condition. Furthermore, analyses of the single tones and BFN conditions demonstrate that the tonal analog of the second formant (T2) is the most perceptually effective in a multimodal context. No perceptual benefit was observed in the cross-modal conditions when the tonal analog of the first formant (T1), third formant (T3), or envelope shaped noise was presented to the listeners. We offer two possible explanations for the observed increase in intelligibility for T2: (1) the variation of the second formant might provide redundant information to the information contained in the dynamic visual display or (2) the tonal analog of the second formant may add non-redundant or complementary information to the information already available in the dynamic visual display. Additional experiments are underway to determine whether the sinewave signal directs the perceiver's attention to information available in the visual display, or whether the two modalities each contribute independent sources of phonetic information conjointly.

The present results indicate that cross-modal integration in speech perception does not depend on speechlike *auditory qualities*, but rather on speechlike *auditory spectral variation*. It is tempting to speculate that cross-modal integration occurs with these unusual acoustic stimuli because we are presenting formant information. However, the tones presented to our listeners are a far cry from the formant structure of natural speech. They contain no fundamental frequency and no harmonic structure. Why are these very simple patterns integrated as if they are formants and why is there such a large gain in performance when this time-varying auditory information is combined with the dynamic information present in the optical display? This is a question that we plan to address in future research.

We believe the present results raise a number of important new questions about speech perception and spoken language processing. It is clear that coherent variation within and across auditory and visual modalities provides the perceiver with reliable information about a unitary perceptual event that is distributed in both time and space. This unitary multi-modal speech event deserves our continued attention in both the mature perceiver and in development. Any complete theory of speech perception and spoken language processing must begin to take these fundamental facts about perception and perceptual systems into account and must offer a coherent framework for explaining multi-modal perception of linguistic events.

5. ACKNOWLEDGMENTS

This research was supported by NIDCD Research Grant DC-00111-18 to Indiana University in Bloomington, IN and DC-00308 to Robert Remez. We gratefully acknowledge the assistance of Luis Hernandez, Bill Svec, and John Karl in various phases of this project.

6. REFERENCES

1. Grant, K.W., Braida, L.D., Renn, R.J., (1994). "Auditory supplements to speechreading: Combining amplitude envelope cues from different spectral regions of speech". *J. Acoust. Soc. Am., 95*, 1065-1073.

2. MacDonald, J., & McGurk, H. (1978). "Visual influences on speech perception process". *Perc. & Psych.* , 24, 253-257.

3. Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Erlbaum: Hillsdale.

4. McGurk, H., & MacDonald, J. (1976). "Hearing lips and seeing voices". *Nature, 264:*, 46-748.

5. Remez R. E. , Fellowes J.M., & Rubin P.E. (in press). Voice identification based on phonetic information. Journ. Exp. Psych: HPP.

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

7. Remez, R. E., Rubin, P.E. (1984). "Perception of intonation in sinusoidal sentences." *Perc. & Psych. 35*, 429-440.

8. Remez, R. E., Rubin, P.E., Berns, S.E., Pardo, J.S., Lang, J.M. (1994). "On the perceptual organization of speech". *Psych. Rev. 101*, 129-136.

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

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

11. Summerfield, A. Q. (1981). "Some preliminaries to a comprehensive account of audio-visual speech perception". In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 3-51). London: Erlbaum.