COGNITIVE SCIENCE SOCIETY July 25-26, 2005, Stresa, Italy, pp. 38-44

TOWARD SOCIAL MECHANISMS OF ANDROID SCIENCE
A COGSCI 2005 WORKSHOP

Prosodic Alignment in Human-Computer Interaction

Noriko Suzuki (noriko@atr.jp)
Yasuhiro Katagiri (katagiri@atr.jp)
ATR Media Information Science Laboratories;
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288 JAPAN

Abstract

Androids replicating humans in forms need also to achieve replication in behaviors to realize high level of believability. It has been observed that people exhibit strong tendency to adjust to each other a number of speech and language features in human-human conversational interactions to obtain communication efficiency and emotional engagement. We investigate in this paper the phenomena of mutual alignment of speech characteristics in human-computer interactions, with particular focus on human to computer alignment of speech prosody features. We found that people exhibit one directional spontaneous short-term alignment of loudness and switching pause durations of their speech to computer produced speech, even without the visual presence of personified characters. We believe this phenomena of prosodic adaptation provides one of the key components for building empathy between humans and androids.

Introduction

The purpose of this study is to investigate how prosody, which includes such non-segmental features in voice as intonation and rhythm, affects human-computer interaction. Our research focuses on the nature of voice interaction and the activation of mutual alignment between a human and a computer or an android by the power of prosody in a computer-voice.

The progress of computer technology has changed the function of robots from a tool for people to use to an interactive partner. Robots have required more social or emotional aspects toward becoming an interactive partner with people. Some key review papers summarize the major finds (e.g., Duffy, 2002; Fong et al., 2003). As they mention, some studies have focused on anthropomorphism for improving the human-like qualities of embodied characters, including robots. One research pursued a reality of expressions for social robots (e.g., Breazeal, 1998). Other research examined minimal cues for inducing interpersonal behavior from user to robots in interaction (e.g., Ono, et al., 2001).

As one approach to establish rapport between people and computers including androids, we focus on the human tendency to synchronize with the speaking style of an interaction partner, especially at the prosodic level. As Pickering and Garrod mentioned, people tend to mirror the conversational behaviors used by their partners at multiple levels (Pickering & Garrod, 2004), from the phonetic representation to the situation level. The communication accommodation theory argues that alignment occurs during conversation in lexical, syntactic, and speech signal features between two people (Giles, et al., 1987). There has been little focus on this human tendency of alignment at the prosodic level in human-computer interaction, although there are many studies in human-human conversation. If this human tendency were applied to human-computer voice interaction, it might be an effective strategy for making interpersonal relations between a human and a computer.

In this paper, we describe a simple experiment to examine prosodic alignment from a human to a computer. We utilize speech amplitude and pause duration as parameters of voice prosody. The congruence of voice intensity between two people has been observed in conventional studies (Meltzer, et al., 1971; Natale, 1975), as has the congruence of pause duration between two people (Matarazzo, et al., 1967; Jaffe & Feldstein, 1970; Welkowitz & Kuc, 1973).

We believe prosodic alignment can be applied to human-computer interaction, even a computer slightly changes the prosody in its voice within a session. A computer may also have a prosodic effect on a participant even when it does not use an animated character as a conversational partner. We examine whether the user's prosodic features of speech amplitude and pause duration are influenced by changes in the these prosodic features in a system's output voice throughout a session. Our voice system changes voice prosody, i.e., amplitude or pause duration, within a session in these ways: (a) increasing, (b) constant, and (c) decreasing. We also assess whether prosodic alignment from a user to a system is bidirectional or increasing or decreasing according to prosodic changes in a system's voice.

Related Works

Prosodic alignment in human-human interaction

In everyday interpersonal conversation, people may consciously or unconsciously change their own voices to adjust the prosodic features in a partner's voice; participants may talk faster and louder as the discussion continues; caregivers with infants may speak more slowly and softly.

Some studies on human-human voice interaction have reported such a synchronous tendency at the prosodic level: utterance duration, pause, speech rate or vocal intensity. For utterance duration, Matarazzo and his

colleagues (1963) found a synchronous tendency that increased or decreased the mean speech duration in interviews as well as in conversations between astronauts and ground communicators (Matarazzo, et al., 1964). Jaffe and Feldstein (1970) stated the congruence of mean pause duration between interviewers and interviewees as well as Mattarazzo and his colleagues (1967). Moreover, Welkowitz and Kuc (1973) examined the congruence of mean switching pause duration in free style conversation between children as well as between university students. They also found correlations between the congruence of switching pause duration and positive evaluation to a partner. Webb (1972) observed that the mean speech rate of an interviewee changed to adjust the change in the interviewer's voice speed. For vocal intensity or speech amplitude, Natale (1975) reported the convergence of mean vocal intensity in both interview and unstructured conversations, while Meltzer et al. (1971) reported the synchrony of vocal amplitude in simultaneous utterances between two people. Couper-Kuhlen (1996) discussed prosodic repetition between two interlocutors.

From these results of conventional studies on human-human interaction, we believe that prosodic changes in a partner's voice have implicit and positive effects on an other's voice prosody through voice interaction. Moreover, we consider the possibility that prosodic alignment of amplitude and pause duration can be observed between a human and a computer through interactive sessions.

Prosodic alignment in human-computer interaction

Personality alignment from a user to a spoken dialogue system has been investigated for speech amplitude (Coulston, et al., 2002), speech latency (Darves, et al., 2002), and speech rate (Bell, et al., 2003). These conventional studies used animated characters as a conversational partner in a computer, which were implemented to output different TTS voices. In other words, each animated character output speech with different prosody.

Nass & Lee (2000) found that a user preferred a computer voice with a similar prosodic personality. They examined personality markers in TTS, i.e., speech rate, volume and pitch range, voices influenced users attitudes and behaviors. They observed that users regarded TTS voices that exhibit the same personality traits as themselves, for example, extroverted or introverted, as more attractive, credible, and informative. These results suggest that the macro level alignment of speech characteristics contributes to establish good social relationships between humans and computers through the disclosure of personality traits,

Suzuki et al. (2003) found both behavioral and psychological effects of voice prosody in human-computer interaction. They found that the micro level utterance-wise alignment of speech characteristics also contributes to fostering emotional engagement. They observed that prosodic mimicry by a computer of a user's voice increased the user's amicability for the computer.

These findings open up the possibility of utilizing the human capacity of voice prosody alignment within the context of spoken dialogue systems to attain both efficient and emotional engagement in interactions, just as in interpersonal conversations. To fully exploit this mutual alignment of prosody in human-computer interaction, we need to establish that humans actually exhibit micro level utterance-wise alignment of speech characteristics to computer voices. Suzuki et al. (2003) observed that some users returned the mimicry by producing similar humming voices with similar prosodic patterns as the preceding computer's voice. To further affirm the human tendency to align at the micro level to computer voice prosody, we conducted experiments in which we systematically manipulate prosodic changes in a computer's voice within one interactive spoken dialogue session. To focus on the effects of voice characteristics, the experiment was designed so that no animated character was used as a visible conversational partner.

Experimental Design

Hypothesis We assume that voice prosody in participants'responses, i.e., speech amplitude or pause duration, will align in the direction heard from the system within one interactive session. Specifically, it is predicted that the voice prosody of participant responses will be louder or longer when the system asks a question with a louder voice or after a longer pause duration, and smaller or shorter when asking a question with a smaller voice or after a shorter pause duration. As conventional studies of prosodic alignment in both interpersonal conversation (Matarazzo, et al., 1963; Natale, et al, 1975) and human-computer interaction (Coulston, et al, 2002; Darves, et al., 2002; Bell, et al., 2003), voice prosody of participants was influenced by changes in the same prosodic features in conversational partners between different conversational sessions. From these findings, the goal of this experiment was to investigate whether prosody in participants' responses during interaction with a system would align bidirectionally: they would be equally likely to align by increasing or decreasing voice prosody in questions from the system; The generality of results regarding prosodic alignment was examined across slight prosodic changes in the system, even though the same system was used to change voice prosody within one interactive session.

Experimental setting In this experiment, we conducted one session that included sixteen Q and A units. In other words, the first and the second half each consisted of eight Q and A units. The quiz system was constructed by a simple acoustic processing function without any speech recognition function that automatically output questions after detecting sound inputs via a microphone in a headset. Figure 1 illustrates the quiz system.

During data collection, answering voices from participants were recorded on both a hard disc recorder (digidesign: PROTOOLS) and a digital video cassette recorder (SONY: DSR-2000) via a headset (AKG:

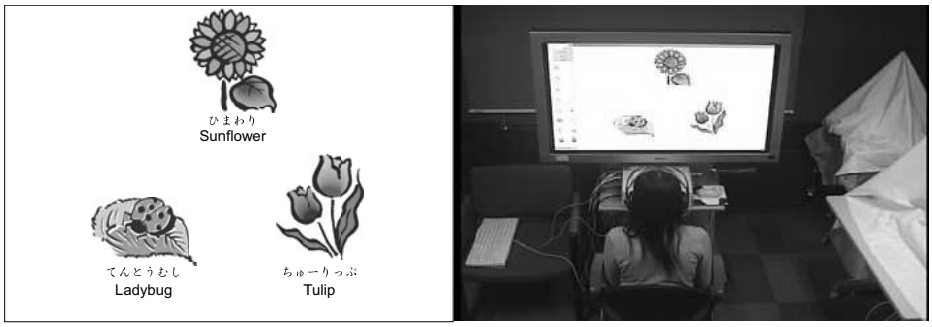

Figure 1: *Task: Multiple choice question, sample photo of the quiz (left), and an interactive experiment (right)*

Table 1: *Acoustic differences under three voice conditions*

Voice Type	Mean Amplitude (dB)	Mean Pitch (Hz)	Mean Mora Duration (mora/sec)	Pause Duration (sec)
STDV	64.5	125.3	0.14	2.10
LOUV	68.0	126.6	0.15	2.10
LPDV	64.3	122.8	0.14	2.40

STDV: Standard Voice, LOUV: Louder Voice, LPDV: Longer Pause Duration Voice

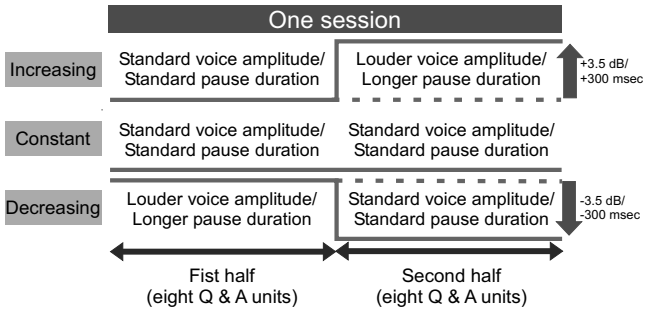

Figure 2: *Conditions: Increasing (top), Constant (middle), and Decreasing (bottom)*

HSO200).

The quiz system output recorded speech. The original voices were read by a male Japanese speaker and recorded in a soundproof studio in our laboratory by a digital audio tape recorder. He read six kinds of sentences for questions that were digitized at a sampling rate of 44.1 kHz.

In this experiment, a question voice with standard amplitude denotes the voice at recorded amplitude, while louder amplitude means +3 dB louder. The loudness of the louder amplitude voice was determined by the results of a perception test conducted before this experiment, which used 20 participants. A question voice with the longer pause duration denotes the voice with 300 msec longer pause duration.

Table 1 summarizes the differences in speech signal profiles between the standard and louder or longer pause duration voices.

Conditions We assigned the following three conditions for different prosody conditions of system voice, i.e., speech amplitude or pause duration (Fig. 2):

Increasing: In the first half of a session, the quiz system outputs a question voice with standard speech amplitude or after standard pause duration; then in the second half the system outputs a question voice with slightly louder or after a slightly longer pause duration.

Constant: In both the first and the second halves, the system outputs the question voice with standard amplitude or after standard pause duration. In other words, the system outputs the question at a constant loudness or pause duration throughout a session.

Decreasing: In the first half of a session, the quiz system outputs the question voice with a louder amplitude or after a longer pause duration; then in the second half the system outputs the question voice with standard amplitude or after standard pause duration in the second half.

Procedure Before starting a session, each participant received instructions while practicing with an experimenter how to interact in front of a 50 inch plasma display (Pioneer: PDK-50HW2). After, the experimenter left, the participants spent approximately five minutes alone in the soundproof room with the quiz system. During this time, participants answered 16 questions (e.g., the system asked "Which object does not belong?" while displaying a sunflower, a tulip, and a ladybug on the screen, in Fig. 1). After the session, the participants answered a questionnaire on their awareness of the slight changes of prosody in the system's question voice. Each participant was as-

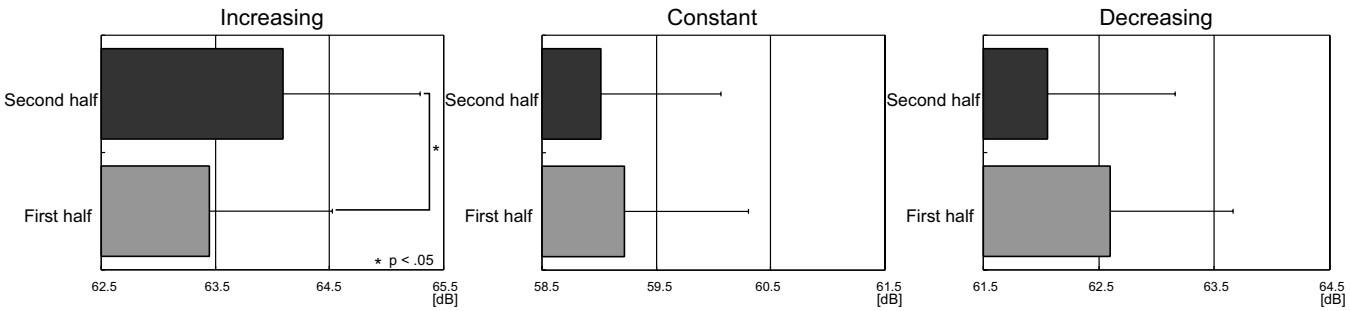

Figure 3: *Result 1: Voice amplitude differences in participants' voices*

signed to two sessions, one for different speech amplitude conditions and the other for different pause duration conditions. They had a short break between the two sessions.

Participants Thirty-nine undergraduate and graduate students from 18 to 25 years-old participated as paid volunteers. They were simply told that they would be participating in a study of how people work with computers when using their voices to accomplish task.

Participants were introduced to a simple quiz conducted via a voice system designed in a multiple choice format. The interface permitted participants to use speech input while it displayed a question by image and speech. There was no animated character on the computer screen.

Results

Speech amplitude

Figure 3 illustrates speech amplitude in the participants' responses in a session and compared amplitude in the first half of a session with the second half.

Three t-tests using within-subjects analysis were performed for each of the speech amplitude conditions of the system's questions. There was a significant difference between speech amplitude in participants' voices in the first half and in the second half under the increasing condition ($\underline{t = 2.30}$, $p < .05$). However, there was no significant difference between these two halves under the constant ($t = .57$, $p = .58$) or decreasing condition ($t = 1.68$, $p = .12$).

While participants may produce louder responses according to slight changes in the increasing speech amplitude in the system's question voice, participants do not always respond congruently with changes in both the increasing and decreasing directions. These results partly support our predictions.

From the results of post-session questionnaires, only two participants of participants in both the increasing and decreasing conditions answered that they were aware of the slight changes in speech amplitude in the system's question voice.

Pause duration

Figure 4 illustrates the pause duration in the participants' responses in a session and compared pause duration between the computer and participant voices in the first half of a session with the second half.

Three t-tests using within-subject analysis were performed for each of the pause duration conditions in the system's questions after the participants' answers. There was a significant difference between pause duration in the participants' voices in the first half and in the second half under the decreasing condition ($\underline{t = 2.87}$, $p < .05$). However, there was no significant difference between these two halves under the constant condition ($t = .38$, $p = .71$) or under the increasing condition ($t = .91$, $p = .38$).

While participants may produce responses after a shorter pause duration according to slight changes in the decreasing pause duration after the system's question voice, the participants do not always respond congruently to this change in both increasing and decreasing directions. These results also partly support our predictions.

From the results of post-session questionnaire, only one participant of the participants in the increasing condition and only two participants in the decreasing condition answered that they were aware of the slight changes in pause duration in the system's question voice.

Discussion

The above results partly support our prediction based on previous studies of prosodic alignment in interpersonal conversations (Matarazzo, et al., 1967; Jaffe & Feltstein, 1970; Welkowitz & Kuc, 1973; Meltzer, et al., 1971; Natale, 1975) and personality alignment in human-computer interaction (Darves, et al., 2002; Coulston, et al., 2002). In this paper, voice prosody in participants' responses during interaction did not clearly align bidirectionally to voice prosody in computer.

The main reason is explained by the parameter settings of prosody, voice amplitude, and pause duration, in the computer voice of this experiment. Both variation widths of prosodic parameter in computer voices were too slight to induce bidirectional prosodic alignment from participants. Therefore, they could align to the prosodic changes in the system unidirectionally,

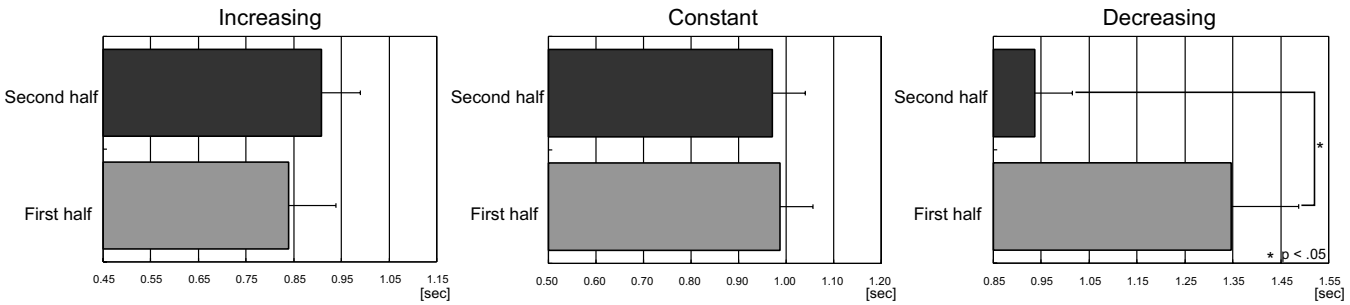

Figure 4: *Result 2: Pause duration differences between computer and participants*

which is easy to adjust for people, for either a louder voice or a shorter pause duration.

There are some theoretical and practical implications of these findings. The key theoretical implication is that prosodic alignment from a user to a spoken dialogue system is powerful and easy to manipulate; even in a simple form, it provides cues to construct mutual alignment rapport at the prosodic level between a human and a computer or an android. The authors have already argued that a computer character that mirrors user intonation patterns affect user attitudes toward the character (Suzuki, et al., 2003). We demonstrated that the more frequently the computer mirrored user intonation by humming, the higher the user evaluated the computer familiarity. We also observed mutual alignment at the prosodic level: the user mirrored the humming of a computer after it mirrored user intonation. From both these previous results and the current preliminary results, prosodic changes in a computer voice could induce prosodic alignment not only from human to computer but mutually. These results also suggest that mutual alignment would contribute a familiarity impression from the user to the computer or an android.

The second theoretical implication concerns the effect of autonomous function with interactiveness or responsiveness in a spoken dialogue system. Previous studies of personality alignment in human-computer interaction (Coulston, et al., 2002; Darves, et al., 2002; Bell, et al., 2003) found that clearer results of prosodic alignment from a user to a system are bidirectional according to the prosodic changes in the system voice; however, they used the Wizard of Oz method. This motivates us to further examine the effects of interactiveness or responsiveness on the direction of prosodic alignment from users to systems.

As a practical implication, this research shows that it is possible to increase speech recognition rates or enhance human-computer interaction by managing prosody in a user's voice. To the extent that androids including spoken dialogue systems, can exploit people's prosodic alignment tendencies to their conversational partner's speech, it may provide a simple but effective tool for error management; an android simply monitors prosodic changes by using acoustic processing and outputs speech by adaptively controlling prosody to guide a user's speech to a speech recognition area. It would also be possible to design an easy to interact android for system novices, including young children and senior citizens.

Conclusions

In this paper, we focused on a human behavior tendency, i.e., prosodic alignment, as a key toward establishing empathic relations between a human and a computer. If this human tendency were applied to human-computer voice interaction, it would also be an effective strategy for constructing mutual alignment at the prosodic level by subtly controlling human voice prosody.

We introduced preliminary results by using a simple experiment to examine prosodic alignment of people to a system according to slight prosodic changes; i.e., speech amplitude or pause duration, which were observed in interpersonal conversations. As a result, we found that participants' speech during interaction with the system aligned at least unidirectionally, even though they did not clearly adapted bidirectionally; while participants produced a louder voice according to slight increases in the speech amplitude of the system's voice, they did not produce a smaller voice according to slight decreases in speech amplitude in the system's voice. While participants produced a shorter pause duration according to a slight decrease in the pause duration in the system's voice, they did not produce a longer pause duration based on slight increases in the pause duration in the system's voice. We confirmed that people tend to align to slight prosodic changes in the system within a session, even when the system does not use any anthropomorphic functions with different voices based on personality.

As future work, we will examine the minimal parameter settings that induce bidirectionally in user prosodic alignment, e.g., both louder and smaller, and both longer and shorter, even without embodied characters. We will also highlight other prosodic features in voice, e.g., modulation and rhythm including pitch range or speech rate, to compare the human synchronous tendency in human-computer interaction with human-human conversations. We will further examine the effects of interactiveness or responsiveness on the direction of prosodic alignment from users to embodied characters including androids.

Acknowledgments

This research reported here was supported in part by a contract with the National Institute of Information and Communications Technology of Japan (NICT) entitled, "A study of innovational interaction media toward a coming high functioned network society".

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. *ICPhS2003*, 2453–2456.

Breazeal, C. F. (1998). Early experiments using motivations to regulate human-robot interaction. *AAAI fall symposium: emotional and intelligent. The tangled knot of cognition.*, 31–36.

Coulston, R., Oviatt, S., & Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. *ICSLP2002*, 2689–2692.

Darves, C., & Oviatt, S. (2002). Adaptation of users' spoken dialogue patterns in a conversational interface. *ICSLP2002*, 561–564.

Duffy, B. R. (2002). Anthropomorphism and the social robot. *IEEE/RSG IROS-2002*.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, *42(3-4)*, 143–166.

Giles, H., Mulac, A., Bradac, J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. In M.L. McLaughlin (Eds.), *Communication yearbook 10*, 13–48, Beverly Hills, CA: Sage.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York: Academic Press.

Matarazzo, J. D., & Wiens, A. N. (1967). Interviewer influence on durations of interviewee silence. *Experimental Research in Personality*, *2*, 56–69.

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. N. (1963). Interviewer influence on durations of interviewee speech. *Verbal Learning and Verbal Behavior*, *1*, 451–458.

Matarazzo, J. D., Wiens, A. N., Saslow, G., Dunham, R. M. & Voas, R. (1964). Speech durations of astronaut and ground communicator. *Science*, *143*, 148–150.

Meltzer, L., Morris, W. N., & Hayes, D. P. (1971). Interruption outcomes and vocal amplitude: explorations in social psychophysics. *Personality and Social Psychology*, *18(3)*, 392–402.

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are social actors. *CHI94*, 72–78.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? *CHI2000*, 329–336.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Personality and Social Psychology*, *32(5)*, 790–804.

Ono, T., Imai, M., & Ishiguro, H. (2001). A model of embodied communications with gestures between humans and robots. *CogSci2001*, 732–737.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Science*, *27*, 169–225.

Suzuki, N., Takeuchi, Y., Ishii, K., & Okada, M. (2003). Effects of echoic mimicry using hummed sounds on human-computer interaction. *Speech Communication*, *40(4)*, 559–573.

Webb, J. T. (1972). Interview synchrony: an investigation of two speech rate measures in an automated standardized interview. In A. A. Siegman & B. Pope (Eds.), *Studies in dyadic communication*. Pergamon Press.

Welkowitz, J., & Kuc, M. (1973). Interrelationships among warmth, genuineness, empathy and temporal speech patterns in interpersonal interaction. *Consulting and Clinical Psychology*, *41*, 472.