1-1-2000

The integration of facial and auditory affect: An emotional McGurk effect?

Teri Jean Forrest
University of Nevada, Las Vegas

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

THE INTEGRATION OF FACIAL AND AUDITORY AFFECT:

AN EMOTIONAL MCGURK EFFECT?

by

Teri Jean Forrest

Bachelor of Arts
University of Nevada, Las Vegas
1998

A thesis submitted in partial fulfillment
of the requirements for the

Master of Arts Degree
Department of Psychology
College of Liberal Arts

Graduate College
University of Nevada, Las Vegas
May 2001

UNLV

Thesis Approval
The Graduate College
University of Nevada, Las Vegas

APRIL 3 , 20 01

The Thesis prepared by

TERI JEAN FORREST

Entitled

THE INTEGRATION OF FACIAL AND AUDITORY AFFECT:

AN EMOTIONAL MCGURK EFFEOT?

is approved in partial fulfillment of the requirements for the degree of

MASTER OF ARTS IN PSYCHOLOGY

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

PR/1017-53/1-00

ABSTRACT

**The Integration of Facial and Auditory Affect:
An Emotional McGurk Effect?**

by

Teri J. Forrest

Dr. Daniel N. Allen, Examination Committee Chair
Assistant Professor of Psychology
University of Nevada, Las Vegas

Research investigating human emotion processing has typically studied either auditory (vocal) or visual (facial) information in isolation. However, speech perception literature supports integration of auditory and visual phonetic information by way of the McGurk effect. This study hypothesized an emotional McGurk effect (i.e., integration of auditory and visual emotional information). To evaluate this hypothesis, emotionally incongruent auditory-visual stimuli (e.g., a joyful voice with a sad face) were developed on a sample of 120 participants. These stimuli were then presented to 30 additional participants who categorized them according to one of eight common emotions. Results indicated significant ($p < .001$) Condition by Emotion interaction effects, suggesting that emotional information from auditory and visual sources is integrated during the perception of auditory emotion. This integration appears to occur in a predictable manner. To understand emotion perception, future studies will need to consider the interaction of auditory and visual emotional information.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

- To Dr. Dan Allen for his invaluable contributions in the areas of leadership, friendship, concern for my mental health (such as it is), SPSS wisdom, and Microsoft Word prowess. I couldn't have done it without you!
- To Dr. Michael Hall for "keeping me busy" and being caring and supportive above and beyond the call of duty.
- To my committee members, Dr. Murray Millar and Dr. Alice Corkill, for their time, their moral support, and for providing such insightful and constructive feedback.
- To Ann Allen for graciously allowing me to borrow her face.
- To my mom and brother for encouraging my academic progress and for never failing to ask about what I did in school today.
- To my friends and classmates for their support and for their commitment to not letting me kill anybody even when I knew that I really needed to.
- To anyone who has listened to more than their fair share of /aba/.

CHAPTER 1

INTRODUCTION

One of the most basic tasks in human communication is the perception of emotion. Research in this area has ranged from the outlining and description of what are thought to be universally interpretable facial expressions (e.g., Ekman, 1993, 1994; Ekman et al., 1987, Tomkins & McCarter, 1964) to delineating acoustic profiles and correlates of vocal and verbal expressions of affect (Banse & Scherer, 1996; Scherer, 1986, 1988; Sobin & Alpert, 1999; Williams & Stevens, 1972). Somewhat less investigated are the interactions between auditory and facial displays of inconsistently conveyed expressions of emotion. Few studies have examined the results of the interaction of facial and nonverbal vocal emotion. Most of the work in this area was conducted more than thirty years ago and attempted to determine whether auditory or visual information was the most heavily relied upon for the interpretation of emotional expressions. Unfortunately, the majority of these studies included spoken words as a verbal channel, which may have inherently conveyed emotional information that could have confounded interpretation by having introduced a spurious source of emotional influence.

However, the interaction between other forms of incongruent visual and vocal information has been investigated. The literature on human perception is replete with

demonstrations of what has been termed the "McGurk effect" (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978). This phenomenon illustrates that visual information heavily influences the interpretation of auditorily presented phonetic information (for example, the synchronous presentation of the visual syllable /ga/ and the vocal syllable /ba/ typically produces the perception of having heard the syllable /da/). This reflects a complex human tendency to intricately integrate discrepant visual and vocal information. It is reasonable to assume that such a phenomenon may also occur within emotional processing. The current study will investigate the hypothesis that incongruent emotional information (sadness and joy) conveyed through vocal (nonverbal) and visual channels and presented in a fashion similar to that employed in standard studies of the McGurk effect will produce integrated emotional responses that are judged to constitute significantly different emotions than either sadness or joy. In order to understand the rationale for such an investigation, research on the interpretation of facial and auditory affect and the integration of emotionally congruent and incongruent information must be thoroughly reviewed. A basic review of the McGurk effect in non-emotional speech perception also is included.

Studies of Facial Affect

Facial expressions of affect represent one of the most heavily and frequently studied areas in human emotion. However, much of the research in this area suffers from methodological shortcomings. Most studies employ an idiosyncratic photographic measure which precludes comparison of results across experiments. These measures have

not been cross-validated with new samples or in different practical contexts. Other drawbacks include differences in the types of discriminations and ratings requested from participants and emotional categories tested in the experimental design (Davitz, 1964). Thus, comparison across studies is often speculative. Nonetheless, a major finding across numerous studies (e.g., Engen, Levy, & Schlosberg, 1958; Gladstones, 1962; Green & Cliff, 1975) is that emotion conveyed by the face can be accurately interpreted across two parsimonious dimensions representing the pleasantness and intensity of the expressed emotion.

Researchers (e.g., Ekman, 1992) have suggested that the deciphering of basic emotions, including happiness, anger, sadness, fear, and disgust, may involve automatic appraisal that lends itself to dealing adaptively with fundamental life-tasks. Some early scientific study in the area suggested that untrained judges of facial affect were generally unable to decipher the intended emotion of a given stimulus (e.g., Woodworth, 1938). However, these researchers noted that errors in the judgment of emotion typically did not produce responses that were grossly inaccurate or representative of a completely different emotion, but were systematic and similar in pleasantness and/or intensity. For example, happiness would never be mistaken for anger, but could be mistaken for pride or love.

Most research suggests that humans do have a specialized ability that aids in the visceral and accurate interpretation of facial expressions of emotion. Child development research has noted that infants as young as 12 days old are able to perceive and mimic the facial expressions of adults (Meltzoff & Moore, 1977). Furthermore, around seven

months of age, infants are able to distinguish between specific facial and auditory representations of emotion (Soken & Pick, 1999), including mixed emotions, such as mock surprise (Ludemann, 1991). When considered in light of studies of the proposed universality of facial affect perception (e.g., Ekman, 1992, 1993) these infant studies suggest that the ability to accurately decipher emotion may be "hardwired" and may lend credibility to the notion of innate, culture-free interpretations of emotion.

Research with adults indicates that this specialized ability to interpret emotion, present in infants, is honed through adulthood. Langfeld (1918) suggested that the tendency to attend to more salient (though perhaps less telling) facial features may contribute to misinterpretation of interpersonal attitudes in everyday interaction. Unfortunately, no numerical, graphical, or statistical data are included in Langfeld's report and he provided no comprehensive summary of the meanings of his findings. However, Langfeld's work strongly suggested a kind of inherent human capacity to decipher emotion and, hence, a standard method for emotion interpretation.

Schlosberg (1954) was one of the first to propose such a method when he proposed a three-dimensional theory of emotion by which any particular emotion could be described in terms of Pleasantness-Unpleasantness, Attention-Rejection, and Sleep-Tension, each along its own continuum of intensity. Subsequently, Engen, Levy, and Schlosberg (1958) developed a set of photographs of female emotional expressions. Over the course of two experiments, approximately 350 undergraduates were asked to judge some or all of the resulting 48 photographs in terms of their respective intensities on each of three continua based on this theory.

Engen and colleagues (1958) found that the judges' ratings were extremely consistent, especially on the Pleasantness-Unpleasantness and Sleep-Tension continua. The lowest interrater correlations were found for the Attention-Rejection dimension, which still had a high median reliability near .85. The highly consistent nature of these findings provide additional evidence for the hypothesis that humans are innately capable of accurately decoding facial emotion in terms of dimensions of pleasantness and intensity. Furthermore, Schlosberg's (1954) three dimensions were, for all intents, cross-validated in these experiments. This implies that regardless of the composition of a sample, participants can be reasonably expected to judge emotions in a consistent manner using general affective descriptors and continua. Engen and colleagues also noted that the variability with which facial expressions were rated was inherently dependent upon the range of stimuli presented.

Abelson and Sermat (1962) noted that early work on emotional interpretation (e.g., Woodworth, 1938) that had revealed erroneous but "rational" emotion deciphering processes could be reconciled with Schlosberg's (1954) theory. They reasoned that, "generally speaking, facial expressions with similar coordinates on Schlosberg's three dimensions are susceptible to confusion with one another, e.g., joy with pleasant surprise, hatred with disgust, etc." (546). To clarify such confusion, the researchers applied a multidimensional scaling procedure to facial affect in order to determine factors vital to the interpretation of emotion. They then attempted to understand the resultant factors through the application of Schlosberg's three-dimensional theory of emotion.

Additionally, the researchers desired to determine whether Schlosberg's theory contained extraneous (or insufficient) factors for emotional interpretation.

Graduate students were instructed to rate the similarity of pairs of photographs. Results indicated that a strong (but unreported) positive relationship existed between the Attention-Rejection and Sleep-Tension dimensions. This suggests that a two-dimensional model of affect recognition may be as suitable and more parsimonious than Schlosberg's (1954) conical model. Abelson and Sermat (1962) also noted that Schlosberg's three-factor theory accounted for 75% of the variance in similarity ratings in their sample (multiple $\underline{R}$=.868). Further analyses revealed that the removal of Attention-Rejection from the multiple regression equation little influenced its predictive power (multiple $\underline{R}$=.854), and that the removal of Sleep-Tension was only slightly more detrimental to prediction (multiple $\underline{R}$=.811). The researchers also undertook an additional exploratory multidimensional scaling procedure. Five factors emerged and were later interpreted based on the Schlosberg model. Dimension I correlated strongly with Pleasant-Unpleasant ($\underline{r}$=.947) and Dimension II correlated with both Attention-Rejection and Sleep-Tension ($\underline{r}$=.878 and .917, respectively).

Thus, Abelson and Sermat (1962) concluded that due to the high positive correlation between the Attention-Rejection and the Sleep-Tension factors, Attention-Rejection could be removed from Schlosberg's (1954) theoretical model without greatly decreasing predictability of emotional interpretation. This would leave Sleep-Tension and Pleasant-Unpleasant as the two sufficient remaining factors of a modified Schlosberg structural model. These findings were also consistent with those of Osgood (1966), who had

developed his own three-dimensional model of emotion. The researchers noted that the only essential dimensions in his model were Pleasantness and Activation.

Gladstones (1962) also undertook a similar multidimensional scaling study in order to verify Schlosberg's (1954) three-dimensional theory. Gladstones instructed undergraduate judges to decide which of two comparison pictures presented more resembled a reference photograph, based solely on similarities between the displayed emotion. Gladstones concluded that no more than three factors were necessary for the interpretation of facial affect. Furthermore, Gladstones' two principal axes correlated highly with Schlosberg's Pleasant-Unpleasant and Sleep-Tension dimensions (r=.97 and .93, respectively). Gladstones hypothesized that as Schlosberg's Attention-Rejection factor had been shown to correlate highly with Sleep-Tension, Attention-Rejection would also correlate highly with his second axis, which it did (r=.82). Therefore, Gladstones concluded that Sleep-Tension was the more fundamental of the two factors.

Gladstones' (1962) third axis correlated only moderately with Attention-Rejection (r=.53), and was therefore considered a largely uninterpretable factor. Based on the physical characteristics displayed in the photographs that placed highly on this factor, Gladstones suggested that it might represent a tenuous dimension of Expressionless-Mobile, suggesting some level of emotional activation. However, it seems that this would likely relate highly to the Sleep-Tension factor, as Gladstones describes the expressions in the photographs along a continuum of "pleased excitement to practically asleep" or blank (p. 99). Gladstones concludes by confirming other findings, including those of

Engen et al. (1958), which noted the interdependence of the Sleep-Tension and Attention-Rejection factors.

Although these studies provide what appears to be robust support for the idea of universal interpretations of facial emotion, through the 1960s there was little standardization in the choice, number, or format of photographs used in studies of facial affect. Paul Ekman and Wallace Friesen (1975) made a significant methodological advance when they set forth what they had identified to be six primary and fundamental emotions (happiness, surprise, fear, anger, sadness, disgust, and neutrality). They then used these emotions as a basis for developing a standardized set of 21 stimuli (1976) for use in facial affect recognition tasks. The set consists of three photographs of different faces for each of the six emotions plus neutrality. Following their development, these photographs have become the standard for use in facial affect research through the present day.

To further address methodological shortcomings of previous research in facial affect, including the use of small, almost exclusively American, samples, Ekman and colleagues (1987) undertook a study encompassing 552 college students from the Estonian S.S.R., Germany, Greece, Hong Kong, Italy, Japan, Scotland, Sumatra, Turkey, and the United States. Their goal was to determine the extent of agreement across cultures in terms of the intensity of emotion displayed in a photograph of a face, what emotion was being displayed when given one choice or up to seven choices of emotion (to account for blends of emotions), and what emotion was perceived as the second strongest in each expression.

Results indicated that across trials, the cultures were agreed on the emotion present in a photograph and its intensity more than 95% of the time. Ekman and colleagues (1987) note limitations to their work, including that as all of the participants were college students, they may have had more access to mass media presentations which could have "taught" them emotional interpretation. This, however, seems unlikely, in that as children, the participants would have been exposed at a much earlier formative age to local and familial influences on emotion perception. In sum, Ekman et al.'s (1987) findings, together with those of many prior studies can be taken as strong evidence for a relatively culture-free, unlearned capacity for both the recognition of facial affect and a small number of fundamental emotions.

Interestingly, one of Langfeld's (1918) participants noticed that the facial features most important to his interpretations of emotion were of the lower part of the face: the lips and jaws were the first features he noticed, followed by nose, cheeks, and eyes. Despite the inclusion of this idiosyncratic data, it was not until the research of Ekman and Friesen that specific muscular and facial actions of emotions were qualitatively explored. Prior to the development of the new set of photographs in 1976, Boucher and Ekman (1975) put forth a concise description of what might be the standard facial movements and gestures associated with these universal expressions. Based on previous components-measurement research (e.g., Ekman, Friesen, & Tomkins, 1971), the researchers hypothesized that the facial area (brow/forehead area, eye/eyelid area, and cheeks/mouth area) most important to the discrimination between emotions would vary on the emotion presented. They predicted that disgust would be best distinguished by

the cheeks/mouth; fear by the eyes/eyelids; sadness by both the brow/forehead and eyes/eyelids; happiness by the cheeks/mouth and eyes/eyelids; anger by cheeks/mouth and brow/forehead; and that surprise would be equally predictable by any of the three regions.

Significant results from separate analyses of variance indicated that for disgust, fear, sadness, and happiness, different parts of the face were indeed more accurate predictors of the intended emotion. As predicted, no main effect was found for surprise, suggesting that all three facial regions contributed significantly to the interpretation of surprise. Contrary to prediction, however, no main effect was found for anger. Hypotheses were supported in that the eye/eyelid area was most predictive of both fear and sadness, and both the eye/eyelid area and cheeks/mouth were predictive of happiness 99% of the time. Furthermore, results across emotions were highly consistent regardless of whether complete or partial faces were viewed by participants, confirming the relative importance of certain facial areas in the differential discrimination of emotions. In Ekman's later work (1992), he notes that the degree of eyebrow and mouth deflection are vital in accurately interpreting positive (e.g., happiness) and negative (e.g., anger) emotions, as well as the intensity of the demonstrated emotion. Massaro and Ellison (1996) report similar findings among a group of Japanese and American participants, in which such deflection greatly influenced decisions about the emotional expressions displayed by a computer-synthesized face. These findings (e.g., Boucher & Ekman, 1975; Ekman and Friesen, 1976; Ekman et al., 1987; Ekman, 1992; Massaro & Ellison, 1996) suggest that the universality of facial affect perception can be linked to facial muscular movements

that are innately understood and universally used to convey standard depictions of particular emotions.

Studies of Auditory Affect

Compared to facial affect expression, the influence of emotional auditory information has been subject to much less inquiry. Scherer (1986) suggests that the late onset of this research may have been due at one time to the inability to store auditory samples prior to the introduction of tape recorders and the cumbersome process of producing and interpreting spectrographic representations of speech. Early attempts at auditory affect research included the use of standard, emotionally neutral statements spoken in various emotional tones or low-pass filtering to mask verbal content while retaining the integrity of other vocal (nonverbal) attributes of the spoken selection, including loudness, pitch, timbre, and rate of speech (Davitz, 1964). However, the research that has been done suggests, similarly to research of facial affect, that vocal expressions of emotion are accurately deciphered along the two continua of pleasantness and intensity.

Davitz and Davitz (1959) attempted to investigate the vocal contribution of content-free speech to emotion perception. It was their contention that emotions are conveyed reliably in a nondiscursive mode by how something is said and less so by the actual semantic content of the statement. The researchers imply a universal method of understanding auditory affect much like that for deciphering facial affect:

> If feelings are communicated in everyday speech with even a moderate degree of efficiency, it seems unlikely that understanding the feelings expressed by another

> person depends entirely upon unique, personal, nonverifiable, and nonduplicable perceptions. Rather, it seems reasonable to assume that within any given speech community there are more or less stereotyped formal aspects of speech associated with the expression of particular feelings (p. 7).

Davitz and Davitz (1959) instructed participants to rate recitations of the alphabet made in each of ten emotional tones (anger, fear, happiness, jealousy, love, nervousness, pride, sadness, satisfaction, and sympathy). Results indicated that although some of the emotions (for example, nervousness, anger, and sadness) were more often judged correctly than others, the frequencies of correct interpretations for all emotions far exceeded chance. They also noted that the frequencies for confusion of certain emotions (especially fear, love, and pride) were far beyond chance. However, the emotions with which they were confused tended to belong to other relatively highly associated psychological constructs (e.g., fear with nervousness and sadness, love with sadness and sympathy, and pride with satisfaction and happiness). Thus, a relatively consistent pattern of errors in emotional identification was revealed. This observation concurred with Woodworth's (1938) description of logical and systematic confusions between psychologically similar emotions. However, this may have been the result of investigating an excessive number of emotional categories, which would likely lead to the presentation of blended or otherwise "psychologically similar" emotions. Such may have also been the case in other research (Havlena, Holbrook, & Lehmann, 1989; Osgood, 1966), in which systematic and frequent confusions were made between conceptually similar emotions.

Williams and Stevens (1972) subjected high-quality recordings of emotional dialogue read by actors, real-life emotional situations, and simulations of similar situations to acoustical analysis. Parameters selected for analysis were chosen on theoretical grounds and focused on fundamental frequency (F_0), which refers to the lowest common frequency relatable to all frequencies being produced at any given time by a speaker. Fundamental frequency, the physical factor most closely related to the perception of vocal pitch, is likely to fluctuate in emotional situations. To this end, it can be influenced by stress, excessive dryness of the mouth or salivation, shortness of breath, or decreased motor control, all of which may be induced by emotional changes. Examples of the dialogues chosen for inclusion were recordings made of a radio announcer present at the scene of the Hindenburg explosion and an actor given a transcript of that announcer's description of the disaster. Various samples of these emotional dialogues were subjected to wide- and narrow-band spectrographic analysis.

Williams and Stevens (1972) concluded that the contour of F_0 vs. time appeared to be the aspect of the speech signal that could provide the clearest vocal indication of emotion. They noted that there is an exceedingly low and narrow range of F_0 associated with sorrow. An increase in utterance duration tended to be associated with a decrease in rate of articulation. The length of duration appeared to result from longer vowel and consonant productions and frequent pauses in speech. Wide-band analysis demonstrated that sorrow was typified by extreme irregularity in volume, as whispered speech was often reduced to noise. On the other hand, neutrality was defined by the absence of noise. Vowels were well defined while consonants, particularly those in unstressed

syllables, were often imprecisely pronounced. Neutral sentences were typically of the shortest durations.

Brighetti and colleagues (1980) attempted to determine the relative contribution of auditory affect to the overall perception of emotion. The researchers wished to determine whether Ekman and Friesen's (1975) seven primary, fundamental emotions (happiness, surprise, fear, anger, sadness, disgust, and additionally contempt) could be deciphered through vocal as well as through facial expressions. The researchers presented 17 male and 17 female university students with depictions of three male and three female actors reading number series in tones conveying each of the seven emotions. Results indicated that the rate of errors made by the participants was far below that which would have been expected by chance. Except for disgust, contempt, and anger, which were correctly identified 58%, 53%, and 78% of the time, respectively, all of the other emotions were correctly identified at least 87% of the time. Furthermore, the researchers' hypothesis that happiness and surprise would be the most easily identified emotions was partially substantiated; in fact, sadness turned out to be the most correctly identified emotion, at 97% accuracy, followed by happiness (89%) and surprise (88%).

Johnson, Emde, Scherer, and Klinnert (1986), in a study theoretically similar to that of Brighetti et al. (1980), noted the importance of determining the contribution of vocal information free of linguistic content and attempted to discover what, other than verbal content, influences the perception of emotion in speech. Participants were instructed to make emotion identifications of samples of an actress reading an unemotional phrase and synthesized voice samples of non-speech stimuli, all of which were intended to convey

anger, sadness, joy, and fear. Tasks included both forced-choice and free-response opportunities.

Results indicated that accuracy in naming the true voice samples was near 100% in the forced-choice portion. Under the free-response option, sadness and anger were correctly identified 91.7% and 97.6% of the time, respectively; however, joy and fear were only correctly identified near 50% of the time. These two emotions were most often misidentified as interest, excitement, or surprise. In terms of the synthesized stimuli, sadness was correctly identified 93% of the time during forced-choice, but the next most often identified emotion, joy, was only identified correctly 65.5% of the time. Anger and fear were correctly identified only 32.1% of the time. As for the free-response portion, only 50% correctly identified sadness, 40% joy, 27% anger, and only 7% fear.

In Johnson et al.'s (1986) second study, conducted one year later, 23 individuals participated in a task similar to the free-response portion of the first experiment. Additionally, those ratings were compared to those they made to the same stimuli after they underwent three acoustic masking procedures. These procedures included: low-pass filtering at 500 Hz, designed to mask voice quality, and loudness; random splicing and reassembly of each stimulus, designed to mask speech pauses and partially mask rhythm, intonation, and tempo; and reverse stimuli, created by playing each original stimulus backward, designed to mask rhythm, intonation, and voice quality.

Results provided support for the claim that the human voice contributes vital emotional information that cannot be replicated by synthesized vocal samples. The intended emotion conveyed by the unaltered human voice was correctly identified 100%

in sadness, 84.8% of the time in joy, and 98.9% in anger. Across the three conditions, the mean accuracies of each emotion were: joy, 75.4%; sadness, 96%, anger, 94.6%, and fear, 40%. As in the first experiment, joy and fear were most often misidentified as interest or excitement. In terms of emotion-specific cues, similar to findings of Williams and Stevens (1972) and Scherer (1986), Johnson et al. (1986) found that sadness seemed to be characterized by a slow rate of speech and a narrow pitch range. In this study, fear could only be associated with a type of arousal, perhaps explaining its frequent misidentification with interest, excitement, and surprise. The same might be said of joy, which has been the target of less research. It is notable that even after being subjected to drastic syntheses, the participants could still correctly identify a true human voice in a free-response format more often than a clear, emotional, voice-synthesized sample. The researchers concluded that voice-synthesized sounds were unable to sufficiently convey vital, emotion-specific cues that would permit accurate emotion recognition.

In a comprehensive review of vocal emotion, Scherer (1986) noted that although study participants could consistently decipher emotion from vocal cues, psychophysical research had not yet identified the defining vocal characteristics that could reliably discriminate between primary emotions. He highlighted this lack of inquiry in light of the fact that definitive facial characteristics had already been largely identified in studies of facial affect. Scherer noted this lack of established vocal correlates of emotion despite the fact that average accuracy in deciphering vocal affect was near 60%, compared to an average chance level of 12%.

Scherer (1986) further suggested a multi-step sequence theory by which the contribution of vocal affect to emotion perception could be quantified (see also Banse and Scherer, 1996). The two steps deemed most important in this theory were novelty and intrinsic pleasantness. Novelty was defined as "whether there is a change in the pattern of external or internal stimulation, particularly whether a novel event occurred or is to be expected" (p. 147). Novelty is related to the orienting response, in which the position of the body and vocal tract may be suddenly and violently altered to orient to a novel stimulus. In turn, this may increase the amplitude and aspiration in speech, related to the preceding deep inhalation and rapid exhalation required. Intrinsic pleasantness was defined as "whether a stimulus event is pleasant, including approach tendencies, or unpleasant, including avoidance tendencies; based on innate feature detectors or on learned associations" (p. 147).

Scherer (1986) proceeded to discuss specific vocal parameters of emotion, which he concluded were hedonic valence, activation, and power. Hedonic valance refers to a pleasantness and approach/avoidance dimension. Vocally, hedonic valence results in either "wide voice" for positive emotions, including enjoyment/happiness and elation/joy, or "narrow voice" for emotions including contempt, sadness, grief, anxiety, fear, irritation, rage, boredom, guilt, and disgust. Activation refers to the arousal or urgency noted in vocal expressions: enjoyment/happiness, and sadness are considered relaxed, while grief/desperation and elation/joy are more tense. Power refers to the "amount of power likely to be perceived by an individual in specific emotional states" and thus high power situations result in a "full voice" and low or no-power situations

result in a "thin voice" (p. 158). Sadness was determined to elicit a thin voice, enjoyment/happiness a slightly full voice, and elation/joy a medium-full voice.

Murray and Arnott (1993), in their review of the literature on vocal affect, note that across a wide range of studies, dimensional models postulated have invariably included factors conceptually identical to Schlosberg's (1954) Attention-Rejection, Pleasantness-Unpleasantness, and Sleep-Tension dimensions. Furthermore, in Scherer's (1986) model, the novelty factor seems highly related to Schlosberg's (1954) Sleep-Tension dimension, in that both imply either quickly attuning to a novel stimulus or remaining in an inert state. Furthermore, Scherer's intrinsic pleasantness step seems similar to both of Schlosberg's Pleasantness-Unpleasantness and Attention-Rejection dimensions, in that a stimulus is evaluated as to its innate attractiveness in both and is, therefore, either approached (given attention) or avoided (rejected). It is interesting that the majority of evidence of the interpretation of vocal affect, as in facial affect research, is highly consistent and largely describable along these dimensions. This suggests that emotion interpretation in both modalities may be largely biologically predisposed and based on quantifiable, physiological muscle actions.

In terms of acoustic qualities, Scherer (1986) noted that joy/elation (a more aroused version of enjoyment/happiness) tends to be consistently defined by increases in mean F_0, F_0 range, F_0 variability, mean intensity, and speech rate, all related to voice tension. Sadness/dejection have been among the most widely studied emotions. Consistent findings have been of decreases in mean F_0, F_0 range, F_0 contour, mean intensity, and speech rate, signifying speech relatively devoid of tension. There is also evidence of

decrease in high-frequency energy and precisely articulated vowel sounds. Although study on grief/desperation has been sparse, the single article cited by Scherer (Costanzo, Markel, & Costanzo, 1969) suggests that these emotions result in an increase in mean F_0.

Murray and Arnott (1993) also concluded that the three major vocal parameters affected by emotion included voice quality, utterance timing, and utterance pitch contour. In terms of specific emotions, they suggest that happiness/joy leads to an increase in pitch and pitch range and note that smiling may increase the fundamental frequency of almost all speakers, while, for some, amplitude and duration may also be increased (see also Williams & Stevens, 1972). Furthermore, sadness is determined by an average or below-average pitch, a narrow range of pitch, and slow tempo, as well as a decrease in speech intensity and a rhythm characterized by irregularly timed pauses. Thus, Murray and Arnott conclude that the level, range, shape, and timing of a particular pitch contour are the most important vocal parameters in differentiating between primary, basic emotions.

Sobin and Alpert (1999) attempted to elucidate the vocal attributes of fear, anger, sadness, and joy. They employed an emotion-induction method for stimulus development, in which many nonactors read a number of emotion-eliciting stories, each containing a standard sentence later extracted for analysis ("It's hard to believe this is real, I can't believe things like this happen"). Twelve women were randomly assigned, three to each emotion, to serve as decoders of that emotion.

For each sentence, 12 acoustic variables related to the encoders' readings of the stories were examined. Through multiple regression analysis, fear was found to be

minimally predicted (multiple R^2=.24) by a decrease in volume, and higher pitch. Anger was moderately predicted (multiple R^2=.45) by increases in volume and pitch variance and decreases in pitch and number of emphasized syllables. Sadness was also moderately predicted (multiple R^2=.51) by a higher duration of speech and lower volume, pitch variance, and number of emphasized syllables. Joy was the most highly predictable emotion (multiple R^2=.72) and was characterized by a decrease in volume variance and correlation of volume and pitch, and increases in pitch variance, number of emphasized syllables, and duration of speech. Although this study examined some different characteristics than were described by Williams and Stevens (1972) or Scherer (1986), most of the general findings were consistent with these prior studies.

Studies of the Relative Contributions of Facial and Auditory Affect

Many studies have been concerned not only with whether specific visual and auditory expressions of emotions can be deciphered in a standardized manner, but also with determining whether the visual or auditory channel is more accurate and efficient in conveying emotion. An early experiment done by Gates (1927) investigated the role of auditory information in children's perception of emotion. She concluded that emotional understanding increased according to age, grade, and intelligence, and that more children correctly identified facial expressions than auditory expressions of emotion. Levitt (1964) also demonstrated the superiority of visual displays of emotion by demonstrating that participants were significantly better at interpreting facial emotion than either vocal emotion or emotion conveyed in a combined auditory-visual condition.

In a later investigation, Zaidel and Mehrabian (1969) proposed that the visual channel would be more dominant in conveying emotion when paired with an auditory channel. Seventy-two undergraduates were instructed to rate 120 photographs and 120 recordings of the neutral words "really" and "maybe" (all conveying various emotions) on a 7-point positivity-negativity scale. Ratings revealed that the visual channel was significantly more effective in the communication of emotion than the auditory channel. Zaidel and Mehrabian explained this by hypothesizing that facial demonstrations of affect may be more effective in communicating affect due to its greater independence from possible confounding influence contributed by any explicit verbal information than is the auditory, nonverbal channel. Furthermore, they suggested that the ability to decipher negative affect, which contributed more to the overall prediction of emotion than did the ability to decipher positive affect, might be due to cultural influence to explicitly verbalize negative affect. Thus, the communication of such socially undesirable information is rendered more subtlety by the nonverbal vocal channel. The nonverbal vocal channel may be more difficult to consciously alter or disguise than information in the verbal and facial channels, and thus may be more relied upon for help in deciphering true affect.

Burns and Beier (1973) attempted to quantify the relative contributions of vocal (sound and pitch) and visual (postural, gestural, and facial expressions) cues to the interpretation of emotion. The researchers developed 36 auditory-visual productions on 16mm sound film of various emotional states, including six each of anger, sadness,

happiness, seductiveness, anxiety, and indifference. Verbal content consisted of any of six sentences chosen as emotionally neutral.

A total of 126 undergraduate students took part in the multiple-choice ratings of the emotional portrayals. Participants were randomly assigned to the auditory-visual group; the filtered auditory-visual group; the auditory group; the visual group; the filtered auditory group; or the content group, which rated only the verbal content. Accuracy scores ranging from 0 to 6 were calculated. Results indicated that, across emotions, the auditory-visual group performed best (4.52), followed by filtered auditory-visual (4.29), visual (4.02), auditory (3.60), filtered auditory (2.35), and content (1.19). An analysis of variance revealed that main effects for emotion and channel were both significant ($p<.01$), as was the interaction between emotion and channel. Happiness appeared easier to judge with only visual cues than with both auditory and visual cues.

Burns and Beier (1973) concluded that the visual and auditory modalities appear to provide independent information for the interpretation of emotion, although the type of information provided and the intensity of each channel varied across emotion. The lack of correlation found between auditory and visual contributions may indicate that people choose one dominant channel to convey important emotional information and that, overall, the visual channel is preferred over the auditory channel to convey information about mood.

Berman, Shulman, and Marwit (1976) attempted to improve upon previous research in the area of vocal affect by focusing primarily upon the reliability of participant responses, rather than the validity of affect-interpretation tasks. Berman and colleagues

noted that reliable decoding of facial and vocal affect had been found to be influenced by stimulus length. Thus, three males and three females read standardized instructions for completing a personality test in either a warm, concerned, and friendly tone, or a cold, aloof, and hostile tone; each stimulus was approximately two minutes in length. Undergraduate raters were then assigned randomly to rate one of the six recordings. Groups were defined by one of three channels (auditory-only, visual-only, and auditory-visual) and one of two types of presentation (warm-cold). Each stimulus was rated on the 11 dimensions of Nowlis' (1964) Mood Adjective Checklist.

Interrater reliabilities were calculated for each of the 11 factors. Aggression, urgency, elation, fatigue, social affection, and vigor were all relatively reliably interpreted across judges. Thus, it was shown that emotions could be reliably rated on scales other than a basic like-dislike or pleasant-unpleasant factor. Furthermore, it was revealed that the auditory channel contributed consistently less information than did either the visual or auditory-visual channels, and the auditory-visual channel proved to not be more reliable than the visual channel alone. Analyses of variance indicated that on every one of the 11 factors, participants were, in fact, able to discriminate between warm and cold presentations.

Thus, the majority of research on auditory-visual emotional presentations suggests that the visual channel is consistently more accurately interpreted than is the vocal channel, or the combined auditory-visual channel, and may suggest that the visual channel may contain more information-bearing cues than does the vocal channel. However, this does not explain the superiority of the visual channel over even the

auditory-visual channel, which would seem to maximize the number of sources from which emotional information can be obtained. Furthermore, as demonstrated by Green and Cliff (1975), auditory-visual presentations of emotion can be described in terms of excitement and pleasantness, similar to auditory- and visual-only presentations, suggesting an innate human ability to combine and decipher emotional information.

Green and Cliff (1975) noted that ratings of both facial expression and tone of voice, when subjected to multidimensional scaling, revealed two strong general dimensions of excitement and pleasantness (excitement for vocal expression being defined as thinness of voice and "height" of pitch). They noted that whereas highly pleasant or unpleasant emotions were rated as high in intensity, emotions rated as low in intensity were not considered either very pleasant or unpleasant. These studies taken together appear to demonstrate the dominance of the visual channel in deciphering auditory-visual communications of emotion. Furthermore, it is interesting that auditory-visual communications, as in studies of facial and auditory affect, can be interpreted along the same two dimensions of pleasantness and intensity. These studies may serve as further evidence of an innate human ability to perceive and decipher emotion.

Studies of Incongruent Facial and Auditory Affect

Although the description of what is perceived when incongruent facial and auditory affective information is simultaneously presented (for example, when a happy face is paired with a sad voice, or vice versa) seems an intriguing area for perception research, surprisingly little study has been conducted in this area. It has been demonstrated that

people naturally interpret facial and auditory affect in standard ways, but it has not been demonstrated what results when natural emotions are presented together in a more unnatural fashion.

Mehrabian and Ferris (1967) were the first to focus on the presentation of incongruent visual and auditory emotions across channels. The researchers hypothesized that a rating of an incongruent facial-vocal communication would most closely equate rating of the facial communication only. Seventeen female undergraduates rated individual visual and auditory stimuli, and 20 others rated the combined visual-auditory stimuli. Stimuli consisted of recordings of a female voice conveying a positive, neutral, or negative attitude while saying "maybe," and photographs of females attempting to convey positive, neutral, or negative facial information. In the congruent combined condition, participants were presented a total of 18 auditory and 18 visual stimuli simultaneously, in attitude-congruent pairs. For the incongruent combined condition, the stimuli were presented simultaneously to all participants, with each of the three visual attitudes being paired with each of the other incongruent auditory attitudes. The pairs were rated on a 7-point positivity-negativity scale.

Results indicated significant main effects for facial and vocal attitude. An investigation of the 95% confidence intervals of the resulting multiple regression coefficients revealed no overlap between the intervals, suggesting that facial attitude produced a significantly greater effect than did vocal attitude. Mehrabian and Ferris (1967) concluded that facial and vocal channels did not interact and that the visual channel was substantially stronger than the auditory channel. The auditory channel still

provided significant information, but at a level about two-thirds the strength of the visual channel. The researchers suggested that providing congruent and, therefore, redundant emotional information in two channels would serve to magnify the emotion being conveyed moreso than when either channel was presented in isolation. However, in the case of incongruent information, the visual channel would predominate.

However, Mehrabian and Ferris' (1967) results were based on tape recordings and photographs presented in a spatially incongruent fashion. Bugental, Kaswan, and Love (1970) moved forward technologically to incorporate videotape into their study of incongruent verbal, nonverbal, and facial emotional interactions. Thirty-two messages were videotaped, containing eight primary messages, varying in terms of whether a positive or negative message, picture, and vocal tone was presented. Participants included 80 children (ages 5-12) and 80 parents. Each participant responded to the four videos for one of the eight primary messages. All participants provided ratings on a 13-point positivity-negativity scale.

Results indicated that children respond with more negativity than adults to joking messages (those containing positive face or vocal tone and negative message). Further, women were generally rated more negatively than men, suggesting a widespread tendency to discount any positive emotion from a woman if any negative components were perceived. Facial expression accounted for almost twice as much variance in ratings than did either message or tone, suggesting that facial expression heavily influenced the interpretation of incongruent messages. Bugental and colleagues (1970) also reported a tendency for the vocal channel to dominate spoken words, but only when the vocal tone

was negative. Through the advent of videotape, which allows for temporally and spatially congruent displays of incongruent attitude, Bugental and colleagues demonstrated that vocal and visual content do indeed influence each other significantly, though the visual channel continued to dominate in conflicting situations.

Argyle and colleagues (1970) investigated the interaction of incongruent emotional information, but in regard to a single major dimension, superiority-inferiority. The researchers undertook two experiments, encompassing 80 participants. Stimuli consisted of three 20-second videotaped statements about participation in psychological experiments, each conveying either a superior, neutral, or inferior message. The nonverbal messages were conveyed by the presence or absence of a smile, vocal tone, and head positioning. Participants rated each performance on ten 7-point scales.

Analyses of variance revealed that the semantic content of the stimuli significantly influenced scores on the Hostile-Friendly, Stable-Unstable, Confusing-Straightforward, Inferior-Superior, Sincere-Insincere, and Submissive-Dominant scales, three at the $p<.001$ level. However, nonverbal content significantly affected scores on all ten scales, nine at the $p<.001$ level. The researchers highlighted this substantially greater impact of the nonverbal channel, noting that the total ratio of nonverbal:verbal variance was 21.7:1, and for the Inferior-Superior dimension alone the ratio of variance was 12.7:1. In sum, the combined nonverbal and verbal cues were approximately 4.3 times as strong as the verbal cues alone. Verbal cues seemed to serve predominately to emphasize the nonverbal communication, but were rendered useless when in conflict with the more powerful nonverbal cues. Argyle and colleagues suggested that the substantially greater

communicative power of nonverbal signals might be due to an intrinsic and universal human ability, much like that evidenced in animals, to recognize and interpret such cues in interpersonal interactions. The researchers also suggested that spoken information may be reserved for conveying other types of information that are unrelated to immediate social interactions, since dialogues concerning the immediacy of interpersonal situations are relatively rare.

Argyle and colleagues (1971) continued their research into incongruent emotional communications by using a similar experimental methodology to examine another major dimension of interpersonal communication, friendliness-hostility. Similar to the prior (1970) experiment, videotaped presentations were created in which actors combined friendly, neutral, and hostile verbal messages about participants in psychological experiments with each of three corresponding nonverbal displays. Thirty education students rated the nine performances on six 7-point scales.

Results indicated that although the nonverbal channel significantly influenced the verbal channel on all three verbal messages, the verbal channel was often dependent on the nature of the nonverbal message with which it was paired. Specifically, no verbal message could override the effects of a hostile nonverbal display. In sum, the nonverbal channel accounted for more than 12.5 times the variance than did the verbal, and in those cases where there was a clear inconsistency between the emotion portrayed by the two channels, the nonverbal was invariably favored. Furthermore, all nonverbal hostile conditions were ranked as more unstable and less sincere than friendly or neutral nonverbal cues, and friendly verbal conditions were widely considered less sincere than

either hostile or neutral messages. Overall, incongruent presentations were considered significantly less stable than congruent conditions ($p<.05$). The hostile verbal-friendly nonverbal was considered significantly more confusing than congruent conditions and the friendly verbal-hostile nonverbal stimulus was rated significantly less sincere than the congruent conditions (both p's$<.05$).

In a second experiment, Argyle and colleagues (1971) adjusted the verbal messages to be stronger than nonverbal conditions and omitted the neutral verbal and nonverbal channels. Thus, four videotapes were made and shown to student participants. Nonverbal cues were again shown to be more effective at conveying emotion on all levels, and again were able to overcome any verbal message when hostility was portrayed through nonverbal means. In these conditions, the nonverbal channel accounted for 1.67 times the variance that the verbal channel did, although the verbal channel had been specifically adjusted to attempt to match the nonverbal channel in terms of emotional intensity. In sum, the researchers concluded that although nonverbal channels seem to invariably dominate emotion perception, especially when cues are incongruent, the strength of the cue also determines whether or not a particular channel will be attended to.

Both of these (1971) experiments demonstrated that the participants appeared to have a standard method for interpreting incongruent emotions: to do so, the truth of the less intense or relevant cue is discarded. Furthermore, when hostile messages are spoken in a friendly manner, one is confused; when a friendly message is spoken in a hostile manner, one is struck by the insincerity of the remark. These findings may suggest a

natural tendency toward evaluating incongruent emotional communications, in that the channel conveying the most dominant information dominates emotional interpretation, although all channels of information are attended to and integrated to some degree.

DePaulo and colleagues (1978) developed a videotaped Nonverbal Discrepancy Test to examine the interaction of incongruent vocal and visual cues. The researchers suggested that although the visual channel had so often been found to be dominant over the vocal, perhaps the vocal channel would gain precedence as the degree of inconsistency increased. In other words, when the message seemed more deceptive, judges would begin to rely more heavily on vocal nonverbal information that were deemed more difficult to consciously alter than facial expression. Two-second videotaped segments of either the face or the body of a female were combined with a filtered auditory channel such that only nonverbal vocal information remained intelligible. For each stimulus, the auditory channel conveyed one of the four possible combinations of positivity-negativity and dominance-submission.

In the first study, the researchers compared auditory-visual interpretations to single-channel auditory and visual judgments. Each stimulus was judged on three 9-point rating scales. In terms of both the dominant-submissive dimension and the positivity-negativity dimension, the auditory-visual ratings more closely resembled visual-only rankings. In all, seven of the eight possible interactions (Condition x Channel x Dimension), the video modality was significantly more influential than information obtained in the auditory channel.

In the second experiment, the auditory-visual presentations were presented to 318 participants (approximately 56% female) across junior high school, high school, and college-aged samples. All three samples invariably displayed strong video primacy ($p<.0001$). Overall, the video channel appeared to convey more information for the body than for the face and for the positivity rather than the dominance dimension, positivity being judged more by the face and dominance more by the body. However, in terms of very incongruent presentations (e.g., positive-dominant auditory and negative-submissive visual), there was less demonstrated visual dominance across samples than for less incongruent presentations, especially when the visual channel centered on the face. This confirmed DePaulo and colleagues' (1978) hypothesis that since it is easier to consciously alter facial expression, as communications became more discrepant, nonverbal auditory information would become more important to accurate interpretation of the presentation.

Hess, Scherer, and Kappas (1988) focused on the importance of facial displays of emotion in their research on the interactive nature of visual and vocal information in deciphering emotional communications. The researchers created one-minute videotaped auditory-visual presentations of a semantically neutral one-sided telephone call. Videos were created, by dubbing independently created auditory and visual tracks, to contain all possible combinations of positive (happy) and negative (angry) facial expressions, vocal quality (characteristics of vocal tract resonance and phonation), and intonation (characteristics of the pitch contour of an utterance). Results indicated that for the emotionally congruent videotapes, information contained in the visual channel was

significantly more influential on participants' ratings than either vocal channel. Furthermore, the concordant positive auditory stimuli (consisting of both positive vocal quality and intonation) was rated as significantly more positive than the concordant negative auditory stimuli were considered negative. Conversely, the positive video channels were rated as significantly less positive than the negative video channels.

In terms of incongruent presentations, the combination of the positive visual channel and the negative auditory channels (thereby incorporating examples of the channels rated as weakest) were rated as more positive and less negative than the counterpart combination of the "strong" negative visual channel and positive auditory channels. This demonstrates the dominance of the visual channel with stimuli conveying less discrepant, more "neutral" information. Based on a three-dimensional model of emotion developed by Osgood (1966), containing dimensions of pleasantness, activity, and control, the positive visual-negative auditory ratings indicated pleasantness judgments closer to those of the video channel, while judgments were closer to those of the auditory channels in terms of activity. On the other hand, the results of the combination of the "strong" negative visual channel and the positive auditory channels indicate ratings more similar to the auditory channel. This demonstrates participants' tendency to rely more heavily on the auditory channel as presentations become increasingly discrepant and deception is suspected, as was shown by DePaulo and coworkers (1978). The combination of the negative visual channel and the positive auditory channels was also rated as significantly more excited, secure, impolite, unfriendly, disapproving, and bored on six bipolar attitude scales than the positive visual-negative auditory combination.

Massaro and Egan (1996) undertook a similar study to Hess et al.'s (1988). Presentations consisted of all possible combinations of happy, angry, and neutral computer-generated faces with happy, angry, and neutral vocal recordings of an emotionally neutral word ("please") by a male amateur actor. During auditory-only, visual-only, and auditory-visual conditions, participants rated the appropriate presentations with a two-alternative forced-choice scale (happy or angry). They found that as the information portrayed in one channel became more ambiguous, the influence of the other channel increased in deciphering the depicted emotion. However, it was noted that the visual channel was influential to some extent across all three levels of vocal affect and thus had a larger overall influence on judgments. Furthermore, neutral presentations were consistently rated as happier rather than angrier.

Massaro and Egan (1996) also noted that participants' reaction time in rating the presentations was significantly faster when the presentations were relatively unambiguous. In sum, "decision time increases as the degree of support for one alternative becomes more similar to the degree of support for the other alternative" (p. 220). The researchers advise that neither instructions nor desire nor intention to answer in a predetermined fashion can preclude the perception of an integrative influence of facial and vocal information.

Massaro (1998) undertook a number of experiments to replicate and extend his previous work (i.e., Massaro & Egan, 1996). The first was created to be quite similar to that described in Massaro and Egan (1996), but included four emotions (happy, angry, fearful, surprised). All four emotions were presented auditorily, visually, and bimodally,

combined in a fully factorial design. The computer-generated faces were designed to hold a static emotional expression for the entire length of the stimulus. Participants were asked to determine what global emotion was being portrayed and were given the four choices (happy, angry, fearful, or surprised) with which to respond. Results indicated that the proportion of correct responses (that is, those matching the expressed auditory emotion) was found to be significantly higher in auditory, visual, and congruent bimodal conditions than in incongruent bimodal conditions, suggesting a strong mutual influence of one channel upon the other. Massaro also noted that anger and happiness appeared to be easier to judge visually and surprise and fear appeared easier to judge given auditory information. In cases when a weaker visual emotion (surprise or fear) or a weaker auditory emotion (happiness or anger) were paired with an incongruent stronger visual or auditory emotion, the stronger emotion would invariably dominate.

In a follow-up study, Massaro (1998) conducted the same investigation, but provided participants with six emotion choices (those given in the above study plus sad and disgusted). Results indicated that provision of additional response choices significantly lowered performance for the isolated auditory condition and performance on the incongruent bimodal condition was again worse than in the isolated auditory and visual conditions. Confusion matrices revealed that a fearful auditory channel was twice as likely to be identified as sad and surprised and fearful visual channels tended to be confused with one another. The "other" response appeared to be rarely utilized.

Massaro (1998) also conducted experiments designed to determine the effect of instructions upon integration of bimodal emotional information. He created three types

of instructions to guide participant intentions. These instructions consisted of the direction, while watching the face and listening to the auditory channel, to make judgments based only on what was heard, on only what was seen, or to integrate the two channels and provide a global rating (bimodal). Given bimodal instructions, facial information was found to exert much more influence than auditory information. Auditory instructions resulted in a much stronger influence of the vocal channel. Visual instructions did not change to a large degree, as the face had already been shown to exert more influence in the bimodal condition. However, the proportion of correct responding was invariably considerably lower in incongruent bimodal conditions. Thus, Massaro concluded that although instruction sets did appear to have a substantial impact on participants' intentions to respond, instructions were never enough to completely prevent the influence of the opposing channel participants were instructed to be ignored.

de Gelder and Vroomen (2000) essentially replicated the above findings in three studies conceptually similar to those of Massaro and colleagues (Massaro & Egan, 1996; Massaro, 1998). The first study investigated the integration of two sources of emotional information. These sources consisted of a still black-and-white photograph of a face taken from a continuum extending between extreme examples of happiness and sadness and the vocal presentation of a sentence ("his girlfriend came by plane") in either a happy or sad tone. Three types of trials were included and consisted of visual-only, auditory-only, and bimodal presentations; in the bimodal presentations, the vocal channel was presented and one of the 11 possible photographs appeared on the computer screen only during the last word. On all three types of trials, participants were

asked to integrate the two sources of information and determine whether the person was happy or sad. Results indicated that vocal information did influence global labeling of the presentations by reducing the proportion of "sad" responses when a happy voice was presented with sad photographs.

de Gelder and Vroomen's (2000) second study was identical except that participants were instructed to judge only the visual channel and ignore vocal information. This was done in order to determine whether the results of the previous study were influenced by the instruction to provide a combination response. Results were similar, and further indicated that the effect of the vocal information was stronger at the midpoint of the visual continuum, where the facial information provided was more ambiguous. In sum, however, it was noted that the effect of the vocal information was smaller in the second compared to the first study, perhaps signaling bias inherent in the instructions. However, the study did suggest that regardless of instruction, participants were unable to fully ignore the information provided in the auditory channel.

The third study involved identical methodology but opposite stimulus construction and instructions than the second experiment. Participants were instructed to judge only a voice taken from a continuum of vocal tones and to ignore whichever of two faces was presented simultaneously. Thus, this experiment included different vocal information than the previous two experiments in that the researchers were unable to create a continuum of happy and sad voices and thus used a continuum of happiness and fear which they decided was easier to obtain. Participants were asked to rate the person presented as either happy or fearful. Results obtained were quite similar to those

obtained in the second study, in that it was impossible for the participants to fully ignore the visual channel, which they had been instructed to ignore. Potential limitations of this study include the limited ecological validity of their presentations and the fact that their research had made no effort to investigate more complex or blended emotions and their potential effects on the integration of emotionally incongruent information.

The above studies, taken as a whole, imply that when emotionally incongruent vocal and facial information are presented simultaneously, patterns of results differ from when emotionally congruent information is presented. Specifically, whereas the visual channel tends to dominate the interpretation of emotionally congruent demonstrations, the nonverbal auditory channel becomes more influential as the perceived discrepancy between the auditory and visual channels increases. These results suggest a tendency to rely on information that is less likely to be successfully consciously altered in order to make the most accurate judgment possible of emotional state.

Studies of Incongruent Facial and Auditory Phonetic Presentations

Emotional perception constitutes only one area of perception in which researchers have investigated the integration of incongruent visual and vocal information. McGurk and MacDonald (1976) demonstrated a visual influence on auditory speech perception. In this phenomenon, for example, the synchronous presentation of the visual syllable /ba/ and the vocal syllable /ga/ produced the perception of having heard the syllable /da/. This phenomenon also occurred with other syllables, including /ka/ and /pa/, which together produce the illusion of having heard the syllable /ta/. This phenomenon results

from the perceptual integration of bimodal information containing pronunciation of consonants differing in their place of articulation (that is, those involving the use of lips [labial, such as /b/] versus speech not involving the use of lips [nonlabial, such as /g/]).

In the initial experiment leading to the report of this finding, McGurk and MacDonald (1976) tested 21 children aged three to four years, 28 children aged seven to eight years, and 54 adults aged 18 to 40 years. The participants were tested individually in two conditions: during the auditory-visual portion, participants were instructed to report the syllable they had heard; similar instructions applied to an auditory-only condition, during which participants were faced away from the screen on which the auditory-visual presentations were displayed. The four test presentations included video recordings of a woman pronouncing each of the four syllables. The stimuli were edited to form new combinations of the following nature: vocal /ba/ and visual /ga/; vocal /ga/ and visual /ba/; vocal /pa/ and visual /ka/; and vocal /ka/ and visual /pa/.

Results indicated that the auditory-only condition yielded high accuracy across participants, with the preschool, primary school, and adult samples achieving accuracy levels of 91%, 97%, and 99%, respectively. However, under the auditory-visual condition, the average accuracy rates for the three groups were 41%, 48%, and 8%, respectively. The effect appeared to be strongest with the /ba/-/ga/ combinations.

McGurk and MacDonald (1976) noted that "where responses are dominated by a single modality, this tends to be the auditory for children and the visual for adults" (p. 747). In sum, the findings indicated that this perceptual illusion, later dubbed the

"McGurk effect," was highly replicable and observable across wide age spans. They also pointed out the interesting fact that

> we ourselves have experienced these effects on many hundreds of trials; they do not habituate over time, despite objective knowledge of the illusion involved. By merely closing the eyes, a previously heard [da] becomes [ba] only to revert to [da] when the eyes are open again (p. 747).

On a final note, McGurk and MacDonald (1976) noted that some participants did not show a bias for either the auditory or visual channel, in which case the participant was left to oscillate between modalities, sometimes hearing different combinations of the syllables, such as /babga/ or /papka/.

MacDonald and McGurk (1978) extended this preliminary research by putting forth a manner-place hypothesis to explain the findings of their previous research. Their hypothesis proposed that in normal conversation, manner of consonant articulation is detected auditorily, while the place of articulation is judged visually. Thus, "at an as yet unknown level of processing, information from the two sources is combined and synthesized, resulting in the 'auditory' perception of a best fit solution" (p. 254). Results of this experiment indicated that the auditory syllables /da, ta, ga, ka, na/, when paired with any visual syllable in the same range, produced a non-illusory perception of the auditory component. The same held for the "labial" syllables /ba, pa, ma/. This supported the manner-place hypothesis in that consonants originating from the same place of articulation can be interchangeably interpreted and that the auditory channel dominates.

However, more interesting are the results of integrations of consonant syllables not originating from the same place of articulation. When auditory labial sounds were combined with non-labial lip movements, accuracy rates for judgments of the information conveyed by the auditory channel averaged 27%, with a range of 0% to 70%. When, alternately, auditory non-labial sounds were combined with labial lip movements, accuracy rates averaged 75%, with a range of 25% to 100% ($p<.05$). Although the errors resulting in the first labial-nonlabial combinations revealed concordance with manner-place theory predictions, the nonlabial-labial errors did not; errors here appeared to suggest a larger visual contribution in that labial consonants were more likely to be reported. Thus, MacDonald and McGurk (1978) suggested the satisfactory fit of their labial-nonlabial data with the manner-place theory and the slightly less satisfactory fit of the nonlabial-labial data.

The McGurk effect has been demonstrated to be a robust perceptual phenomenon in Western cultures, occurring equally and as often in adults as in prelinguistic English-exposed infants as young as five months old (Rosenblum, Schmuckler, & Johnson, 1997), suggesting that the phenomenon may be largely unlearned and "hardwired". The McGurk effect has also been investigated over a range of languages and cultures. Native speakers of Finnish showed a strong McGurk effect when exposed to Finnish syllables, words, and words presented in sentence context (Sams, Manninen, Surakka, Helin, & Kaettoe, 1998). However, results of studies involving nonnative speakers of American English have been more equivocal. Research has suggested that native speakers of Japanese (Sekiyama & Tohkura, 1991) and Chinese (Sekiyama, 1997a) are less

susceptible to making errors of the McGurk type whether exposed to English or Japanese stimuli. This could be due to Asian cultural aversions to observing the lip and mouth movements of others. Familiarity with the language in which the McGurk fusion syllables are presented may be the most influential factor in determining the probability of a participant's susceptibility to the effect: the importance of familiarity has been noted by several researchers, including Hardison (1996) and Sekiyama (1997b).

Other researchers have attempted to demonstrate McGurk effect-like responses in speech stimuli manipulated in other fashions, including auditory and visual differing in terms of conveyed vowel information (Green & Gerdman, 1995) and gender information (Green, Kuhl, Meltzoff, & Stevens, 1991). These results suggest that information gleaned from both channels about specific vocal characteristics are normalized even before the actual semantic information is integrated and normalized. Thus, these vocal characteristics are "compromised" before phonetic information is assessed and are, therefore, unable to interfere with the later phonetic processing stage.

Study of the McGurk effect has not only been limited to speech research. Other studies (e.g., Saldaña and Rosenblum, 1993) have noted that the integration of visual and vocal information can be observed in the integration of visual and non-speech auditory information (e.g., integrating auditory and visual plucks and bows on a cello). Although the McGurk effect demonstrated in this study was not as strong for non-speech stimuli as for speech stimuli, it was nevertheless observable. The fact that the finding was not as significant in non-speech stimuli may be due to several factors, including the use of

different, non-speech stimuli or the fact that the participants had not been identified as having any musical experience.

However, it is likely that humans have more occasion to need to integrate speech modalities than to integrate musical characteristics. Such occasions may arise in order to interpret potential threats or deceptions or to gather information conveyed on a nonverbal level through tone of voice, facial expression, or body posturing (as suggested by DePaulo and colleagues [1978]). These types of needs are likely to be innate and universal, as they are crucial tasks for interpreting one's surroundings and for avoiding dangerous situations. Thus, the integration of emotional information is nearly as important in interpersonal communication as is the deciphering of phonetic information. A greater emotional McGurk effect (that is, the unconscious fusion of emotionally discrepant vocal and facial information) may be found than when this type of integration is investigated with other, more obscure kinds of non-speech information (such as music).

It seems a logical extension of this argument that if phonetic speech information is integrated and normalized at an unconscious level, emotional speech information likely is as well. At some basic level, it appears that there is an innate human ability to decipher facial and auditory affect in a standardized fashion. Furthermore, the McGurk effect demonstrates an unconscious and uncontrollable predisposition to integrate speech information. Studies of emotion (Hess, Scherer, & Kappas, 1988; Massaro & Egan, 1996; Massaro, 1998; de Gelder & Vroomen, 2000) have shown that participant responses differ when they are presented with emotionally incongruent information than

with emotionally congruent information. However, none of these studies have used stimuli that are ecologically valid (that is, stimuli that are realistic, naturally produced, and dynamic) and edited in a manner similar to that employed in standard studies of the McGurk effect. Although some studies (e.g., de Gelder & Vroomen, 2000) claim that their results support the idea of an emotional McGurk effect, their stimuli have not been produced or edited in a way that permits comparison with speech studies concerning the McGurk effect.

The current study aims to produce natural, ecologically valid videotaped stimuli conveying joy and sadness. The stimuli will then be edited in a manner similar to the paradigm employed by McGurk and MacDonald (1976, 1978). Computer editing software will be used to create stimuli consisting of isolated auditory-only or visual-only information as well as emotionally incongruent auditory-visual stimuli (that is, the auditory channel will convey either joy or sadness, while the visual channel will provide information conveying the opposite emotion). Emotionally congruent auditory-visual productions (that is, stimuli in which the two channels convey the same emotion) will also be produced and differences between participant ratings of the four types of stimuli (auditory-only, visual-only, emotionally congruent auditory-visual, and emotionally incongruent auditory-visual) will be investigated. Specifically, the hypotheses put forth in this study are as follows:

1. Auditory-visual presentations of incongruent emotional information (sadness and joy), conveyed through vocal (nonverbal) and visual channels, will produce

integrated emotional responses that are judged to constitute significantly different emotions than either sadness or joy.

2. As the emotionally incongruent auditory-visual presentations become more incongruous in terms of each channel's emotional intensity, the auditory channel will be more heavily relied upon to determine the resulting emotional production.
3. However, as the combined presentations are more similar in terms of their emotional intensity, the visual channel will be predominant in the deciphering of the resulting emotional production.

CHAPTER 2

METHOD

Participants

Participants included 150 University of Nevada, Las Vegas, undergraduate students (ranging in age from 18 to 45 with no reported hearing deficits) recruited from the Psychology Department Subject Pool during the Spring, Summer, and Fall 2000 semesters by way of voluntary sign-up for research credit. This research credit served as partial fulfillment of course requirements or as course-specific extra credit. Recruiting volunteers in this manner provided equal opportunity to obtain both male and female participants and thus helped to assure a relatively unbiased and random sample. All participants provided fully informed written consent prior to participating in any of the tasks. Participants were 65% female and 56% Caucasian. Participation was limited to those whose native language was American English. Ten volunteers were recruited for each of the six auditory-only conditions (Task I; N=60), 10 for each of the six visual-only conditions (Task II; N=60), and 10 for each of the three combined auditory-visual conditions (Task III; N=30). These conditions will be described below under Procedure. Each participant was involved in only one of the three tasks (auditory-only, visual-only, or auditory-visual). All tasks were completed in a quiet room with the University of

Nevada, Las Vegas Auditory Perception Laboratory). A personal computer equipped with Music Experiment Development System (MEDS 99-C; Kendall, 1999) delivered stimuli and subsequently collected and summarized all responses. Visual stimuli were presented on a 19" computer monitor within the Auditory Perception Laboratory.

Stimulus Development

A large set of dynamic auditory-visual stimuli were recorded on S-VHS videotape in the University of Nevada, Las Vegas' Audio-Visual Studio. One native female speaker of American English was recorded while producing utterances intended to convey one of two opposing emotions (sadness or joy). The speaker was asked to act as if she were happy or sad as she conveyed the emotions. Utterances consisted of one of the three vowel-consonant-vowel (VCV) clusters /aba/, /ada/, and /aga/. These three clusters were chosen because they represent both labial and nonlabial consonants, a common consideration in studies of the McGurk effect. Nonsense syllables were chosen for the verbal content as to reduce the chance of participants having an emotional association with the word. Furthermore, Burns and Beier (1973) suggested that using non-emotional verbal content maximizes the accurate interpretation of vocal emotion without sacrificing acoustic quality that may be lost through filtering semantic content from an emotional production.

Ten of the clearest recordings of each of the six intended combinations of VCV cluster and emotion were selected by consensus of the author, chairperson, and first examining committee member to be included in the tasks. Clarity of recordings was

determined by perceived clarity of vocal articulation of the target cluster and clarity of physical characteristics (e.g., absence of shadow); thus, all stimuli selected were considered the ten best representations produced in each of the six VCV cluster/emotion combinations. All of the 60 resulting auditory-visual presentations were captured from the video cassette recorder using a MiroVideo DC-Plus 30 video capture card operating at NTSC-standard 30 frames per second playback. The presentations were subjected to editing using Adobe Premier 4.2 (see below).

Auditory-Only Stimuli

The auditory-only stimuli were created by separating the auditory channels from their original corresponding visual productions and saved as .wav files. The resulting auditory stimuli were all standardized at a length of two seconds after being digitized using Cool Edit 96 software and a Hewlett-Packard filter in the personal computer on which the editing was accomplished. The Hewlett-Packard filter used a cutoff of 90 hertz to eliminate background noise remaining from the recording process and the 60 hertz cycle noise contributed by the overhead fluorescent lighting in the room in which the recordings were made. Editing in Cool Edit 96 also allowed for standardization of stimulus volume. The final auditory signals were of a sample rate of 44.1 kilohertz and of 16-bit stereo quality.

Visual-Only Stimuli

After separating the video channels from their corresponding auditory channels, the video segments were then edited by superimposing black mattes of a length of six frames each onto both the beginning and end of each segment and saved as .avi files. The video

stimuli were all standardized at a length of two seconds (which included the lengths of both the video channel and the black mattes). The .avi files were compressed using Adobe Premier 4.2 utilizing the Indeo 3.2 codec at 100% quality. The data rate for presentation was fixed at 1000 kilobytes/second.

Auditory-Visual Presentations

Selected visual and auditory stimuli determined to vary with respect to perceived quality as an example of its respective emotion in Tasks I and II were combined to make the presentations for Task III. The emotionally congruent auditory-visual presentations were created by combining in a factorial manner the auditory and visual presentations selected as either the best or poorest representations of the same emotion during Tasks I and II. The emotionally incongruent auditory-visual presentations were created by combining these same auditory and visual segments into presentations that conveyed incongruous emotional information (i.e., a visual stimulus conveying joy with an auditory stimulus conveying sadness, or a visual stimulus conveying sadness with an auditory stimulus conveying joy). Thus, four auditory-visual presentations (two emotionally congruent and two emotionally incongruent) were created using each auditory-only stimulus. However, for all auditory-visual presentations, the VCV cluster was held constant to maintain phonetic synchrony. The emotionally congruent and incongruent presentations were created by synchronizing the consonantal bursts of the auditory and visual stimuli. This was done in order to produce the effect of natural speech and to prevent perception of temporal displacement. The combined auditory-visual stimuli

were edited to constitute the same auditory and visual quality and presented at the same filtered settings as the isolated auditory-only and visual-only signals.

Procedure

All data was collected by the author, as well as by undergraduate laboratory assistants under the direction of the author. Prior to data collection, all involved laboratory assistants were thoroughly trained in the use of software necessary to run the tasks and in the particulars of the study, including the ethical treatment of human participants. Additionally, approval for research involving human subjects was obtained from the University of Nevada, Las Vegas Institutional Review Board prior to commencing data collection.

Task I: Auditory Affect (Auditory-Only Presentation).

The goal of this task was to determine the best and poorest auditory representations of joy and sadness for each of the three VCV clusters through goodness and similarity ratings. The best and poorest auditory and visual representations (described in Task II) would then be combined to create the auditory-visual stimuli for inclusion in Task III.

The first task involved auditory-only stimuli and consisted of two parts. Part A determined the extent to which the stimuli varied with respect to perceived quality of the expressed emotion (either joy or sadness). Part B determined the perceived similarity between stimuli within each emotional category. Before beginning the task, participants were given a brief familiarization period during which they were asked to listen carefully to each stimulus (presented randomly) in order to become familiar with the stimuli of

which their respective tasks would consist. Participants were given the option to repeat this familiarization task a second time before beginning the tasks. All parts of this task were completed in a single session, lasting no more than one hour. Opportunities for rest were provided following the completion of Part A and, during Part B, between each block of trials.

In Part A, each of the 10 participants were presented with 10 auditory stimuli constituting one of the three VCV groups (/aba/, /ada/, or /aga/) and one of the two emotional categories (sadness or joy). Thus, there were six auditory-only conditions and 10 participants per group (N=60). Auditory stimuli were presented at a comfortable listening level [approximately 80 dB (A)] through Sennheiser HD 25-headphones and were low-pass filtered through a Krohnhite filter at five kilohertz. Participants rated the quality of each stimulus as an example of the intended emotion on a seven-point scale (e.g., 1 = very poor, through 7 = very good). Part B of the task consisted of one block of 100 trials (10 randomized repetitions of the 10 selected stimuli).

In Part B, participants subsequently rated the similarity of all possible pairs of the same stimuli (amounting to 90 randomly presented stimulus pairs) in order to determine the relative perceptual distance between stimuli within the emotion category (1 = very dissimilar through 7 = very similar). This part of the task consisted of five blocks of 90 trials each, totaling 450 trials.

Ratings from Part B of Task I were submitted to multidimensional scaling analyses, which provided best-fit maps of the perceived similarity relationships between all stimuli in the six categories. These maps were used to determine the physical acoustical

characteristics of the auditory stimuli that were relied upon for the judgments of goodness and similarity. These characteristics were evaluated by examining the pitch contours and spectrograms of each auditory stimulus using Computerized Speech Research Environment software (CSRE). Spectrographic settings utilized Moderated Covariance Parameters (window size = 128; number of bands = 256; overlap % = 90; order = 15; preemphasis = 98).

Task II: Facial Affect (Visual-Only Presentation)

Similar to Task I, the goal of this task was to determine the best and poorest visual representations of joy and sadness for each VCV cluster through goodness and similarity ratings. The best and poorest auditory and visual representations would then be combined to create the auditory-visual stimuli for inclusion in Task III.

This task involved visual-only stimuli and also consisted of two parts. Part A determined the extent to which the stimuli varied with respect to perceived quality of the expressed emotion (either joy or sadness). Part B determined the perceived similarity between stimuli within each emotional category. Before beginning the task, participants were given a brief familiarization period during which they were asked to carefully watch each stimulus (presented randomly) in order to become familiar with the stimuli of which their respective tasks would consist. Participants were given the option to repeat this familiarization task a second time before beginning the tasks. All parts of this task were completed in a single session, lasting no more than one and one-half hours. Opportunities for rest were provided following the completion of Part A and during Part B between each block of trials.

Each of the 10 participants were presented with 10 visual stimuli constituting one of the three VCV groups (/aba/, /ada/, or /aga/) and one of the two emotional categories (sadness or joy). Thus, there were six visual-only conditions and 10 participants per group (N=60). Other aspects of the procedure were conducted in the same manner as were the two sections of Task I.

Task III: Facial and Auditory Affect (Combined Presentation).

This task was designed to determine whether visual emotional expression can influence perception of auditory emotion, indicating the integration of facial and auditory affect. Thirty participants were included in this task (10 for the /aba/ condition, 10 for the /ada/ condition, and 10 for the /aga/ condition). Before beginning any part of the task, participants were given a brief familiarization period during which they were shown the stimuli of which the respective part of the task would consist in a random order. Participants were given the option to repeat this familiarization task a second time before beginning the actual task. All parts of this task were completed in a single session, lasting no more than one hour. Opportunities for rest were provided following the completion of Parts A and B and during Part C between each block of trials.

A set of four auditory-only and a set of four visual-only stimuli was selected for each VCV group for inclusion in Task III. Each set included a good example of joy, a good example of sadness, a poor example of joy, and a poor example of sadness, as determined through goodness ratings in Task I [auditory] and Task II [visual]). Isolated auditory-only and visual-only conditions consisting of these stimuli were included to provide a baseline measure of the perception of vocal and visual emotion and as a validity check to

ensure that the original stimuli conveyed the correct and intended emotion (joy or sadness). In the isolated auditory- and visual-only conditions, participants classified the four auditory-only (Part A) and four visual-only (Part B) stimuli as either acceptance, anger, disgust, expectancy, fear, joy, sadness, or surprise. The eight emotion categories were based upon the second level of Plutchik's revised (1980) circular model of eight primary emotions. Research (Fromme & O'Brien, 1982; Havlena, Holbrook, & Lehmann, 1989; Osgood, 1966) has suggested an empirically sound basis for both the choice of and relationships between adjacent emotions in Plutchik's circumplex model. After classifying each stimulus as representing one of the eight emotions, participants then rated the intensity of that emotion (1 = low intensity through 7 = high intensity). Each of the auditory-only and visual-only portions consisted of 60 trials (four stimuli presented randomly 15 times).

Following the isolated auditory- and visual-only conditions, the selected auditory and visual stimuli were combined in a factorial manner to make the auditory-visual presentations for Part C of Task III. In half of the presentations, the visual and auditory channels were congruent in terms of the expressed emotion, with varying differences in the quality of the expressed emotion (emotionally congruent condition). The other half of the presentations constituted the emotionally incongruent condition and consisted of emotionally incongruous auditory and visual channels. Part C was comprised of three blocks of eighty randomly presented auditory-visual trials containing these emotionally congruent and incongruent presentations for a total of 240 trials. On these trials, participants were instructed to classify the perceived auditory emotion while observing

the accompanying speech production. Ratings were made in the same manner as they were in the isolated auditory-only and visual-only conditions. Each stimulus was presented randomly 15 times over the course of the 240 trials. Thus, participants made 15 emotion classifications and 15 intensity ratings for each auditory-visual stimulus.

Data Analysis

To investigate the overall study design, multivariate analyses of variance (MANOVA) with three within-subjects factors and one between-subjects factor were used to evaluate the overall study design (see Table 1).

Table 1

Levels of Between- and Within-Subjects Factors Included in Analyses

Within-Subjects			Between-Subjects
Quality[a]	Emotion[b]	Condition[c]	Consonant[d]
Poor	Joy	Auditory + Congruent Good-Quality Visual (CG)	/aba/
Good	Sadness	Auditory + Congruent Poor-Quality Visual (CP)	/ada/
		Auditory Only (AO)	/aga/
		Auditory + Incongruent Good-Quality Visual (IG)	
		Auditory + Incongruent Poor-Quality Visual (IP)	

Notes. [a]Refers to the quality of the original auditory stimulus. [b]Refers to the emotion conveyed by the original auditory stimulus. [c]Letters in parentheses refer to condition type. [d]Refers to the vowel-consonant-vowel (VCV) cluster conveyed by the stimuli.

The between-subjects factor was the consonant used in the VCV cluster (/aba/,

/ada/, /aga/) (Consonant). The first within-subjects factor was the quality of the auditory stimuli chosen for inclusion in the presentations for Task III, as originally rated by participants in Task I (Quality). The second within-subjects factor reflected the originally intended emotion conveyed by the auditory channel (Emotion). The third within-subjects factor was Condition, consisting of the four types of emotionally congruent and incongruent auditory-visual presentations, as well as the isolated auditory-only stimuli (from this point forward, the two-letter descriptor in parentheses will be used in place of condition name). The dependent variables in these analyses were the participants' classifications of the perceived auditory affect conveyed by the five different conditions (e.g., CG, CP, AO, IG, IP).

In a first MANOVA, frequencies for the emotion classification choice "joy" served as the dependent variable. In a second MANOVA, frequencies for the emotion classification choice "sadness" served as the dependent variable. The classifications of "joy" and "sadness" were chosen for analysis to correspond to the two emotions originally conveyed by the auditory stimuli. The choice of these two categories also served to eliminate the dependence of the number of classifications in one category on the other seven categories. Since each participant provided 15 classifications for each stimulus, the number of classifications in each category was dependent upon the number of classifications in the other seven categories. Thus, selecting only classifications of "joy" and "sadness" for further analysis eliminated category dependence and permitted the number of classifications in each of the eight categories (otherwise categorical data) to be utilized as fractions (number of classifications out of 15 possible) and thus as

continuous data. The number of emotion classifications in each of the eight emotion categories was converted to a fraction, which reflected the number of times (out of a total of 15) that the participants classified the emotion as such. Post hoc univariate analyses (t-tests) were used to examine main effects and interaction effects identified by the MANOVAs. As the Greenhouse-Geisser statistic did not provide any additional statistical information beyond that provided by the basic F ratio, the basic F ratio was utilized.

Analysis of Specific Hypotheses

Hypothesis 1

Auditory-visual presentations of incongruent emotional information (sadness and joy), conveyed through nonverbal vocal and visual channels, will produce integrated emotional responses that are judged to constitute significantly different emotions than either sadness or joy.

To investigate Hypothesis 1, eight paired-samples t-test comparisons (one for each of the eight possible emotion classifications [acceptance, joy, expectancy, surprise, anger, disgust, sadness, fear]) were made. These comparisons were made for each of the eight emotions between the average number of times (out of 15 possible) the emotion was perceived in the emotionally congruent and incongruent auditory-visual presentations in which the auditory and visual conditions were of varying quality. This set of eight t-tests were run once for auditory-visual presentations involving auditory channels conveying joy and once for auditory-visual presentations involving auditory channels conveying sadness. This was done in order to examine the differential pattern of

classifications resulting from the inclusion in a stimulus of either a pleasant or unpleasant auditory emotion. To reduce the chance of incorrectly rejecting the null hypothesis, Bonferroni corrections for multiple comparisons were made. The critical alpha level for each set of eight comparisons was set at (.05/8) = $p<.00625$. These analyses were designed to determine whether there was a significant shift within emotion classification categories depending on whether the auditory-visual presentations conveyed emotionally congruent or incongruent information.

Hypothesis 2

As the emotionally incongruent auditory-visual presentations become more incongruous in terms of each channel's emotional intensity, the auditory channel will be more heavily relied upon to determine the resulting emotional production.

Hypothesis 2 was analyzed in a similar fashion in order to determine whether emotion classifications of isolated auditory-only stimuli and emotion classifications of emotionally incongruent auditory-visual presentations consisting of auditory and visual channels of varying intensity would produce a significant shift within each of the eight emotion classification categories. Classifications of the auditory-only stimuli were compared with the auditory-visual presentations including the same auditory stimulus. These analyses also employed paired-samples t tests with Bonferroni corrections for multiple comparisons. The Bonferroni correction set the critical alpha level for each set of eight comparisons at (.05/4) = $p<.0125$. Statistical nonsignificance, which would demonstrate no significant difference between frequency of classifications of the isolated auditory-only stimuli and classifications of the emotionally incongruent auditory-visual

presentations consisting of the same auditory stimulus and visual stimuli of a varying intensity, was required to confirm this hypothesis.

Hypothesis 3

However, as the combined presentations are more similar in terms of their emotional intensity, the visual channel will be predominant in the deciphering of the resulting emotional production.

Hypothesis 3 was analyzed in a fashion similar to that of Hypothesis 2. Four paired-samples t-tests with Bonferroni corrections for multiple comparisons were undertaken to determine whether emotion classification of isolated auditory-only stimuli and emotion classification of emotionally incongruent auditory-visual presentations consisting of same-quality auditory and visual channels would produce a significant shift within each of the eight emotion classification categories. Emotion classification of the auditory-only stimuli were compared with the auditory-visual presentations including the same auditory stimulus. The Bonferroni correction set the critical alpha level for each set of eight comparisons at $(.05/4) = p<.0125$. In this case, statistical significance, which would demonstrate a significant difference between emotion clasification of auditory-only stimuli and emotion classification of the emotionally incongruent auditory-visual presentations, was required to confirm this hypothesis.

CHAPTER 3

RESULTS

Multidimensional Scaling

A mean on the 7-point goodness scale was calculated for the ten stimuli in each of the six auditory-only conditions and six visual-only conditions. This mean was derived from all ratings provided for each stimulus in Part A of Tasks I (auditory) and II (visual). These means were used to determine which stimuli were perceived by the participants as being of the best and the poorest quality. Both good- and poor-quality stimuli were chosen in order to determine if the perceived quality of the channels would have an impact on the perceived emotion of the combined auditory-visual stimuli in Task III, as addressed in Hypotheses 2 and 3. The stimulus with the highest mean goodness rating within its respective category was chosen as the good-quality stimulus, and the stimulus with the lowest mean goodness rating within its respective category was chosen as the poor-quality stimulus. However, in cases when one or more mean goodness ratings were very similar or were identical, good- and poor-quality stimuli were chosen based on similarity ratings made in Tasks I and II. Thus, stimuli judged to be of identical quality were examined in terms of how dissimilar they were rated relative to other stimuli; in such instances, the selected stimulus was the one judged to be most dissimilar to the

remaining stimuli in the category. This method (choice of good- or poor-quality stimulus based on goodness ratings as well as dissimilarity ratings) was utilized in five categories in which there were stimuli with identical goodness ratings. These categories all involved stimuli conveying sadness (auditory /aba/, auditory /aga/, visual /aba/, visual /ada/, and visual /aga/). Dissimilarity rating matrices for these five categories are presented in Figures 1 through 5, located in Appendix II. For example, as shown, for auditory /aba/ stimuli conveying sadness, stimulus 3 was chosen over stimulus 15 as the poor-quality stimulus. This was because although both received a mean goodness rating of 4.1, the comparison of stimulus 3 to stimulus 12 (the stimulus chosen as the good-quality representative) had a dissimilarity rating of 3.4, while stimuli 15 and 12 had a dissimilarity rating of only 2.3. Thus, since stimulus numbers 12 and 3 had appeared to Task I participants to be more dissimilar from each other, these were the stimuli chosen to represent the good- and poor-quality representations, respectively, of auditory /aba/ stimuli conveying sadness. The stimuli for the other 4 categories mentioned (auditory /aga/, visual /aba/ , visual /ada/, and visual /aga/) were chosen in a similar fashion.

Based on these mean goodness ratings and where applicable, dissimilarity ratings, the selected auditory and visual stimuli were thus chosen and combined in order to create the presentations for the emotionally congruent and incongruent auditory-visual presentations for Task III. Tables 10 and 11 present mean ratings for the auditory and visual stimuli rated as best and poorest in their respective categories and chosen for inclusion in the auditory-visual presentations for Task III. (Tables 15 through 20, located in Appendix II, provide complete listings of mean goodness ratings for each individual

stimulus.) Tables 12 and 13 present frequencies of emotion classifications of these auditory and visual stimuli as perceived by participants in Task III. These emotion classification tasks served as a validity check to ensure that the selection of these stimuli as the best and poorest examples of their respective intended emotion was warranted.

As shown in Table 14 (located in Appendix II), a multidimensional scaling procedure produced two- or three-dimensional best-fit solutions for all six auditory-only categories. This procedure was conducted in order to facilitate investigation of how many and which acoustic characteristics were relied upon by participants to judge the emotion conveyed by the sixty auditory stimuli. One-dimensional solutions are also provided for comparison, since for two categories (/aba/ conveying joy and /aga/ conveying sadness), only one acoustic characteristic (discussed below) could be found, which best fit a one-dimensional solution. Tables 15 through 20 (located in Appendix II) present measurements of acoustic characteristics and Task I goodness ratings taken for all sixty auditory stimuli. Table 21 presents Task I goodness ratings for all sixty visual stimuli. The acoustic characteristics investigated include F_0 measurements in hertz at six specific points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). These six measurements provide a rough estimate of the shape of the pitch contour over time. For a specific point to be considered a valid measurement, it was required to fall within 20 Hz of the previous point (that is, it could not represent a sudden shift in direction or intensity) and within a uniform pitch contour. Uniform pitch contour was determined by a visual inspection of the general shape of the pitch contour, excluding points of high

hertz value which would suggest content more indicative of noise or unvoiced articulations (such as breathing). Also included are duration measurements in milliseconds of first and second syllables and total duration of each stimulus. These measurements were taken based on the suggestion in prior literature (e.g., Williams & Stevens, 1972; Scherer, 1986; Murray & Arnott, 1993) that characteristics of an auditory signal's pitch contour, as well as its duration, may be instrumental in judges' perception of what emotion is being conveyed by the signal.

After examining all auditory productions using CSRE, it was determined that no single dominant major physical acoustic characteristic could have been relied upon by participants for judging the goodness or similarity of either the sad or joyful auditory stimuli across all three VCV clusters. However, similarity ratings of auditory examples of /aba/ conveying joy appeared to be judged on the basis of peak F_0 during the first syllable. Stimuli with higher first syllable peaks were judged as better representations of joy than were those with lower first syllable peaks (see Table 15). Furthermore, similarity ratings of auditory examples of /aga/ conveying sadness appeared to be judged on the basis of rate of F_0 change as a function of time (Hz per second). This was determined by subtracting the "valley" F_0 measurement of the second syllable from the "peak" F_0 measurement of the first syllable, dividing the number of milliseconds elapsed between the two points, and multiplying the result by 1000) (see Table 20).

To determine if any stimuli within a particular category were rated as significantly better than others, which would serve as a justification for investigating the basis of perceptual differences between stimuli, one-way ANOVAs were conducted for the six

sets of 10 auditory stimuli. Although results of all six analyses were significant ($p <$ .0001), suggesting significant differences within the set of stimuli, very few stimuli were actually judged as significantly better (as determined by post hoc Scheffé tests) than others within their respective categories. For /aba/ stimuli conveying joy, only stimuli 11, 2, and 9 were rated as significantly better than the other seven stimuli in that category. This finding corresponds to the multidimensional scaling solution that suggests that these three stimuli involved a much higher peak F_0 in the first syllable. For /ada/ stimuli conveying joy, stimuli 3 and 2 were rated as significantly better than all other stimuli. For /aga/ stimuli conveying joy, stimuli 2 and 15 were rated as significantly better than all other stimuli. For /aba/ stimuli conveying sadness, only stimuli 12 and 4 were rated as significantly better than the three lowest-rated stimuli. For /ada/ stimuli conveying sadness, only stimulus 1 was rated as the three lowest-rated stimuli, and stimuli 2, 7, and 9 were rated as significantly better than the lowest-rated stimulus. For /aga/ stimuli conveying sadness, stimuli 26 and 15 were rated as significantly better than other stimuli. These results demonstrate that there were very few differences in similarity ratings between stimuli within the six categories.

Evaluation of Study Hypotheses

Fully factorial multivariate analyses of variance (MANOVA) with three within-subjects factors (Quality x Emotion x Condition) and one between-subjects factor (Consonant) were used to evaluate the overall study design (see Table 1). In the first MANOVA (see Table 2), frequency of the emotion choice "joy" served as the

dependent variable. In the second MANOVA (see Table 3), frequency of the emotion choice "sadness" served as the dependent variable. Post hoc univariate analyses (t-tests) were used to examine main effects and interaction effects identified by the MANOVAs. Results using the Greenhouse-Geisser statistic were examined for purposes of identifying a potential lack of sphericity in the data, but are not provided because they did not significantly alter any of the effects or conclusions from the original analyses.

All significant results of the first MANOVA are presented in Table 2. There were significant main effects for Condition and Emotion, as well as significant interactions for Condition x Consonant, Condition x Emotion, Condition x Quality, and Emotion x Quality. There was also a three-way interaction effect for Condition x Emotion x Quality. Results of the second MANOVA (based on frequency of the emotion choice "sadness") (see Table 3) indicated significant main effects for Condition and Emotion, as well as significant interaction effects for Condition x Consonant and Condition x Emotion. There was also a significant three way interaction (Condition x Emotion x Consonant, which is further described in the Discussion section under Additional Findings). Of central importance to primary hypotheses of the current investigation are the Condition x Emotion interactions. The relations of each of these interactions to study hypotheses are considered in the following sections.

Hypothesis 1

Auditory-visual presentations of incongruent emotional information (sadness and joy), conveyed through vocal (nonverbal) and visual channels, will produce integrated

emotional responses that are judged to constitute significantly different emotions than either sadness or joy.

Hypotheses 1 was supported by the significant Emotion x Condition interactions present when frequency of either the emotion choice "joy" (Table 2) or "sadness" (Table 3) were used as dependent variables. These interaction effect are presented graphically in Figure 6 (joy) and Figure 7 (sadness). In these figures, Conditions CG and CP comprised emotionally congruent auditory-visual presentations, Condition AO comprised the auditory-only stimuli, and Conditions IG and IP comprised the emotionally incongruent auditory-visual presentations. Figure 6 indicates that auditory stimuli conveying joy were most frequently rated as joy when included in the emotionally congruent auditory-visual presentations, but were infrequently rated as joy when included in the emotionally incongruent auditory-visual presentations. When the auditory stimuli conveyed sadness, the opposite pattern was noted. Conditions CG and CP were most infrequently rated as joy, as was also the case in Condition AO. However, when a happy face was paired with a sad voice (Conditions IG and IP), a shift in classification was present and the presentations were more frequently classified as joy.

The opposite pattern occurred when frequency of the emotion choice "sadness" was used as the dependent variable (see Figure 7). As expected, Conditions CG and CP were most frequently classified as sad. Little change in frequency of classifications was noted for Condition AO. However, when a sad auditory stimulus was paired with a happy visual stimulus (Conditions IG and IP), there was a marked shift in the perception of the

emotion, as indicated by a decrease in the number of times these emotionally incongruent presentations were classified as sad.

The significance of the interaction effects presented in Figures 6 and 7 were further evaluated using paired *t*-tests with the Bonferroni correction for multiple comparisons (alpha level = $p<.00625$). Even with this correction, results of the paired *t*-tests further confirmed the significance of the shift in emotions suggested by the significant interaction effects (see Tables 4 and 5). Results indicated a significant change in the frequencies of choice across emotion categories when the auditory and visual information both conveyed congruent emotional information compared to when the auditory and visual channels conveyed incongruent emotional information.

In order to determine which specific emotions were endorsed most often depending on Condition, difference scores were calculated in which the frequency of emotional classifications for the emotionally incongruent presentations were subtracted from those of the emotionally congruent presentations. To increase reliability, the frequencies for Conditions CG and CP and Conditions IG and IP were combined prior to deriving the difference scores, as preliminary inspection indicated minimal differences between the difference scores for Conditions CG and CP or between Conditions IG and IP. In the end, this procedure yielded eight difference scores (one for each of the eight possible emotion classification choices that participants could have used).

Figure 8 presents the difference scores for each level of the factor Emotion for the presentations including an auditory stimulus conveying joy and those conveying sadness.

The table below presents a hypothetical example that helps to clarify the rationale behind use of difference scores in this situation:

	AC	JO	EX	SU	AN	DI	SA	FE
C	0	15	0	0	0	0	0	0
I	0	0	0	0	15	0	0	0
C – I	0	15	0	0	-15	0	0	0

Note. C = emotionally congruent; I = emotionally incongruent. C – I = emotionally congruent – emotionally incongruent. AC = acceptance; JO = joy; EX = expectancy; SU = surprise; AN = anger; DI = disgust; SA = sadness; FE = fear.

This table presents "ideal" hypothetical data concerning auditory-visual stimuli containing an auditory signal conveying joy. Difference scores are presented in the C – I row of the table. In the emotionally congruent conditions, a participant classified the stimulus as joy 15 out of 15 times. However, in the emotionally incongruent conditions, when the joyful voice was paired with a sad face, the participant classified the stimulus as anger 15 out of 15 times. Thus, the positive difference score of 15 under the "joy" column suggests the direction in which the congruent conditions were classified because the positive score would suggest that the emotionally congruent conditions had received more endorsements of the emotion. Alternatively, the negative difference score of –15 under the "anger" column suggests the direction in which the incongruent conditions were classified, because the negative score would suggest that the emotionally incongruent conditions had received more endorsements of the emotion.

Similarly, for the purpose of these analyses, difference scores with positive values tended to indicate the most common classification of the emotion conveyed by the emotionally congruent auditory-visual presentations. On the other hand, negative difference scores tended to indicate the most common emotion classification of the emotionally incongruent auditory-visual presentations. These negative difference scores across the various emotion choices are informative regarding how the incongruent presentations were perceived.

As can be seen from Figure 8, auditory-visual presentations including an auditory stimulus conveying joy tended to be perceived as either surprise or joy in the emotionally congruent conditions but as either anger, disgust or sadness, and to a lesser degree, fear, in the emotionally incongruent conditions. Auditory-visual presentations including auditory stimuli conveying sadness tended to be perceived overwhelmingly as sadness in the emotionally congruent conditions but as either joy, acceptance, or surprise, and, to a lesser extent, expectancy, in the emotionally incongruent conditions.

To evaluate the statistical significance of the shift in the difference scores portrayed in Figure 8, a repeated-measures analysis of variance (ANOVA) was employed. In this analysis, the most frequently chosen emotion in the emotionally congruent conditions was compared to the emotions where a shift in classification, as evidenced by a negative difference score, was present. Only negative difference scores were used because the negative scores indicated the emotion choices in which the largest shift in emotion perception was present. For the data presented in Figure 8 representing presentations including an auditory channel conveying joy, the ANOVA was significant ($F(4,116) =$

56.36; $p<.001$). Simple within-subjects contrasts indicated that classifications of joy in the emotionally congruent conditions differed significantly from classifications of anger ($F(1,29) = 122.11$; $p<.001$), disgust ($F(1,29) = 98.83$; $p<.001$), sadness ($F(1,29) = 78.77$; $p<.001$), and fear ($F(1,29) = 82.32$; $p<.001$) in the emotionally incongruent conditions. Furthermore, in order to determine any significant difference in frequency of classifications between these four emotions most frequently endorsed in the emotionally incongruent conditions, a one-way ANOVA was undertaken. Post hoc Scheffé tests indicated that anger was perceived significantly more often than disgust, sadness, or fear in the emotionally incongruent conditions (all p's$<.01$).

For the data presented in Figure 8 representing presentations including an auditory channel conveying sadness, the ANOVA was significant ($F(4,116) = 28.37$; $p<.001$). Simple within-subjects contrasts indicated that classifications of sadness in the emotionally congruent conditions differed significantly from classifications of acceptance ($F(1,29) = 55.30$; $p<.001$), joy ($F(1,29) = 46.28$; $p<.001$), expectancy ($F(1,29) = 40.97$; $p<.001$), and surprise ($F(1,29) = 47.85$; $p<.001$) in the emotionally incongruent conditions. Furthermore, in order to determine any significant difference in frequency of classifications between these four emotions most frequently endorsed in the emotionally incongruent conditions, a one-way ANOVA was undertaken. Post hoc Scheffé tests indicated that none of these emotions were perceived significantly more often than any of the other three in the emotionally incongruent conditions (all p's$>.05$).

In summary, results of these analyses largely confirmed Hypothesis 1 by demonstrating a significant shift in emotion perception from emotionally congruent to

emotionally incongruent conditions. Additionally, results also suggest a directionality in this shift: a joyful face paired with a sad voice is most often perceived as joy or acceptance, while a sad face with a joyful voice is most often perceived as anger.

Hypothesis 2

As the emotionally incongruent auditory-visual presentations become more incongruous in terms of each channel's emotional intensity, the auditory channel will be more heavily relied upon to determine the resulting emotional production

The within-subjects Quality factor, representing the judgments of Task I participants of the quality of auditory stimuli chosen for inclusion in Task III as either poor or good, was selected as an analogue to emotional intensity. This decision was based on some support in the literature (e.g, Green & Cliff, 1975; Ekman, 1992) that suggests that judges' perception of the quality of an emotional presentation is highly related to their perception of its emotional intensity. However, Task III participants were unable to distinguish differences in intensity in the auditory-visual stimuli, particularly for the sad stimuli (p's all $> .05$). It is thus possible that the distinction between quality and intensity may be somewhat arbitrary. For example, participants could rate an emotion as being of "good quality" because it represented a higher intensity emotion and rate another emotion as being of "poor quality" because it represented a lower intensity emotion.

Analyses of Hypothesis 2 focused on evaluating the difference in number of classifications between a set of isolated auditory-only stimuli and the emotionally incongruent auditory-visual presentations including the same auditory-only stimuli.

Because Hypothesis 2 was concerned with the emotionally incongruent auditory-visual presentations, emotionally congruent presentations were not considered in these analyses. The results of these analyses are further explicated in Tables 22 and 23. Four paired *t*-tests (2 for auditory stimuli conveying joy and 2 for auditory stimuli conveying sadness) were conducted for each of the eight emotion response categories with Bonferroni corrections for multiple comparisons. This was done in order to determine whether a substantially different pattern of responses would be found between emotion classifications of isolated auditory-only stimuli and the emotionally incongruent auditory-visual presentations including the same auditory-only stimuli and a visual stimulus differing in quality. One set of paired *t*-tests was undertaken to evaluate these patterns (see Table 6). Results from these analyses did not support Hypothesis 2, in that only 3 of the 16 comparisons made between the auditory-only stimuli conveying joy and the emotionally incongruent auditory-visual presentations including the same auditory-only stimuli and a visual channel differing in quality were deemed statistically nonsignificant after taking into account the revised significance level ($p<.0125$) derived through the Bonferroni correction for multiple comparisons.

Similar to the emotion classifications of auditory-visual presentations involving auditory channels conveying joy, results from the analysis of auditory-visual presentations involving auditory channels conveying sadness also failed to support Hypothesis 2. As can be seen from Table 7, only 5 of the 16 comparisons made between auditory-only stimuli and emotionally incongruent auditory-visual presentations including the same auditory-only stimuli and a visual channel differing in quality were

statistically nonsignificant after taking into account the revised significance level ($p<.0125$) derived through the Bonferroni correction for multiple comparisons.

Hypothesis 3

However, as the combined presentations are more similar in terms of their emotional intensity, the visual channel will be predominant in the deciphering of the resulting emotional production.

Hypothesis 3 was confirmed, in that 13 of the 16 *t*-test comparisons between emotion classifications of emotionally incongruent auditory-visual presentations composed of two same-quality signals and the isolated same-quality auditory-only stimuli conveying joy were statistically significant after taking into account the revised significance level ($p<.0125$) derived through the Bonferroni correction for multiple comparisons (see Table 8). Furthermore, 12 of the 16 *t*-test comparisons made between emotion classifications of emotionally incongruent auditory-visual presentations composed of two same-quality channels and emotion classifications of the same isolated auditory-only stimuli conveying sadness were significant after taking into account the revised significance level ($p<.0125$) derived through the Bonferroni correction for multiple comparisons (see Table 9). These results suggest findings similar to those of Hypothesis 2, but in the opposite fashion: participants were unable to ignore emotionally incongruent visual information and focus on simply the auditory channel. The visual channel influences emotion perception regardless of the quality of either the auditory or visual channel.

An additional finding involves the interactions between Consonant and Condition, which were significant in both MANOVAs, as well as the significant Consonant x Emotion x Condition interaction in the second MANOVA run on the emotion classifications of "sadness" (see Tables 2 and 3). Graphical representations of these interactions are provided in Figures 9 (stimuli including an auditory signal conveying joy) and 10 (stimuli including an auditory signal conveying sadness). Since there was no significant main effect of Quality on emotion classification in either MANOVA, emotion classification of "joy" for good- and poor-quality auditory stimuli were averaged to obtain mean emotion choice frequencies shown in Figure 9. Furthermore, emotion classifications of "joy" for auditory-visual stimuli containing both good- and poor-quality auditory and visual signals were averaged to obtain mean emotion choice frequencies for emotionally congruent auditory-visual stimuli conveying joy. Figure 10 represents the same information for emotion classifications of "sadness."

Figure 9 suggests that although in the isolated Auditory-Only conditions, each consonant cluster received a nearly identical number of classifications of joy, in the emotionally congruent auditory-visual conditions, the /aga/ cluster, and to a smaller extent, the /ada/ cluster, received many more endorsements of "joy." Figure 10 suggests a nearly opposite pattern for stimuli involving an auditory signal conveying sadness. In the isolated Auditory-Only conditions, the mean number of classifications of sadness were quite similar; however, in the emotionally congruent auditory-visual conditions, those stimuli involving the /aba/ cluster received more endorsements of "sadness," while those involving the /ada/ and /aga/ clusters received fewer endorsements of "sadness." Paired t-

tests were run in order to determine if the distances between the two values in each VCV cluster condition were significant. As seen in Figure 9, both the distances between mean classifications of "joy" in the /ada/ and /aga/ conditions were significant ($p < .01$); however, none of the distances between mean classifications of "sadness" were significant (see Figure 10).

Figures 11 (stimuli including an auditory signal conveying joy) and 12 (stimuli including an auditory signal conveying sadness) present similar relationships between auditory-only stimuli and the emotionally incongruent auditory-visual stimuli. Again, there had been no significant main effect of Quality on emotion classification in either MANOVA. Thus, emotion ratings for good- and poor-quality auditory stimuli and emotion ratings for auditory-visual stimuli containing both good- and poor-quality auditory *and* visual signals were averaged to obtain mean classifications for emotionally incongruent auditory-visual stimuli.

Figure 11 suggests that although in the isolated Auditory-Only conditions, each consonant cluster received a nearly identical number of classifications of joy, in the emotionally incongruent auditory-visual conditions, the /ada/ cluster, and to a smaller extent, the /aga/ cluster, received more endorsements of "joy" than did the /aba/ condition, even though the stimuli were now conveying incongruent emotional messages. Figure 12 suggests a nearly opposite pattern for stimuli involving an auditory signal conveying sadness. In the isolated Auditory-Only conditions, the mean number of classifications of sadness were quite similar. However, in the emotionally incongruent auditory-visual conditions, those stimuli involving the /aba/ cluster received more

endorsements of "sadness," while those involving the /ada/ and /aga/ clusters received many fewer endorsements of "sadness," again despite the fact that the stimuli were conveying incongruent emotional messages. Paired t-tests were run in order to determine if the distances between the two values in each VCV cluster condition were significant. As seen in Figure 11, all distances between mean classifications of "joy" conditions were significant at at least the $p < .05$ level; however, only the distances between mean classifications of "sadness" for the /ada/ and /aga/ conditions were significant ($p < .05$) (see Figure 12).

CHAPTER 4

DISCUSSION

Integration of Auditory and Facial Affect

Previous research has suggested the dominance of auditory information in cases when incongruent emotional information differing in quality or intensity across modalities is presented simultaneously (for example, DePaulo et al., 1978). In contrast, the results of this investigation suggest that regardless of the quality of the emotionally incongruent visual information presented in conjunction with the auditory channel, the visual information cannot be ignored by participants and significantly influences the perception of the resulting auditory emotions. This can be readily argued by examining the results of Hypotheses 2 and 3 of the current investigation. Results of these hypotheses demonstrated that despite instruction to focus exclusively on the emotion conveyed by the auditory signal of an emotionally incongruent auditory-visual stimulus, participants were unable to ignore the information conveyed by the visual channel and invariably incorporated it into their interpretation of the auditory emotion. These results support those of previous research (e.g., Burns & Beier, 1973; Zaidel & Mehrabian, 1969) that also suggest that the visual channel is invariably more relied upon than the auditory channel.

Furthermore, data from the current investigation reinforce the existing literature on the integration of conflicting auditory and visual information and extend it to more specifically investigate interactions between incongruent emotional information. As hypothesized, a true emotional McGurk effect occurred when incongruent emotional information (sadness and joy), conveyed through vocal (nonverbal) and visual channels, was combined to produce an emotionally incongruent auditory-visual production. These emotionally incongruent auditory and visual sources appear to have been perceptually integrated and were judged to constitute significantly different emotions than either sadness or joy. When paired with visual displays of sadness, an auditory example of joy was most likely to be interpreted as anger followed by disgust. Conversely, although sadness remained the most common emotion classification by a small margin when a sad auditory stimulus was paired with a visual example of joy, classifications of acceptance and joy were much more common.

Earlier work (e.g., Massaro & Egan, 1996; Massaro, 1998; de Gelder & Vroomen, 2000) has demonstrated that visual emotional information influences the interpretation of auditory emotion. However, the present study extends this work by constructing stimuli in a fashion similar to that employed in standard studies of the McGurk effect and by allowing a wider range of emotional choices by which participants could convey a better sense of what emotion was truly perceived. The current results suggest that the combination of two emotionally incongruent auditory and visual stimuli produce the perception of an auditory emotion that is entirely different than that conveyed by either of the sources in isolation. This differs from previous claims (e.g., Massaro & Egan,

1996; Massaro, 1998; de Gelder & Vroomen, 2000) in several important ways. For example, the provision to participants of emotional categories with which to judge emotionally incongruent stimuli is crucial. Massaro and Egan (1996) and de Gelder and Vroomen (2000) instructed participants to classify incongruent stimuli given only two choices. Thus, participants had to render one channel useless in their classifications in order to describe their global impression of the stimulus. This method seriously limits the usefulness of these studies in that a great deal of vital information concerning possible emotional integration is lost through the use of such a small number of emotion classification choices.

Although Massaro (1998) increased the number of emotion classification categories to 4 and 6 for similar stimuli, this change did not significantly alter participant classifications. This lack of improvement might be explained by examining a second way in which the current study improves over previous research: the use of dynamic, ecologically valid stimuli. Previous research has incorporated the use of synthesized "talking heads" (Massaro & Egan, 1996; Massaro, 1998) or black-and-white still photographs (de Gelder & Vroomen, 2000). Furthermore, stimulus content has included spoken words, which may in themselves convey emotional meaning, which could obscure the interpretation of the emotion conveyed by the pure, nonverbal vocal signal and the visual signal. These methodological concerns limit the validity and generalizability of the results of these studies. On the other hand, the current investigation employed short (two-second) dynamic color videotaped stimuli of a human speaker pronouncing nonsense syllables in order to reduce the potential biasing effects of

actual words. The use of these ecologically valid stimuli is important in supporting the usefulness and meaningfulness of the results presented here. It is likely that the difference in results between this study and others (e.g., Massaro & Egan, 1996; Massaro, 1998; de Gelder & Vroomen, 2000) is largely due to the use in the current investigation of more natural stimulus materials.

Multidimensional Scaling

The limited findings from the multidimensional scaling portion of this investigation may be explained by a number of factors, including the lack of variability in stimulus quality and thus the questionable implications of the resulting multidimensional scaling solutions. The process by which a "good" or "poor" quality stimulus were endorsed as such in Tasks I and II and selected for inclusion in Task III stimuli was inherently dependent on the variation present in the original set of 10 auditory and 10 visual stimuli. The speaker in this study was instructed only to act as if she were sad, and was not instructed to vary her expression as a function of emotional intensity. Had the speaker been instructed to act "very sad" or "slightly sad," the variability of the stimuli within each category would undoubtedly have been much greater. By definition this suggests that all stimuli in each category were likely to represent examples of each emotion that were very similar in terms of intensity. This is also likely considering the fact that a large number of stimuli were created within a very short time period (less than 30 minutes). Furthermore, these reasons could explain the unsuccessful attempts by the participants to rate stimulus intensity: Engen and colleagues (1958) did note that the

variability with which facial expressions were rated was inherently dependent upon the range of stimuli presented. Thus, the multidimensional scaling solutions provided in this investigation may have been "stressed" in order to account for small differences between similarity ratings within categories.

These findings are nevertheless important. Research (e.g., Scherer, 1986; Murray & Arnott, 1993) has only suggested what might differentiate stimuli in different emotion categories (that is, what makes a sad stimulus appear sad when compared to a joyful stimulus?). No known previous research has investigated what differentiates stimuli in terms of goodness within one particular emotion category (that is, what makes one sad stimulus appear sadder than another sad stimulus?). Furthermore, the fact that there were some significant differences between stimuli within each category, as demonstrated by the ANOVAs run in each category, suggests that participants were still able to make judgments to some extent even within a small range of quality. Furthermore, the dimensions determined to relate to judges' perception of emotion in the joyful /aba/ condition (peak pitch value in the first syllable) and the sad /aga/ condition (rate of F_0 change as a function of time) are consistent with prior findings (e.g., Williams & Stevens, 1972; Murray & Arnott, 1993) which emphasize the importance of the pitch contour in interpreting emotional classification. The fact that the specific acoustic characteristics relied upon to judge emotion in the other four categories are as yet unclear may be because the intra-category differences were too small to determined within the constraints of a spectral analysis. Nonetheless, future research in this area needs to

include applying a multidimensional scaling procedure to stimuli which span a greater range of emotional intensity within each emotion.

Effects of Consonant on Perceived Emotion

The significant Consonant x Condition interactions in the two MANOVAs (as well as the significant Consonant x Emotion x Condition interaction in the second MANOVA) raise an intriguing question for future research concerning the role of phoneme pronunciation and articulation in the perception of emotion. As was demonstrated in Figure 9, the emotionally congruent audio-visual conditions (CG and CP) for /ada/ and /aga/ received a greater number of classifications of "joy" than did the corresponding conditions for /aba/, although all places of articulation received similar numbers of "joy" endorsements in the isolated Auditory-Only conditions. Furthermore, Figure 10 reveals that the emotionally congruent auditory-visual /aba/ condition received the most endorsements of "sadness" of any of the three categories, although the three VCV conditions received similar numbers of endorsements of "sadness" in the isolated Auditory-Only conditions. Similarly, in the emotionally incongruent auditory-visual conditions, the /aba/ condition continued to more often classified as sad and less often as joyful, while in the /ada/ and /aga/ conditions, the opposite situation was the case. A possible explanation for these results can be posited both by research into the universal emotions and by observable facial characteristics during phoneme pronunciation. For example, during pronunciation of the nonlabial VCV clusters /ada/ and /aga/, a speaker is most likely to refrain from making wide jaw movements and lip compressions. Thus, it is

more likely in these two conditions that teeth will be bared. These aspects of expression, in conjunction with the smiles produced by the speaker, are very likely to convey a prototypical expression of happiness (Ekman & Friesen, 1975). In contrast, the pronunciation of the labial cluster /aba/ requires lip closure, which is also required in order to frown. Thus, upon visual inspection of the pronunciation of /aba/, it is more likely that this might be interpreted as conveying the prototypical frown of sadness seen in the photographs of Ekman and Friesen (1975). Future research should address this possible phoneme-emotion connection and determine which VCV clusters might be the most appropriate for the investigation of certain emotions. This also lends credence to the supposition that facial and vocal displays of emotion should not be investigated in isolation or in non-dynamic displays. It is evident from the results of this investigation, as well as those of other bimodal speech and emotion perception studies, that visual information strongly influences the interpretation of vocally presented information.

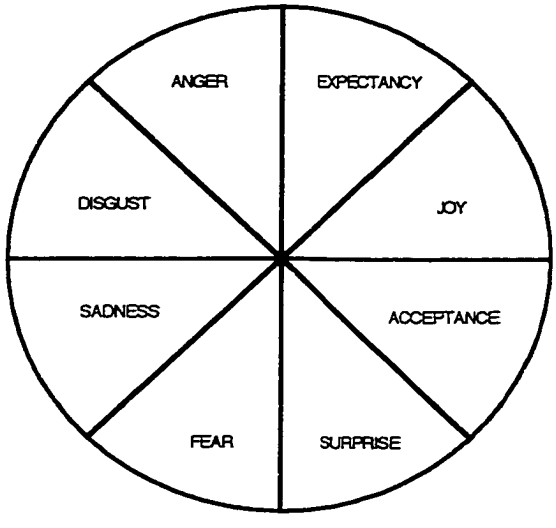
Implications for Models of Emotion Processing

An additional outgrowth of this research involves the possible misclassification of surprise as a negative emotion. In the second circular level of Plutchik's revised (1980) three-dimensional model of eight primary emotions (those used as the emotion choices for participant classification in Task III), surprise is located between sadness and fear in the emotion circumplex, implying that surprise carries an unpleasant connotation along the lines of shock. However, in the present study, participants invariably grouped surprise together with three other positive emotions; thus, participants appeared to

subconsciously dichotomize the eight emotion choices provided into two groups of emotions: positive (surprise, joy, acceptance, and expectancy) and negative (anger, disgust, fear, and sadness). This suggests that, at least in the current investigation, surprise may be misplaced in Plutchik's model and that it may be more appropriately classified near other positive emotions (acceptance and joy). However, future work may show that surprise can be interpreted in two different ways: as a pleasant surprise, as is suggested in this study, or as a more negative emotion indicative of shock, as it is currently displayed in Plutchik's model. These preliminary findings suggest that future research should take into account that surprise may be interpreted in two (or more) different ways.

Furthermore, work concerning Plutchik's original model does not take into account how judges should interpret combinations of emotions that are construed as polar opposites. The only work in this area has involved speculation on what emotion might result from combining two adjacent emotions. As it stands, if one were to attempt to reconcile opposite emotions with Plutchik's current circumplex, the model could only predict two things. On one hand, if two polar opposites of equal intensity were combined, they would meet near the center of the circle, suggesting a state of diffuse emotion (encompassing a small area of each of the eight emotions). In other words, this state would represent no emotion at all (as Plutchik's three-dimensional conical model suggests that the cone's point represents a state of no emotion). Alternatively, if one combined two polar opposites of unequal intensity, Plutchik's model would suggest that the most salient emotion would continue to be such, but in a less intense form.

Thus, these two possibilities would constitute very different outcomes than those implied in the results of the current investigation, which suggest that one may be able to predict what emotion will be perceived when emotional vocal information is paired with a visual representation of an emotionally incongruent polar opposite. Thus, a revision of Plutchik's second-level circumplex (and indirectly, the other levels as well) may be in order to explain the results of this investigation. The following restructuring is proposed, with surprise placed near other positive emotions in a more accurate position (based on the results of this study):



This revised version of Plutchik's model accurately accounts for the majority of results of Hypothesis 1, the main hypothesis of this study. One could have predicted these results by determining the location of the emotion of the dominant signal (in this case, the visual signal, with the auditory signal constituting the polar opposite) and then locating the emotion two wedges clockwise. For example, a sad visual signal combined

with a joyful auditory signal would most likely be perceived as anger (two wedges clockwise from sadness). Conversely, a joyful visual signal combined with a sad auditory signal would most likely be interpreted as surprise (two wedges clockwise from joy). Future research is necessary which would incorporate the revised Plutchik circumplex proposed here to predict what might be the emotion perceived when other polar opposites (for example, anger and surprise, disgust and acceptance, or fear and expectancy) are combined in a manner similar to that used in this investigation. Before these preliminary findings can be accepted, future research will also need to investigate the potential usefulness of this revised circumplex and replicate findings concerning the prediction of fusion responses using the polar opposites sadness and joy.

The revised circumplex adequately described the results of the combination of a joyful auditory signal and a sad visual signal. However, contrary to what would be predicted by the model, surprise was *not* the most frequently perceived emotion when participants experienced the fusion of a joyful visual signal and a sad auditory signal. This may be due to the emotional quality of the auditory and visual stimuli rated as the best and poorest representations of joy and sadness by Task I and II participants. It may also be due to the fact that when Task III participants classified the same stimuli using the eight emotion choices, the "best" and "poorest" representations of an emotion did not always even fall within the intended emotional category. For example, the visual stimuli chosen as the best examples of joy for /aba/, /ada/, and /aga/ by Task II participants were only endorsed as joy 58%, 29%, and 36% of the time, respectively, by Task III participants. The same stimuli were also predominately perceived as conveying

surprise by Task III participants (18%, 47%, and 58%, respectively). As a further illustration, the visual /aga/ stimulus chosen as the *poorest* example of joy actually received more endorsements of joy by Task III participants (55% of the endorsements possible) than the 36% received by the stimulus endorsed by Task II participants as the *best* representation of joy. This suggests that the reason why a prediction based on such a revised model may not have worked for such data in this case, as the visual stimuli which were supposed to have represented joy (based on Task I and II data) did not represent joy to Task III participants. This also relates back to the limitation inherent in the low variability range of the stimuli presented to Task I and II participants for evaluation and how such a limitation affected the production and subsequent emotional interpretation of the Task III auditory-visual stimuli. These findings further suggest that having two distinct groups of participants rate the same stimuli in two different manners (that is, on a goodness continuum or using eight categorical emotion choices) may reveal that the stimuli endorsed by one group as the best or poorest example of one emotion may not even be perceived as representing that emotion to the same degree by the second group.

Although this study provides intriguing insight into possibly unconscious combination of incongruent auditory and visual information, limitations are present within its methodology. Aside from the aforementioned low range of variability of stimulus quality within each of the six VCV/emotion categories, the use of only one female speaker raises questions about possible gender influences on conveyed emotion. Future research will want to address whether the pattern of results outlined in this

investigation may be idiosyncratic to this speaker or whether the results can be reliably replicated across speakers and participants.

Another possible direction for future research would be to have participants respond on a "believability" or confidence-rating dimension after making an emotion classification of emotionally incongruent auditory-visual stimuli. As suggested by McGurk and MacDonald (1976, 1978), participants were able to notice a discrepancy between the incongruent phonetic information included in the presentations, even though such knowledge was unable to prevent the fusion of such information. It would be interesting to investigate whether participants are able to notice that the auditory-visual stimuli conveyed emotionally incongruent information, or whether the integration of the incongruent information occurs at an unconscious and unrecognizable level. Although participants in this study were not asked whether they had noticed a discrepancy between the two channels, it is likely that is such a discrepancy is noticed in phonetic McGurk effect studies, it would also be noticed in emotional McGurk effect work.

The results of this investigation also tentatively support the supposition in previous research (e.g., Ekman, 1992) that methods of emotional interpretation may be innate and universal. In this study, classifications of auditory stimuli almost invariably shifted to those of an opposite emotion on a pleasantness-unpleasantness dimension when paired with an emotionally incongruent visual stimulus. To illustrate, a joyful auditory stimulus tended to be perceived as negative when combined with a sad visual stimulus, whereas a sad auditory stimulus tended to be perceived as conveying a positive emotion when paired with a joyful visual stimulus. One possible explanation for the current results is

that humans have a specific, unconscious method for perceiving and interpreting emotionally incongruent and potentially deceptive communications in order to avoid dangerous and potentially life-threatening situations. These results were obtained using 30 different participants, some with diverse backgrounds, across stimuli incorporating three different VCV clusters. Thus, it is likely that the current results would generalize to other populations of American adults. However, it has been noted in the literature (e.g., Sekiyama & Tohkura, 1991; Sekiyama, 1997a) that native speakers of Japanese and Chinese are less susceptible to making McGurk-type phonetic errors, which could possibly be due to Asian cultural aversions to observing the lip and mouth movements of others. Thus, future research will want to extend the current investigation to include more culturally diverse participants in order to more adequately answer to the possible innateness and universality of emotion perception.

The current investigation is significant not only in its contribution to theoretical implications of emotion perception and interpretation, but it also has import for applied fields with interest in emotional processing. As a clinical method, stimuli similar to those involved in this study could function as a diagnostic instrument to assess for deficits in emotional functioning, as have been observed in disorders such as schizophrenia or autism. As a research instrument, these findings could have implications for better understanding of the perceptual and conscious levels at which emotion processing (and integration) appear to occur. Future research in both of these areas will not only help to clarify the nature and processes of emotional perception and interpretation but may also serve an applied purpose in the diagnosis and treatment of mental disorder.

REFERENCES

Abelson, R.P., & Sermat, V. (1962). Multidimensional scaling of facial expressions. Journal of Experimental Psychology, 63, 546-554.

Argyle, M., Alkema, F., & Gilmour, R. (1971). The communication of friendly and hostile attitudes by verbal and non-verbal signals. European Journal of Social Psychology, 1, 385-402.

Argyle, M., Salter, V., Nicholson, H., Williams, M., & Burgess, P. (1970). The communication of inferior and superior attitudes by verbal and non-verbal signals. British Journal of Social and Clinical Psychology, 9, 222-231.

Banse, R., & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 614-636.

Berman, H.J., Shulman, A.D., & Marwit, S.J. (1976). Comparison of multidimensional decoding of affect from audio, video, and audiovideo recordings. Sociometry, 39, 83-89.

Boucher, J.D., & Ekman, P. (1975). Facial areas and emotional information. Journal of Communication, 25, 21-29.

Brighetti, G., Ladavas, E., & Ricci-Bitti, P.E. (1980). Recognition of emotion expressed through voice. Italian Journal of Psychology, 7, 121-127.

Bugental, D.E., Kaswan, J.W., & Love, L.R. (1970). Perception of contradictory meanings conveyed by verbal and nonverbal channels. Journal of Personality and Social Psychology, 16, 647-655.

Burns, K.L., & Beier, E.G. (1973). Significance of vocal and visual channels in the decoding of emotional meaning. Journal of Communication, 23, 118-130.

Costanzo, F.S., Markel, N.N., & Costanzo, P.R. (1969). Voice quality profile and perceived emotion. Journal of Counseling Psychology, 16, 267-270.

Davitz, J.R. (1964). Auditory correlates of vocal expressions of emotional meanings. In J.R. Davitz (Ed.), The Communication of Emotional Meaning (pp. 101-112). New York: McGraw-Hill.

Davitz, J.R., & Davitz, L.J. (1959). The communication of feelings by content-free speech. Journal of Communication, 9, 6-13.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. Cognition and Emotion, 14, 289-311.

DePaulo, B.M., Rosenthal, R., Eisenstat, R.A., Rogers, P.L., & Finkelstein, S. (1978). Decoding discrepant nonverbal cues. Journal of Personality and Social Psychology, 36, 313-323.

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6, 169-200.

Ekman, P. (1993). Facial expression and emotion. American Psychologist, 48, 384-392.

Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology, 53, 712-717.

Ekman, P., Friesen, W.V., & Tomkins, S.S. (1971). Facial affect scoring technique: A first validity study. Semiotica, 3, 37-58.

Engen, T., Levy, N., & Schlosberg, H. (1958). The dimensional analysis of a new series of facial expressions. Journal of Experimental Psychology, 55, 454-458.

Fromme, D.K., & O'Brien, C.S. (1982). A dimensional approach to the circular ordering of the emotions. Motivation and Emotion, 6, 337-363.

Gates, G.S. (1927). The role of the auditory element in the interpretation of emotion. Psychological Bulletin, 24, 175. (Abstract)

Gladstones, W.H. (1962). A multidimensional study of facial expression of emotion. Australian Journal of Psychology, 14, 95-100.

Green, K.P., & Gerdman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. Journal of Experimental Psychology: Human Perception and Performance, 21, 1409-1426.

Green, K.P., Kuhl, P.K., Meltzoff, A.N., & Stevens, E.B. (1991). Integrating speech information across talkers, genders, and sensory modality: Female faces and male voices in the McGurk effect. Perception and Psychophysics, 50, 524-536.

Green, R.S., & Cliff, N. (1975). Multidimensional comparisons of structures of vocally and facially expressed emotion. Perception and Psychophysics, 17, 429-438.

Hardison, D. (1996). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. Language Learning, 46, 3-73.

Havlena, W.J., Holbrook, M.B., & Lehmann, D.R. (1989). Assessing the validity of emotional typologies. Psychology and Marketing, 6, 97-112.

Hess, U., Scherer, K.R., & Kappas, A. (1988). Multichannel communication of emotion: Synthetic signal production. In K.R. Scherer (Ed.), Facets of Emotion: Recent Research (pp. 161-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, W.F., Emde, R.N., Scherer, K.R., & Klinnert, M.D. (1986). Recognition of emotion from vocal cues. Archives of General Psychiatry, 43, 280-283.

Kendall, R.A. (1999). Music Experiment Development System (1999) [Computer software]. Los Angeles: University of California, Los Angeles.

Langfeld, H.S. (1918). The judgment of emotion from facial expressions. Journal of Abnormal Psychology, 13, 172-184.

Levitt, E.A. (1964). The relationship between abilities to express emotional meanings vocally and facially. In J.R. Davitz (Ed.), The Communication of Emotional Meaning (pp. 87-100). New York: McGraw-Hill.

Ludemann, P.M. (1991). Generalized discrimination of positive facial expressions by seven- and ten-month-old infants. Child Development, 62, 55-67.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. Perception and Psychophysics, 24, 253-257.

Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, MA: The MIT Press.

Massaro, D.W., & Egan, P.B. (1996). Perceiving affect from the voice and the face. Psychonomic Bulletin and Review, 3, 215-221.

Massaro, D.W., & Ellison, J.W. (1996). Perceptual recognition of facial affect: Cross-cultural comparisons. Memory and Cognition, 24, 812-822.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Mehrabian, A., & Ferris, S.R. (1967). Inference of attitudes from nonverbal communication in two channels. Journal of Consulting Psychology, 31, 248-252.

Meltzoff, A.N., & Moore, M.K. (1977). Imitation of facial and manual gestures in human neonates. Science, 198, 75-78.

Murray, I.R., & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America, 93, 1097-1108.

Osgood, C.E. (1966). Dimensionality of the semantic space for communication via facial expressions. Scandinavian Journal of Psychology, 7, 1-30.

Plutchik, R. (1980). Emotions: A psychoevolutionary synthesis. New York: Harper & Row.

Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (1997). The McGurk effect in infants. Perception and Psychophysics, 59, 347-357.

Saldaña, H.M., & Rosenblum, L.D. (1993). Visual influences on auditory pluck and bow judgments. Perception and Psychophysics, 54, 406-416.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kaettoe, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. Speech Communication, 26, 75-87.

Scherer, K.R. (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99, 143-165.

Schlosberg, H. (1954). Three dimensions of emotion. Psychological Review, 61, 81-88.

Sekiyama, K. (1997a). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. Perception and Psychophysics, 59, 73-80.

Sekiyama, K. (1997b). Audiovisual speech perception and its inter-language differences. Japanese Journal of Psychonomic Science, 15, 122-127.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America, 90, 1797-1805.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. Journal of Psycholinguistic Research, 28, 347-365.

Soken, N.H., & Pick, A.D. (1999). Infants' perception of dynamic affective expressions: Do infants distinguish specific expressions? Child Development, 70, 1275-1282.

Tomkins, S.S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. Perceptual and Motor Skills, 18, 119-158.

Williams, C.E., & Stevens, K.N. (1972). Emotions and speech: Some acoustical correlates. The Journal of the Acoustical Society of America, 52, 1238-1250.

Zaidel, S.F., & Mehrabian, A. (1969). The ability to communicate and infer positive and negative attitudes facially and vocally. Journal of Experimental Research in Personality, 3, 233-241.

APPENDIX I

TABLES AND FIGURES REFERENCED IN TEXT

Table 2

Repeated Measures Analyses of Variance F Ratios for Emotion Classifications (Auditory Stimuli Conveying Joy)

Source	df	SS	MS	F
		Between Subjects		
Consonant	2	176.94	88.47	2.10
Error	27	1135.93	42.07	
		Within Subjects		
Condition	4	120.86	48.55	3.93*
Cond x Cons	8	166.52	33.45	2.71*
Emotion	1	888.17	888.17	18.70***
Error	27	1282.51	47.50	
Cond x Emot	4	2084.38	950.65	40.60***
Error	108	1386.04	23.41	
Cond x Qual	4	55.24	16.49	10.32***
Error	108	144.50	1.60	
Emot x Qual	1	17.34	17.34	6.65*
Error	27	70.43	2.61	

Note. Cond = Condition; Cons = Consonant; Qual = Quality; Emot = Emotion. *$p < .05$; **$p < .01$; ***$p < .001$.

Table 3

Repeated Measures Analyses of Variance F Ratios for Emotion Classifications (Auditory Stimuli Conveying Sadness)

Source	df	SS	MS	F
		Between Subjects		
Consonant	2	105.21	52.61	.90
Error	27	1579.38	58.50	
		Within Subjects		
Condition	4	300.56	92.97	6.63***
Cond x Cons	8	485.32	75.06	5.36***
Emotion	1	4771.44	4771.44	58.98***
Error	27	2184.32	80.90	
Cond x Emot	4	1674.89	749.52	24.53***
Error	108	1843.58	30.56	
Cond x Emot x Cons	8	508.83	113.85	3.73**
Cond x Emot x Qual x Cons	8	37.33	5.80	2.37*

Note. Cond = Condition; Cons = Consonant; Qual = Quality; Emot = Emotion. *$p < .05$; **$p < .01$; ***$p < .001$.

Table 4

Group Differences of Emotion Classifications of Stimuli for Hypothesis 1 (Auditory Signals Conveying Joy)

Emotion	Congruent Auditory-Visual		Incongruent Auditory-Visual		t (29)
	M	SD	M	SD	
Acceptance	2.77	2.77	1.34	2.11	2.69
Joy	6.23	3.42	1.23	2.91	9.14**
Expectancy	1.19	1.43	1.13	1.46	.18
Surprise	4.37	2.18	1.39	1.94	6.87**
Anger	.08	.29	4.85	3.59	-7.22**
Disgust	.13	.59	2.21	1.80	-6.83**
Sadness	.18	.68	2.13	2.52	-4.22**
Fear	.05	.15	.73	.98	-3.95**

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.00625$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.00125$ using Bonferroni correction critical alpha value).

Table 5

Group Differences of Emotion Classifications of Stimuli for Hypothesis 1 (Auditory Signals Conveying Sadness)

Emotion	Congruent Auditory-Visual		Incongruent Auditory-Visual		t (29)
	M	SD	M	SD	
Acceptance	.74	1.40	3.52	3.00	-4.88**
Joy	.12	.33	3.26	3.48	-5.10**
Expectancy	.58	1.04	1.66	2.11	-2.33
Surprise	.11	.24	1.88	2.26	-4.50**
Anger	1.51	1.67	.14	.34	4.33**
Disgust	1.86	1.76	.49	.96	4.01**
Sadness	8.50	3.38	3.68	4.96	6.85**
Fear	1.58	1.77	.38	.84	3.95**

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.00625$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.00125$ using Bonferroni correction critical alpha value).

Table 6

Group Differences of Emotion Classifications of Stimuli for Hypothesis 2 (Auditory Signals Conveying Joy)

Emotion	Good Auditory-Only		Incongruent Good Auditory Poor Visual		t (29)
	M	SD	M	SD	
Acceptance	2.60	2.46	.73	1.20	3.75**
Joy	5.43	2.89	1.20	3.22	7.32**
Expectancy	.90	.85	.67	1.52	.74
Surprise	4.80	2.91	2.27	3.73	3.54**
Anger	.67	1.52	5.93	4.89	-6.13**
Disgust	.27	.69	1.33	1.49	-3.82**
Sadness	.03	.18	2.03	3.57	-3.07*
Fear	.30	.70	.83	1.46	-1.89

Emotion	Poor Auditory-Only		Incongruent Poor Auditory Good Visual		t (29)
	M	SD	M	SD	
Acceptance	5.33	3.20	2.23	3.64	4.58**
Joy	3.20	2.55	1.23	3.03	4.41**
Expectancy	3.07	2.27	1.57	2.54	2.19
Surprise	1.90	2.07	.77	1.81	2.66*
Anger	.60	1.16	3.67	3.95	-4.47**
Disgust	.63	.89	2.90	3.06	-4.16**
Sadness	.17	.53	2.07	2.91	-3.48**
Fear	.10	.31	.57	1.14	-2.54*

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.0125$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.0025$ using Bonferroni correction critical alpha value).

Table 7

Group Differences of Emotion Classifications of Stimuli for Hypothesis 2 (Auditory Signals Conveying Sadness)

Emotion	Good Auditory-Only		Incongruent Good Auditory Poor Visual		t (29)
	M	SD	M	SD	
Acceptance	.37	.77	4.57	3.65	-5.97**
Joy	.50	1.33	2.73	3.89	-3.45**
Expectancy	.37	.62	1.63	2.16	-3.38**
Surprise	.10	.31	.93	1.80	-2.53
Anger	.53	1.14	.00	.00	2.57
Disgust	1.70	2.05	.60	1.25	2.59
Sadness	9.03	4.79	4.13	5.10	4.32**
Fear	2.40	3.28	.40	1.22	3.50**

Emotion	Poor Auditory-Only		Incongruent Poor Auditory Good Visual		t (29)
	M	SD	M	SD	
Acceptance	.80	1.35	2.73	2.89	-3.30*
Joy	.10	.31	3.33	4.11	-4.42**
Expectancy	.70	1.12	1.76	2.74	-1.91
Surprise	.20	.41	2.90	3.99	-3.66**
Anger	.37	.72	.13	.57	1.49
Disgust	3.03	2.86	.30	.75	5.27**
Sadness	8.10	4.79	3.47	5.24	4.57**
Fear	1.70	1.93	.37	1.13	3.32**

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.0125$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.0025$ using Bonferroni correction critical alpha value).

Table 8

Group Differences of Emotion Classifications of Stimuli for Hypothesis 3 (Auditory Signals Conveying Joy)

Emotion	Good Auditory-Only		Incongruent Good Auditory Good Visual		t (29)
	M	SD	M	SD	
Acceptance	2.60	2.46	1.30	2.26	1.87
Joy	5.43	2.86	1.50	3.18	7.90**
Expectancy	.90	.85	1.83	2.61	-1.85
Surprise	4.80	2.91	2.10	3.29	3.69**
Anger	.67	1.52	2.80	3.70	-3.12*
Disgust	.27	.69	2.53	3.28	-3.65**
Sadness	.03	.18	2.10	3.91	-2.89*
Fear	.30	.70	.83	1.66	-2.08

Emotion	Poor Auditory-Only		Incongruent Poor Auditory Poor Visual		t (29)
	M	SD	M	SD	
Acceptance	5.33	3.20	1.10	2.99	6.62**
Joy	3.20	2.55	.97	2.83	5.35**
Expectancy	3.07	2.27	.43	1.10	5.27**
Surprise	1.90	2.07	.43	1.33	4.21**
Anger	.60	1.16	7.00	5.18	-6.77**
Disgust	.63	.89	2.07	2.33	-3.46**
Sadness	.17	.53	2.30	3.78	-3.12*
Fear	.10	.31	.70	1.21	-2.90*

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.0125$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.0025$ using Bonferroni correction critical alpha value).

Table 9

Group Differences of Emotion Classifications of Stimuli for Hypothesis 3 (Auditory Signals Conveying Sadness)

Emotion	Good Auditory-Only		Incongruent Good Auditory Good Visual		t (29)
	M	SD	M	SD	
Acceptance	.37	.77	2.83	3.62	-3.54**
Joy	.50	1.33	3.20	3.82	-3.94**
Expectancy	.37	.62	1.43	1.94	-2.87*
Surprise	.10	.31	2.60	3.74	-3.66**
Anger	.53	1.14	.20	.48	1.78
Disgust	1.70	2.05	.77	1.61	2.42
Sadness	9.03	4.79	3.57	5.06	5.29**
Fear	2.40	3.28	.40	1.10	3.41**

Emotion	Poor Auditory-Only		Incongruent Poor Auditory Poor Visual		t (29)
	M	SD	M	SD	
Acceptance	.80	1.35	3.93	3.40	-4.51**
Joy	.10	.31	3.77	4.38	-4.67**
Expectancy	.70	1.12	1.80	2.72	-2.08
Surprise	.20	.41	1.07	1.53	-3.31**
Anger	.37	.72	.23	.77	.66
Disgust	3.03	2.86	.30	.84	5.49**
Sadness	8.10	4.79	2.53	5.13	4.34**
Fear	1.70	1.93	.37	.81	3.64**

Note. *$p<.05$ ($p<.05$ is equivalent to $p<.0125$ using Bonferroni correction critical alpha value); **$p<.01$ ($p<.01$ is equivalent to $p<.0025$ using Bonferroni correction critical alpha value).

Table 10

Goodness and Distance Ratings of Auditory Stimuli Chosen for Use in the Creation of Auditory-Visual Stimuli for Task III

VCV Cluster		Joy		Sadness	
		M	SE	M	SE
/aba/	Poor	3.5	.35	4.1	.32
	Good	5.6	.16	5.2	.28
	Distance	4.4	.29	3.4	.37
/ada/	Poor	3.8	.30	3.7	.42
	Good	5.4	.23	5.4	.38
	Distance	3.3	.40	4.8	.32
/aga/	Poor	3.9	.37	3.3	.30
	Good	6.2	.23	5.3	.44
	Distance	4.6	.23	3.7	.37

Note. Ratings were made on a 7-point scale (1 = *very poor*, 7 = *very good*). Distance scores were measured on a 7-point scale (0 = *very similar*, 6 = *very dissimilar*) and refer to the mean distance between the two chosen stimuli as rated by participants in Task I.

Table 11

Goodness and Distance Ratings of Visual Stimuli Chosen for Use in the Creation of Auditory-Visual Stimuli for Task III

VCV Cluster		Joy		Sadness	
		M	SE	M	SE
/aba/	Poor	2.7	.47	4.6	.33
	Good	5.3	.22	3.7	.14
	Distance	3.8	.51	3.6	.70
/ada/	Poor	3.1	.35	4.5	.17
	Good	5.2	.36	3.6	.29
	Distance	3.6	.28	3.7	.59
/aga/	Poor	3.0	.35	4.4	.77
	Good	5.6	.30	3.8	.30
	Distance	3.9	.21	3.8	.48

Note. Ratings were made on a 7-point scale (1 = *very poor*, 7 = *very good*). Distance scores were measured on a 7-point scale (0 = *very similar*, 6 = *very dissimilar*) and refer to the mean distance between the two chosen stimuli as rated by participants.

Table 12

Emotion Classifications of Auditory Stimuli Used in the Creation of Stimuli for Task III

VCV Cluster		Joy		Sadness	
		Emotion	%[a]	Emotion	%[a]
/aba/	Poor	Acceptance	40	Sadness	59
		Expectancy	23	Disgust	22
		Joy	21		
		Surprise	12		
/aba/	Good	Surprise	39	Sadness	59
		Joy	35	Disgust	17
		Acceptance	18	Fear	15
/ada/	Poor	Acceptance	27	Sadness	54
		Joy	22	Disgust	17
		Expectancy	21	Fear	14
		Surprise	16		
/ada/	Good	Joy	34	Sadness	68
		Acceptance	26	Fear	12
		Surprise	24		
/aga/	Poor	Acceptance	40	Sadness	49
		Joy	21	Disgust	21
		Expectancy	17	Fear	11
		Surprise	10		
/aga/	Good	Joy	40	Sadness	59
		Surprise	33	Fear	21
		Anger	11	Disgust	11

Note. Emotion choices were derived from the second level of Plutchik's revised (1980) circular model of eight primary emotions. [a]Results are presented for emotion categories receiving at least 15 endorsements (10%; 150 total endorsements across participants for each stimulus).

Table 13

Emotion Classifications of Visual Stimuli Used in the Creation of Stimuli for Task III

VCV Cluster		Joy		Sadness	
		Emotion	%[a]	Emotion	%[a]
/aba/	Poor	Expectancy	31	Anger	38
		Acceptance	26	Sadness	33
		Joy	23	Fear	15
		Surprise	19	Disgust	13
/aba/	Good	Joy	58	Sadness	67
		Surprise	18	Disgust	16
		Expectancy	13	Fear	10
		Acceptance	11		
/ada/	Poor	Joy	31	Anger	39
		Acceptance	30	Sadness	29
		Expectancy	21	Disgust	17
		Surprise	15	Fear	14
/ada/	Good	Surprise	47	Sadness	48
		Joy	29	Anger	23
		Expectancy	12	Disgust	19
		Acceptance	11		
/aga/	Poor	Joy	55	Sadness	71
		Acceptance	24	Disgust	14
		Surprise	13		
/aga/	Good	Surprise	58	Fear	31
		Joy	36	Expectancy	25
				Acceptance	14
				Disgust	11

Note. Emotion choices were derived from the second level of Plutchik's revised (1980) circular model of eight primary emotions. [a]Results are presented for emotion categories receiving at least 15 endorsements (10%; 150 total endorsements across participants for each stimulus).

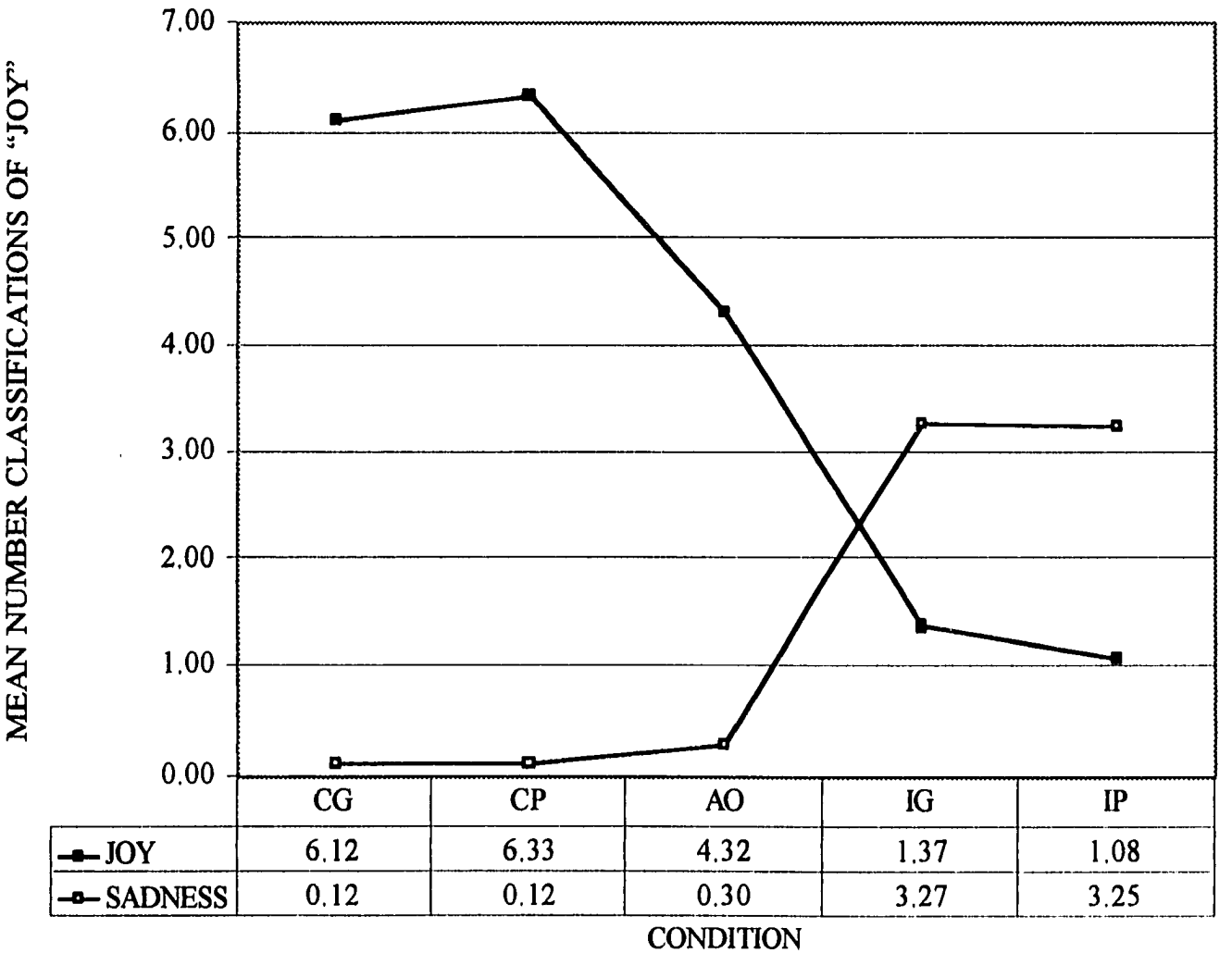
Table 21

Task I Goodness Ratings of the 60 Visual Stimuli

Joy						Sadness					
/aba/		/ada/		/aga/		/aba/		/ada/		/aga/	
Stimulus	M	Stimulus	M	Stimulus	M	Stimulus	M	Stimulus	M	Stimulus	M
11	5.3	10	3.1	10	3.6	11	4.1	1	4.3	11	3.8
15	2.7	15	5.2	11	3.0	12	3.9	14	3.6	12	4.0
16	3.9	18	4.3	15	4.7	15	4.6	15	3.8	15	3.9
17	3.6	2	4.9	16	5.6	16	3.8	2	4.3	20	4.4
2~1	4.4	21	4.1	19	3.9	18	4.1	20	3.7	21	4.2
2~11	5.1	24	4.1	2	3.4	2	3.9	25	3.9	22	3.8
2~4	5.2	25	4.2	20	5.0	3	3.7	4	3.8	23	3.9
2	5.2	29	4.5	21	3.6	4	3.9	6	4.4	24	3.9
23	5.1	3	3.9	8	3.7	8	4.0	7	4.5	26	3.9
9	5.2	9	4.6	9	5.1	9	3.8	9	4.5	3	3.9

Figure 6

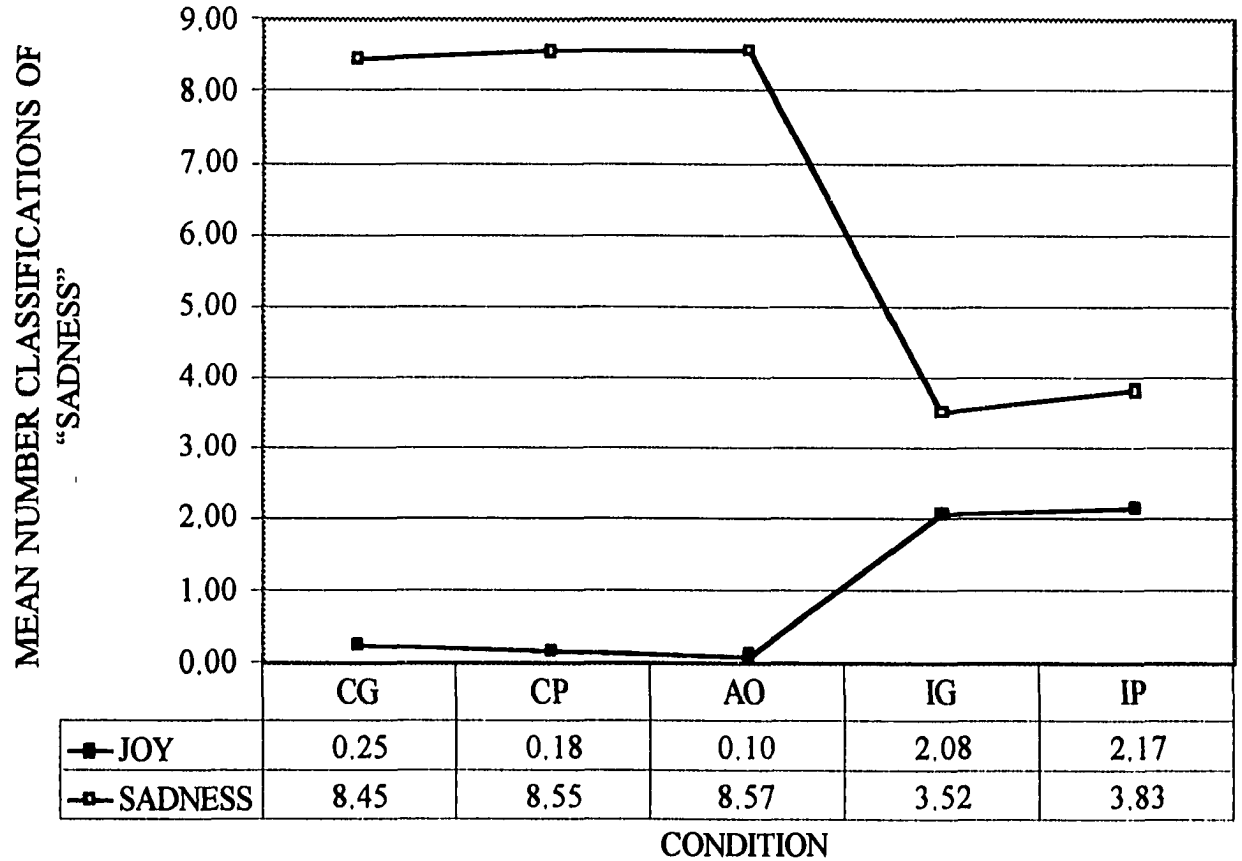
Mean number of classifications of "joy" across five conditions for auditory stimuli conveying joy and sadness.



	CG	CP	AO	IG	IP
JOY	6.12	6.33	4.32	1.37	1.08
SADNESS	0.12	0.12	0.30	3.27	3.25

Figure 7

Mean number of classifications of "sadness" across five conditions for auditory stimuli conveying joy and sadness.



	CG	CP	AO	IG	IP
JOY	0.25	0.18	0.10	2.08	2.17
SADNESS	8.45	8.55	8.57	3.52	3.83

Figure 8

Change in emotion category classifications of auditory stimuli conveying joy and sadness across auditory-visual conditions.

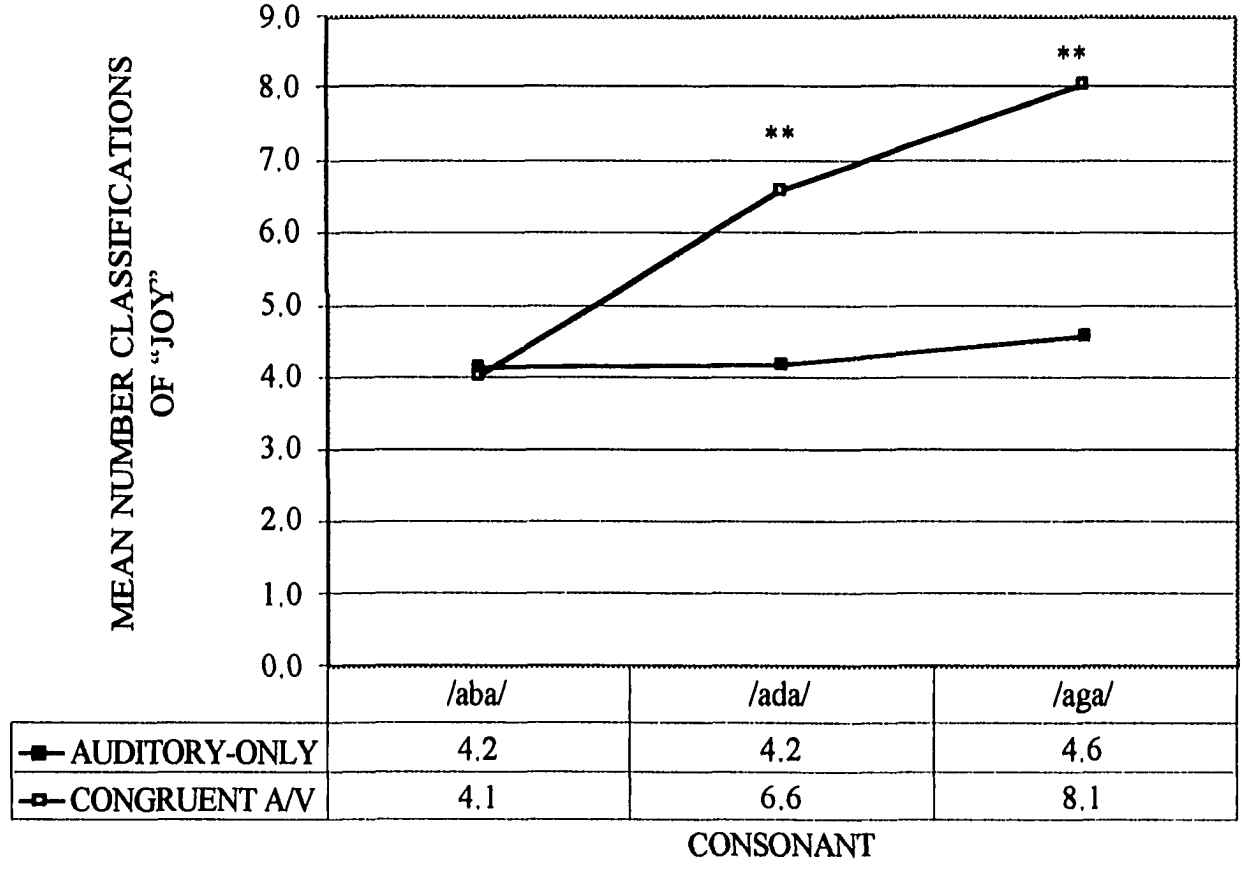
	AC	JO	EX	SU	AN	DI	SA	FE
JOY	1.43	5.00	0.07	2.98	-4.77	-2.08	-1.94	-0.68
SADNESS	-2.78	-3.14	-1.08	-1.77	1.37	1.37	4.83	1.20

EMOTION

Note. Difference scores were calculated by subtracting the mean number of emotion classifications of the emotionally incongruent auditory-visual conditions for each emotion from the main number of emotion classifications of the emotionally congruent auditory-visual conditions. AC = acceptance; JO = joy; EX = expectancy; SU = surprise; AN = anger; DI = disgust; SA = sadness; FE = fear.

Figure 9

Condition x Consonant interaction for emotionally congruent auditory-visual and auditory-only stimuli conveying joy



Note. **Significantly higher at $p < .01$.

Figure 10

Condition x Consonant interaction for emotionally congruent auditory-visual and auditory-only stimuli conveying sadness

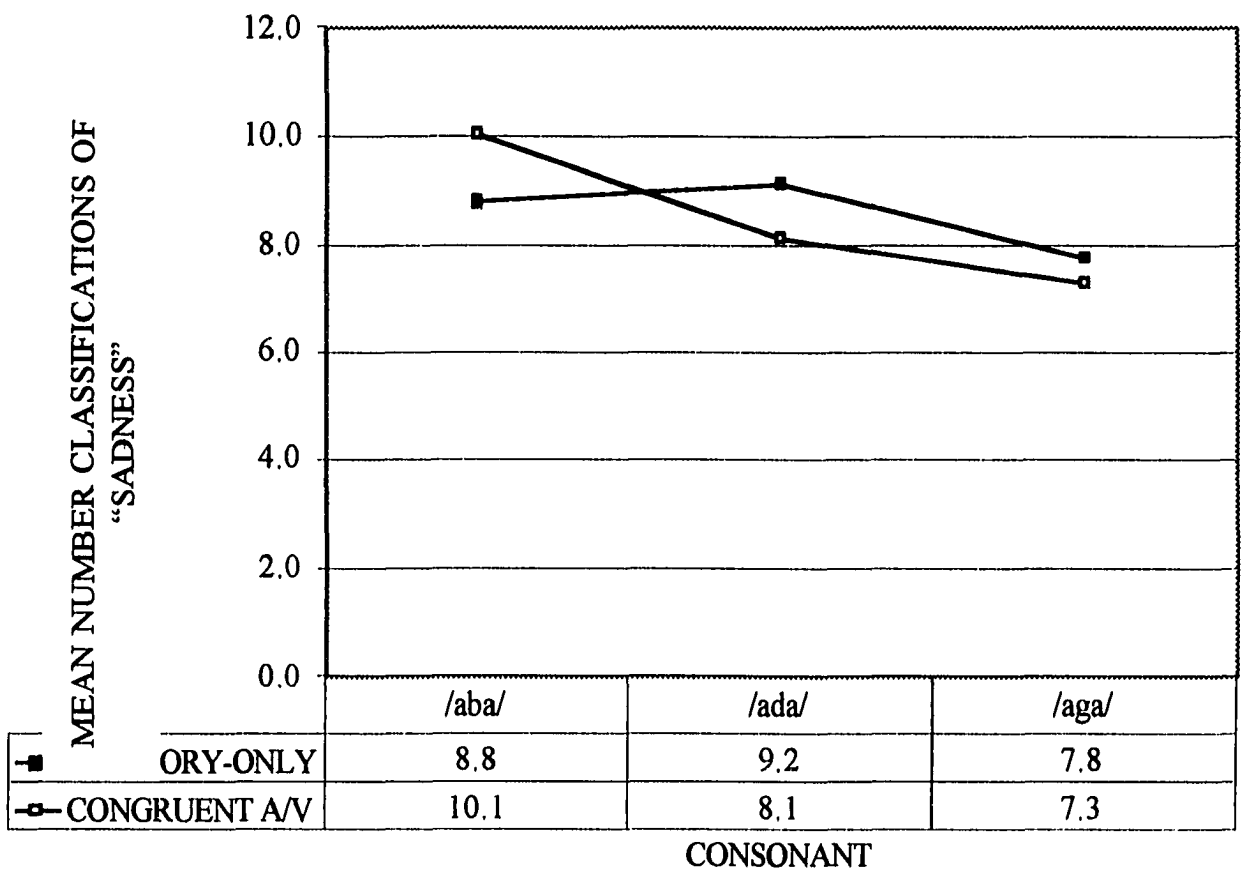
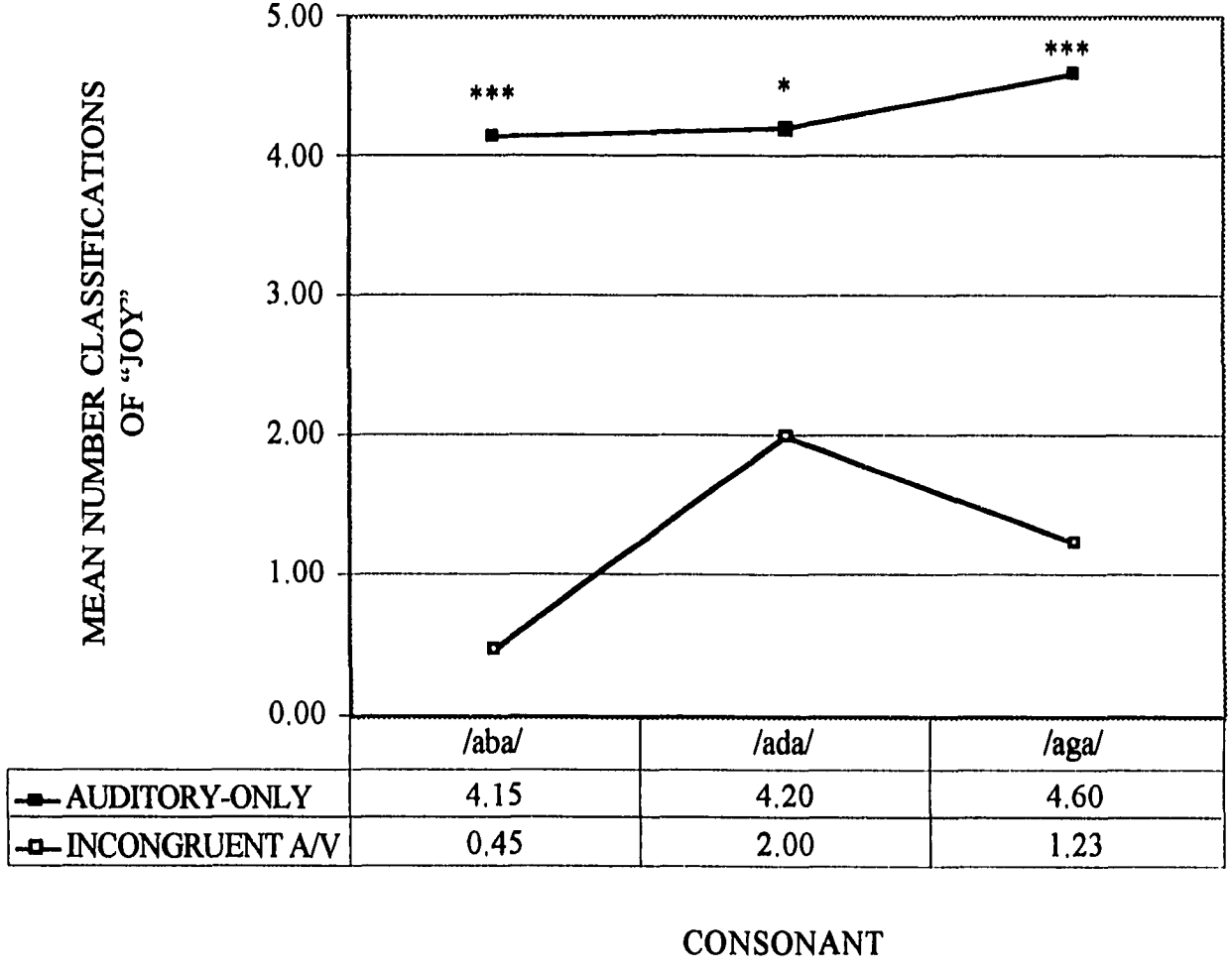


Figure 11

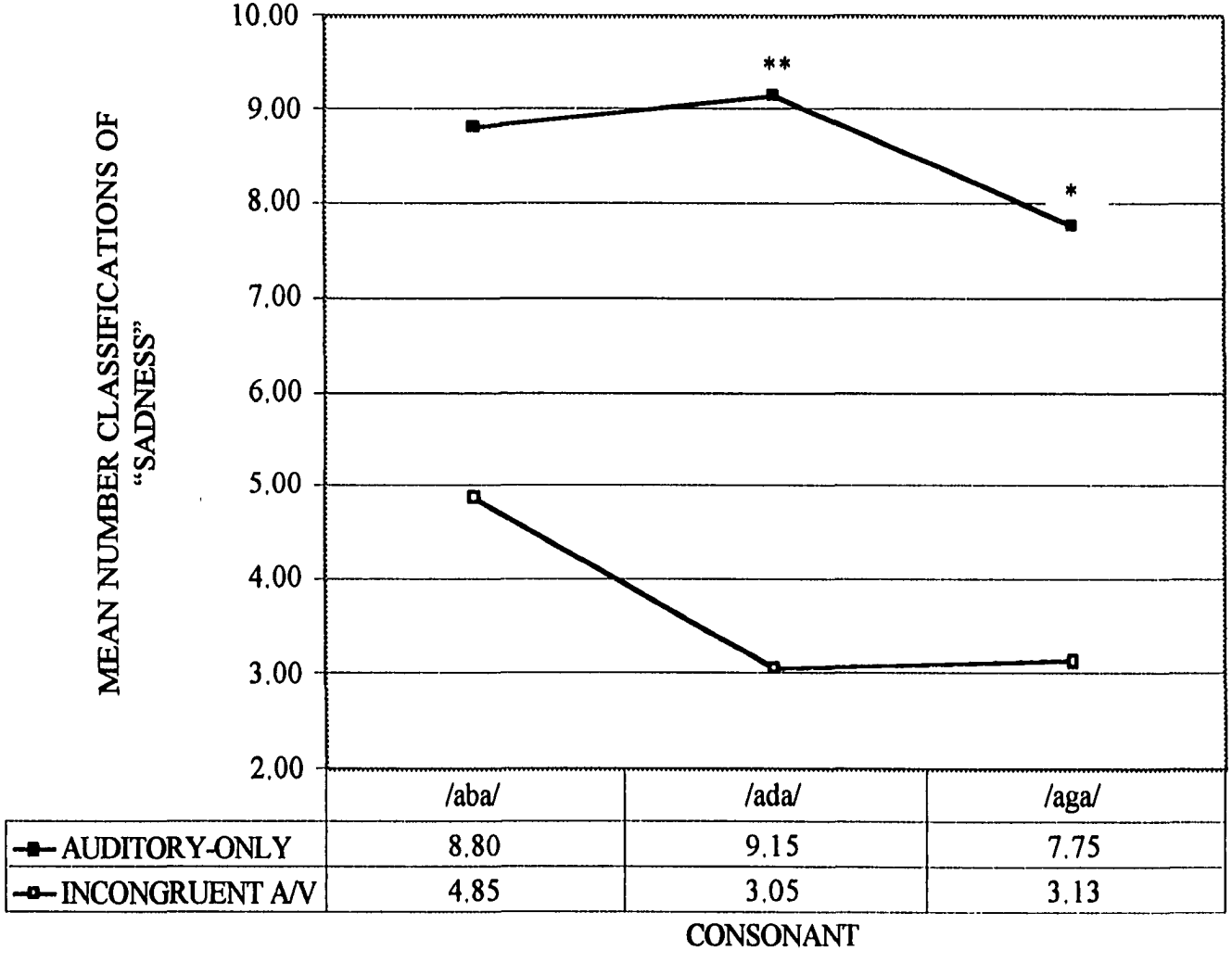
Condition x Consonant interaction for emotionally incongruent auditory-visual and auditory-only stimuli conveying joy



Note. *Significantly higher at $p < .05$; ***Significantly higher at $p < .001$.

Figure 12

Condition x Consonant interaction for emotionally incongruent auditory-visual and auditory-only stimuli conveying sadness



Note. *Significantly higher at $p < .05$; **Significantly higher at $p < .01$.

APPENDIX II

TABLES AND FIGURES RELATED TO

MULTIDIMENSIONAL SCALING PROCEDURES

Table 14

Derived Multidimensional Scaling Solutions for the Six Auditory-Only Categories

	Multidimensional Scaling Solutions			
Category	Number of Dimensions[a]	Stress Level[b]	R^2	Number of Iterations
/aba/ Joy	2	.13	.92	5
	1	.19	.89	4
/aba/ Sadness	2	.11	.93	9
	1	.30	.76	4
/ada/ Joy	3	.10	.95	10
	1	.25	.86	4
/ada/ Sadness	2	.07	.98	7
	1	.14	.95	12
/aga/ Joy	3	.10	.95	11
	1	.38	.77	3
/aga/ Sadness	2	.09	.97	5
	1	.29	.80	10

Note. [a]Values for one-dimensional solutions are included for comparison. [b]Stress values serve as a "badness-of-fit" measure where a value of 0 indicates perfect fit and a value of 1 indicates complete lack of fit.

Table 15

Goodness Ratings and Acoustic Characteristics of /aba/ Auditory Stimuli Conveying Joy

Stimulus	M	$F_{0(1)}$	$Time_1$	$F_{0(2)}$	$Time_2$	$F_{0(3)}$	$Time_3$	$F_{0(4)}$	$Time_4$	$F_{0(5)}$	$Time_5$	$F_{0(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
11[a]	5.6	225.0	18.4	275.6	182.9	262.5	191.4	143.2	343.1	121.1	567.1	139.6	639.4	334	337	671
2~1[b]	3.5	208.0	24.4	225.0	102.2	208.0	228.9	162.1	386.1	132.8	560.0	139.6	582.9	367	358	725
2	5.5	225.0	9.8	282.7	74.5	220.5	211.0	149.0	331.9	132.8	559.7	132.8	559.7	317	351	668
9	5.2	220.5	14.4	275.6	160.9	262.5	229.8	151.0	377.7	141.4	451.0	143.2	649.1	355	327	682
2~2	3.7	204.2	12.1	216.2	84.5	186.9	239.9	157.5	380.2	136.1	519.1	143.2	579.4	374	447	821
23	4.4	216.2	11.6	239.7	34.7	208.0	203.8	167.1	328.0	125.3	419.1	128.2	656.1	313	373	686
2~4	4.2	204.2	9.1	229.7	153.9	186.9	236.9	157.5	380.2	136.1	519.1	143.2	579.4	358	359	717
15	4.3	220.5	19.9	245.0	163.2	216.2	224.2	164.6	349.0	132.8	516.4	143.2	648.4	321	353	674
16	4.4	245.0	10.0	245.0	85.7	229.7	184.4	159.8	347.3	121.1	608.8	131.3	640.2	314	365	679
17	4.4	229.7	26.7	245.0	175.4	225.0	204.8	162.1	331.1	121.1	520.5	136.1	632.7	322	342	666

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as good representative of the category. [b]Stimulus endorsed as poor representative of the category.

Table 16

Goodness Ratings and Acoustic Characteristics of /ada/ Auditory Stimuli Conveying Joy

Stimulus	M	$F_{0\,(1)}$	$Time_1$	$F_{0\,(2)}$	$Time_2$	$F_{0\,(3)}$	$Time_3$	$F_{0\,(4)}$	$Time_4$	$F_{0\,(5)}$	$Time_5$	$F_{0\,(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
3[a]	5.4	234.6	25.9	268.9	137.9	220.5	202.1	149.0	364.5	126.7	427.3	136.1	475.0	308	341	649
29[b]	3.8	196.9	42.8	256.4	200.9	190.1	238.8	139.6	382.1	131.3	461.2	132.8	599.5	355	425	780
10	4.0	208.0	24.8	234.6	87.5	212.0	227.5	164.6	361.6	128.2	463.7	137.8	648.9	339	351	690
15	4.2	196.9	14.3	250.6	151.9	229.7	199.3	153.1	351.2	126.7	481.7	132.8	512.0	312	367	679
18	4.2	220.5	24.4	234.6	113.0	208.0	217.1	175.0	363.6	126.7	531.6	132.8	577.4	343	380	723
2	5.0	225.0	41.6	250.6	113.0	225.0	205.0	155.3	327.1	131.3	420.8	139.6	612.6	301	403	704
21	4.0	229.7	12.1	239.7	66.3	216.2	193.8	159.8	316.6	128.2	548.7	134.4	661.7	322	392	714
24	4.3	220.5	8.1	234.6	138.9	220.5	220.8	141.4	347.3	128.2	547.6	136.1	579.9	321	444	765
25	4.3	212.0	21.4	245.0	140.1	229.7	176.1	157.5	336.2	123.9	568.6	139.6	604.9	345	436	781
9	4.5	220.5	26.3	245.0	98.8	245.0	98.8	193.4	329.2	132.8	474.1	137.8	521.8	324	456	780

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as good representative of the category. [b]Stimulus endorsed as poor representative of the category.

Table 17

Goodness Ratings and Acoustic Characteristics of /aga/ Auditory Stimuli Conveying Joy

Stimulus	M	$F_{0\,(1)}$	$Time_1$	$F_{0\,(2)}$	$Time_2$	$F_{0\,(3)}$	$Time_3$	$F_{0\,(4)}$	$Time_4$	$F_{0\,(5)}$	$Time_5$	$F_{0\,(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
2[a]	6.2	204.2	64.7	256.4	202.3	220.5	275.1	153.1	399.7	118.5	616.5	126.7	697.4	357	409	766
11[b]	3.9	200.4	21.5	234.6	136.2	212.0	240.1	180.7	376.3	122.5	770.5	145.1	811.7	347	502	849
10	4.1	204.2	23.4	216.2	135.4	193.4	287.6	172.3	371.2	132.8	520.0	155.3	568.5	340	452	792
15	5.1	175.0	15.0	229.7	133.7	225.0	225.6	159.8	374.4	132.8	558.2	149.0	616.7	324	468	792
16	4.2	177.8	28.1	229.7	167.1	225.0	242.1	196.9	363.9	132.8	512.3	136.1	610.7	329	411	740
19	4.5	208.0	23.2	225.0	112.7	200.4	227.1	172.3	343.2	134.4	472.5	149.0	591.9	298	487	785
20	4.5	183.7	14.4	212.0	142.8	200.4	234.3	162.1	335.4	136.1	526.4	159.8	601.9	303	457	760
21	4.2	190.1	27.7	220.5	184.3	196.9	215.3	180.7	340.9	136.1	440.5	147.0	649.3	309	464	773
8	4.1	196.9	32.2	225.0	243.1	208.0	273.7	177.8	399.3	136.1	681.1	143.2	692.4	371	390	761
9	4.3	175.0	40.5	220.5	208.5	220.5	208.5	159.8	326.9	134.4	479.9	139.6	523.4	286	424	710

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as good representative of the category. [b]Stimulus endorsed as poor representative of the category.

Table 18

Goodness Ratings and Acoustic Characteristics of /aba/ Auditory Stimuli Conveying Sadness

Stimulus	M	$F_{0\,(1)}$	$Time_1$	$F_{0\,(2)}$	$Time_2$	$F_{0\,(3)}$	$Time_3$	$F_{0\,(4)}$	$Time_4$	$F_{0\,(5)}$	$Time_5$	$F_{0\,(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
12[a]	5.2	186.9	46.6	196.9	52.7	177.8	261.2	162.1	449.6	139.6	846.5	145.1	911.3	430	527	957
3[b]	4.1	186.9	26.2	204.2	31.8	186.9	245.0	167.1	389.0	149.0	605.9	164.6	826.5	364	522	886
11	4.5	193.4	40.6	208.2	46.2	177.8	280.7	164.6	446.9	157.5	635.3	157.5	679.6	418	457	875
15	4.1	190.1	23.9	193.4	31.3	177.8	259.4	175.0	391.9	155.3	660.6	164.6	719.5	378	494	872
16	4.5	180.7	35.9	190.1	77.7	183.7	277.0	169.6	436.4	155.3	717.3	159.8	753.2	418	526	944
18	4.4	190.1	35.5	208.0	46.8	177.8	286.1	167.1	430.1	153.1	718.2	159.8	738.7	405	481	886
2	4.2	180.7	40.3	200.4	80.6	186.9	252.8	162.1	404.9	143.2	480.0	164.6	729.1	377	491	868
4	5.0	208.0	27.9	208.0	27.9	180.7	264.4	162.1	426.4	149.0	582.8	157.5	726.2	386	496	882
8	4.6	169.6	30.3	183.7	256.3	164.6	275.9	169.6	389.8	147.0	475.3	164.6	761.9	371	472	843
9	4.6	164.6	11.1	196.9	27.6	186.9	296.7	164.6	433.1	147.0	497.6	137.8	845.9	412	461	873

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as good representative of the category. [b]Stimulus endorsed as poor representative of the category.

Table 19

Goodness Ratings and Acoustic Characteristics of /ada/ Auditory Stimuli Conveying Sadness

Stimulus	M	$F_{0\,(1)}$	$Time_1$	$F_{0\,(2)}$	$Time_2$	$F_{0\,(3)}$	$Time_3$	$F_{0\,(4)}$	$Time_4$	$F_{0\,(5)}$	$Time_5$	$F_{0\,(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
1[a]	5.4	159.8	69.8	196.9	279.1	177.8	281.0	355.6	513.6	315.0	676.4	315.0	802.4	460	458	918
25[b]	3.7	196.9	21.2	204.2	44.5	190.1	286.1	190.1	540.4	162.1	779.9	172.3	858.3	505	499	1004
14	4.4	183.7	27.4	208.0	39.1	190.1	299.1	473.1	367.5	306.3	533.7	315.0	795.6	427	499	926
15	4.6	204.2	39.1	204.2	39.1	186.9	323.2	355.6	549.6	315.0	689.6	324.3	833.7	499	476	975
2	5.1	175.0	40.1	200.4	253.0	180.7	282.5	175.0	497.6	151.0	649.4	155.3	794.9	463	536	999
20	4.3	183.7	37.1	196.9	91.9	190.1	271.7	175.0	453.5	164.6	602.1	169.6	762.4	449	475	926
4	4.6	190.1	29.4	208.0	69.2	193.4	337.8	175.0	505.7	149.0	717.6	162.1	809.9	465	529	994
6	4.2	177.8	17.7	193.4	35.5	190.1	285.7	169.6	490.6	155.3	705.4	183.7	798.0	446	488	934
7	4.8	200.4	30.7	208.0	269.7	208.0	269.7	164.6	545.6	157.5	627.3	169.6	895.0	477	491	968
9	4.8	183.7	40.6	204.2	275.5	204.2	275.5	169.6	514.7	155.3	670.6	196.9	833.0	471	541	1012

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as poor representative of the category. [b]Stimulus endorsed as good representative of the category.

Table 20

Goodness Ratings and Acoustic Characteristics of /aga/ Auditory Stimuli Conveying Sadness

Stimulus	M	$F_{0\,(1)}$	$Time_1$	$F_{0\,(2)}$	$Time_2$	$F_{0\,(3)}$	$Time_3$	$F_{0\,(4)}$	$Time_4$	$F_{0\,(5)}$	$Time_5$	$F_{0\,(6)}$	$Time_6$	$Syll_1$	$Syll_2$	Dur_{tot}
26[a]	5.3	24.5	177.8	32.7	212.0	279.9	177.8	512.9	172.3	531.3	157.5	784.7	162.1	447	521	968
20[b]	3.3	21.0	172.3	68.3	190.1	262.7	175.0	413.3	167.1	663.8	149.0	667.3	164.6	379	451	830
11	4.0	32.9	183.7	92.8	196.9	278.5	193.4	443.0	164.6	719.6	159.8	773.7	169.6	394	511	905
12	4.3	34.1	180.7	225.0	196.9	243.0	193.4	439.9	169.6	466.0	157.5	799.4	177.8	383	569	952
15	4.5	19.3	216.2	140.7	282.7	254.5	190.1	435.7	172.3	614.9	159.8	703.6	167.1	381	532	913
21	3.3	20.5	186.9	31.7	193.4	247.6	175.0	394.7	164.6	662.9	153.1	692.7	162.1	353	529	882
22	3.7	36.9	183.7	46.1	193.4	237.9	180.7	387.2	157.5	577.2	153.1	689.7	157.5	340	434	774
23	3.5	36.4	183.7	61.3	196.9	264.4	183.7	419.7	175.0	448.4	155.3	808.6	177.8	386	522	908
24	4.0	9.6	190.1	54.0	193.4	235.5	183.7	403.4	164.6	430.4	153.1	727.6	175.0	352	562	914
3	4.0	8.9	190.1	74.6	204.2	246.8	177.8	381.7	169.6	554.0	151.0	683.6	159.8	349	492	841

Note. The first column presents mean goodness ratings (7 being the highest possible rating). Subsequent columns represent fundamental frequency (F_0) in kilohertz at six points in time (onset of first syllable, peak of first syllable, endpoint of first syllable, onset of second syllable, valley of second syllable, endpoint of second syllable). Also included is duration in milliseconds of first ($Syll_1$) and second syllables ($Syll_2$) and total duration (Dur_{tot})of the stimulus. [a]Stimulus endorsed as poor representative of the category. [b]Stimulus endorsed as good representative of the category.

Figure 1

Dissimilarity Ratings Matrix for Auditory /aba/ Stimuli Conveying Sadness

	11	12	15	16	18	2	3	4	8	9
11	0.0	-	-	-	-	-	-	-	-	-
12	2.1	0.0	-	-	-	-	-	-	-	-
15	2.0	2.3	0.0	-	-	-	-	-	-	-
16	2.6	2.1	1.9	0.0	-	-	-	-	-	-
18	3.2	4.0	3.3	2.8	0.0	-	-	-	-	-
2	2.2	2.7	2.2	2.8	4.0	0.0	-	-	-	-
3	3.4	3.4	2.4	3.2	4.2	2.0	0.0	-	-	-
4	2.6	3.2	3.0	3.0	4.0	2.2	2.4	0.0	-	-
8	2.2	2.8	3.3	3.6	4.5	2.3	2.9	3.3	0.0	-
9	1.7	2.2	2.5	2.0	3.7	1.7	3.1	2.6	2.2	0.0

Note. Dissimilarity ratings range from 0 (very similar) to 7 (very dissimilar).

Figure 2

Dissimilarity Ratings Matrix for Auditory /aga/ Stimuli Conveying Sadness

	11	**12**	**15**	**20**	**21**	**22**	**23**	**24**	**26**	**3**
11	0.0	-	-	-	-	-	-	-	-	-
12	1.5	0.0	-	-	-	-	-	-	-	-
15	3.1	2.5	0.0	-	-	-	-	-	-	-
20	3.2	3.1	4.7	0.0	-	-	-	-	-	-
21	2.1	2.6	4.4	1.8	0.0	-	-	-	-	-
22	2.5	2.4	4.0	2.1	1.6	0.0	-	-	-	-
23	1.4	1.5	3.6	2.2	2.3	2.3	0.0	-	-	-
24	1.8	1.9	3.5	2.7	2.0	2.2	1.7	0.0	-	-
26	2.8	2.6	4.2	3.7	3.1	3.5	3.0	2.9	0.0	-
3	2.0	1.7	3.1	3.1	2.3	2.4	1.7	1.9	3.1	0.0

Note. Dissimilarity ratings range from 0 (very similar) to 7 (very dissimilar).

Figure 3

Dissimilarity Ratings Matrix for Visual /aba/ Stimuli Conveying Sadness

	11	**8**	**4**	**3**	**2**	**18**	**16**	**15**	**12**	**9**
11	1.9	-	-	-	-	-	-	-	-	-
8	2.6	1.8	-	-	-	-	-	-	-	-
4	3.3	2.1	1.5	-	-	-	-	-	-	-
3	3.0	2.4	2.3	1.6	-	-	-	-	-	-
2	3.1	2.5	2.3	2.0	2.0	-	-	-	-	-
18	2.6	2.7	2.9	2.9	3.4	2.2	-	-	-	-
16	2.6	2.6	3.0	3.0	3.4	2.2	1.9	-	-	-
15	3.5	3.4	3.7	3.6	3.9	2.8	2.8	1.2	-	-
12	2.5	2.2	2.2	2.7	2.9	2.0	2.4	3.0	2.2	-
9	2.4	1.9	2.1	2.9	3.0	2.8	2.6	3.3	2.1	1.4

Note. Dissimilarity ratings range from 0 (very similar) to 7 (very dissimilar).

Figure 4

Dissimilarity Ratings Matrix for Visual /ada/ Stimuli Conveying Sadness

	1	**14**	**15**	**2**	**20**	**25**	**4**	**6**	**7**	**9**
1	1.3	-	-	-	-	-	-	-	-	-
14	3.6	1.1	-	-	-	-	-	-	-	-
15	3.6	1.5	1.5	-	-	-	-	-	-	-
2	1.5	2.8	3.2	1.3	-	-	-	-	-	-
20	3.9	1.4	1.5	3.2	0.9	-	-	-	-	-
25	3.5	1.7	1.9	2.8	1.3	1.0	-	-	-	-
4	3.8	3.3	3.4	3.3	3.5	4.1	1.1	-	-	-
6	1.9	3.1	3.3	1.7	3.4	3.4	2.8	1.4	-	-
7	1.5	3.6	3.6	1.5	3.5	3.7	3.3	1.5	1.2	-
9	1.9	3.7	3.8	1.7	3.7	3.4	3.2	1.4	1.2	1.0

Note. Dissimilarity ratings range from 0 (very similar) to 7 (very dissimilar).

Figure 5

Dissimilarity Ratings Matrix for Visual /aga/ Stimuli Conveying Sadness

	11	**12**	**15**	**20**	**21**	**22**	**23**	**24**	**26**	**3**
11	1.5	-	-	-	-	-	-	-	-	-
12	1.5	1.2	-	-	-	-	-	-	-	-
15	1.5	1.5	1.4	-	-	-	-	-	-	-
20	2.1	1.7	1.8	1.6	-	-	-	-	-	-
21	2.2	1.7	1.6	1.9	1.3	-	-	-	-	-
22	4.5	4.1	4.0	3.8	4.2	1.0	-	-	-	-
23	1.8	2.2	2.2	2.5	2.2	4.2	1.7	-	-	-
24	1.6	2.0	1.8	2.2	2.2	4.4	1.7	1.3	-	-
26	4.9	4.6	4.7	4.3	4.6	2.0	4.9	4.6	0.7	-
3	1.4	1.7	1.8	2.1	2.0	4.3	1.8	1.6	4.8	1.7

Note. Dissimilarity ratings range from 0 (very similar) to 7 (very dissimilar).

VITA

Graduate College
University of Nevada, Las Vegas

Teri J. Forrest

Permanent Address:
2182 Caravelle St.
Las Vegas, NV 89142-1748

Degrees:
Bachelor of Arts, Psychology, 1998
University of Nevada, Las Vegas

Publications:

Forrest, T.J., Allen, D., & van Kammen, D. (2000, November). Validity of facial affect processing tasks in schizophrenia. Poster session presented at the Annual Meeting of the National Academy of Neuropsychology, Orlando, FL.

Forrest, T., Bocchieri, L., & Meana, M. (2000, March). Expectations and beliefs about painful intercourse in college women. Paper presented at the Annual Conference of the Society for Sex Therapy and Research, Santa Rosa, CA.

Meana, M., Nunnink, S., Forrest, T., & Bocchieri, L. (2001, March). Perceived etiology of dyspareunia in college women. Poster session presented at Annual Meeting of the North American Society for Psychosocial Obstetrics and Gynecology, Waikoloa, HI.

Special Honors:
Graduate Student Research Award, UNLV Department of Psychology, April 2001

Thesis Title:
The Integration of Facial and Auditory Affect: An Emotional McGurk Effect?

Thesis Examination Committee:
Chairperson, Daniel N. Allen, Ph.D.
Committee Member, Michael D. Hall, Ph.D.
Committee Member, Murray G. Millar, Ph.D.
Graduate Faculty Representative, Alice J. Corkill, Ph.D.