Acoustical Correlates of Affective Prosody

Kurt Hammerschmidt and Uwe Jürgens

Göttingen, Germany

Summary: The word "Anna" was spoken by 12 female and 11 male subjects with six different emotional expressions: "rage/hot anger," "despair/lamentation," "contempt/disgust," "joyful surprise," "voluptuous enjoyment/sensual satisfaction," and "affection/tenderness." In an acoustical analysis, 94 parameters were extracted from the speech samples and broken down by correlation analysis to 15 parameters entering subsequent statistical tests. The results show that each emotion can be characterized by a specific acoustic profile, differentiating that emotion significantly from all others. If aversive emotions are tested against hedonistic emotions as a group, it turns out that the best indicator of aversiveness is the ratio of peak frequency (frequency with the highest amplitude) to fundamental frequency, followed by the peak frequency, the percentage of time segments with nonharmonic structure ("noise"), frequency range within single time segments, and time of the maximum of the peak frequency within the utterance. Only the last parameter, however, codes aversiveness independent of the loudness of an utterance.

Key Words: Emotion—Vocal expression—Nonverbal behavior—Aversion.

INTRODUCTION

One and the same word pronounced in an angry, friendly, or fearful manner sounds different. Obviously, information in spoken language is coded not only in the verbal component, but also in a nonverbal component expressing the real or feigned emotional state of the speaker. Transcultural studies have shown that the specific affect underlying an emotionally spoken sentence can be identified correctly by a listener even if the language is unknown to him/her.[1–4] They also showed that the way specific affects are expressed acoustically is very similar across different cultures. The latter finding suggests that the vocal expression of emotion is, to some extent, innate. As there are similarities not only between different human cultures but also between humans and nonhuman primates,[5,6] we may conclude that the vocal expression of emotion has very deep-reaching phylogenetic roots.

The present study will give a detailed acoustic characterization of six emotional prosodies, using an acoustic analysis system (*LMA 9.2*) that allows the extraction of more than 90 parameters. Using this high number, a more differentiated description of emotional prosodies and their differences is possible than it was in previous studies, which used a much smaller set of parameters.[7–12] A second aim of the present study was to find out whether there are acoustic features distinguishing the vocal expression of aversive emotional states from that

Accepted for publication March 8, 2006.
From the German Primate Center, Göttingen, Germany.
Address correspondence and reprint requests to Uwe Jürgens, German Primate Center, Kellnerweg 4, 37077 Göttingen, Germany. E-mail: ujuerge@gwdg.de
Journal of Voice, Vol. 21, No. 5, pp. 531–540
0892-1997/$32.00

doi:10.1016/j.jvoice.2006.03.002

of pleasurable states. In other words, are there acoustic features common to various aversive states? The following aversive states were analyzed: ‘‘rage/hot anger,’’ ‘‘despair/lamentation,’’ and ‘‘contempt/disgust.’’ Hedonistic counterparts were ‘‘joyful surprise,’’ ‘‘voluptuous enjoyment/sensual satisfaction,’’ and ‘‘affection/tenderness.’’ All utterances used in the present study were simulated, not true, emotional expressions to guarantee recordings under standardized and, thus, comparable acoustic conditions. The utterances consisted of single words, instead of whole sentences. There were two reasons for this. First, we assumed that emotions are not expressed in each word of a sentence to the same degree. An acoustical analysis spanning a whole sentence, therefore, would ‘‘dilute’’ the acoustic characteristics of a specific emotion. Second, the present study was intended as a basis for future interspecies comparisons of the vocal expression of emotion between humans and nonhuman primates. As the vocalizations of nonhuman primates have a duration in the range of a short word, rather than a whole sentence, we decided to use single words.

METHOD

Recordings

Twenty-three students of dramatic art (12 women, 11 men) during their third year at the Academy of Performing Arts, Berlin, were asked to pronounce the name ‘‘Anna’’ with six different prosodies, expressing the following emotional states: (1) rage/hot anger, (2) despair/lamentation, (3) contempt/disgust, (4) joyful surprise, (5) affection/tenderness, and (6) voluptuous enjoyment/sensual satisfaction. Each prosody was produced twice by each speaker. Recordings were made on a digital audio tape (DAT) recorder (TCD-D8; Sony, Tokyo, Japan) with a condenser microphone (ECM-959 A; Sony) at a constant microphone-to-speaker distance (0.4 m).

After recording, the speech samples were played back to a group of 10 listeners (five women, five men; none of the listeners was identical with any of the speakers). This perception test served to make sure that the intended emotional expressions were, in fact, achieved. The test showed that each of the six emotions was recognized better than chance. Depending on the specific emotion, however, the hit rate varied between 43% (contempt/disgust) and 87% (rage/hot anger); the random probability was 16.7%.

Acoustic analysis

To extract acoustic parameters correlating with different emotions, we carried out a multiparametric analysis. First, we copied the speech samples from the DAT to a computer and transformed them to the WAV file format (44.1 kHz, 16 bit). We inspected the wave files for quality using *RTS 2.0* (Engineering Design, Belmont, MA). Samples of good quality were obtained for the 12 female and 11 male speakers for all six emotions, except one missing for contempt/disgust and two missing for voluptuous enjoyment/sensual satisfaction.

Amplitudes were measured in relative terms by calculating the amplitude of one sample relative to others, based on the logarithmic root mean square method, using *SIGNAL 3.0* (Engineering Design). For a better frequency resolution of the low-pitched male speech samples, we reduced the sampling frequency to 5500 Hz. The subsequent fast Fourier transformations (1024 points), made with *SIGNAL 3.0*, resulted in a frequency range of 2200 Hz and a frequency resolution of approximately 5 Hz. The time resolution was 5 milliseconds. The resulting frequency-time spectra were analyzed with *LMA 9.2* (developed by K.H.). *LMA* is a software tool to extract different sets of call parameters from acoustic signals.[13] In the present study, altogether 94 acoustic parameters were extracted from the speech samples.

One set of parameters (n = 18) related exclusively to tonal time segments. *LMA* determined the fundamental frequency (F0) by calculating an autocorrelation function of the frequency-amplitude spectra for each single time segment (segment length, 5 milliseconds). Depending on the number of peaks and periodicity of the autocorrelation function, each time segment was classified as noisy (no peaks could be detected), complex (some peaks could be detected, but they were not periodic), or tonal (periodic peaks). In case the time segment could be classified as tonal, we determined the value of F0. Additionally, we calculated for each