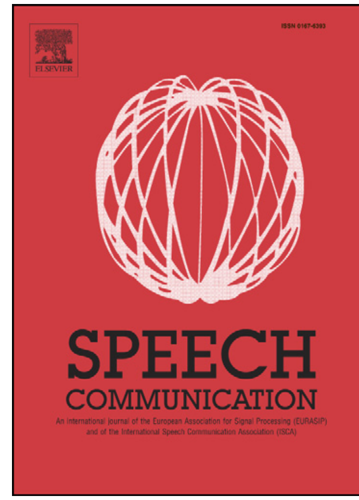
The Vocal Communication of Different Kinds of Smile

Amy Drahota, Alan Costall, Vasudevi Reddy

Accepted Manuscript



The Vocal Communication of Different Kinds of Smile

Amy Drahota, Alan Costall, Vasudevi Reddy

PII: S0167-6393(07)00173-2
DOI: 10.1016/j.specom.2007.10.001
Reference: SPECOM 1670

To appear in: *Speech Communication*

Received Date: 27 July 2007
Revised Date: 18 October 2007
Accepted Date: 19 October 2007

Please cite this article as: Drahota, A., Costall, A., Reddy, V., The Vocal Communication of Different Kinds of Smile, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.10.001

The Vocal Communication of Different Kinds of Smile[1]

Amy Drahota[a2], Alan Costall[a], and Vasudevi Reddy[a]

[a]Department of Psychology, University of Portsmouth, King Henry Building, King Henry 1[st] Street, Portsmouth, PO1 2DY, UK.

Correspondence concerning this article should be addressed to Amy Drahota, School of Health Sciences & Social Work, University of Portsmouth, James Watson Hall, 2 King Richard 1st Road, Portsmouth, Hampshire, PO1 2FR, UK. Telephone: +44 23 92 84 4432. Facsimile: +44 23 92 84 4402. E-mail: amy.drahota@port.ac.uk

[1]Portions of this work were presented in "The Auditory Discrimination of Socially Mediated Smiles," Proceedings of 10th European Conference: Facial Expression, Measurement and Meaning, Rimini, Italy, September 2003.

[2]Amy Drahota is now at the School of Health Sciences & Social Work, University of Portsmouth, James Watson Hall, 2 King Richard 1st Road, Portsmouth, Hampshire, PO1 2FR, UK

Abstract

The present study investigated the vocal communication of naturally occurring smiles. Verbal variation was controlled in the speech of 8 speakers by asking them to repeat the same sentence in response to a set sequence of 17 questions, intended to provoke reactions such as amusement, mild embarrassment, or just a neutral response. After coding for facial expressions, a sample of 64 utterances was chosen to represent Duchenne smiles, non-Duchenne smiles, suppressed smiles and non-smiles. These audio clips were used to test the discrimination skills of 11 listeners, who had to rely on vocal indicators to identify different types of smiles in speech. The study established that listeners can discriminate different smile types and further indicated that listeners utilize prototypical ideals to discern whether a person is smiling. Some acoustical cues appear to be taken by listeners as strong indicators of a smile, regardless of whether the speaker is actually smiling. Further investigations into listeners' prototypical ideals of vocal expressivity could prove worthwhile for voice synthesizing technology endeavoring to make computer-simulations more naturalistic.

Key words:

smiles; non-verbal communication; emotion; speech characteristics; auditory perception

The Vocal Communication of Different Kinds of Smile

1. Introduction

Affect is expressed throughout the body (Trevarthen & Malloch, 2000), and is detectable through different senses (De Gelder & Vroomen, 2000; Scherer et al., 1986). Therefore, different channels (or modes) of expression are likely to be involved in a single communicative act. The ability to discriminate audibly between vocal expressions of different categorical emotions has been found in different cultures and in some different languages (Scherer et al., 2001; Scherer & Wallbott, 1994). Investigating the auditory detection of facial expression is crucial not just for our understanding of perceptual processes but also because it could be vital for helping people with sensory deficits. People who are blind, for instance, may rely heavily on distinctions in emotional tone and cadence in voice in order to facilitate communication.

Although research on the vocal communication of affect is growing (Douglas-Cowie et al., 2003; Juslin & Laukka, 2001; Ladd et al., 1985), the data are often limited to the use of “unnatural” speech samples involving either synthetic manipulation or production by actors who have been asked to focus on “pure” and “intense” exemplars of emotional expression (Juslin & Laukka, 2003; Lieberman & Michaels, 1962; Scherer, 2003). This artificiality has been criticized in research on facial displays of affect (see Fernández-Dols & Ruiz-Belda, 1997). Keltner and Ekman (2000) argue that this concentration on “universal, prototypical facial expressions, … ignor(es)…individual variation in such expressions” (p. 244). It is equally problematic for vocal expressions of affect: the use of prototypical examples obscures the nuances that accompany the regulation and moderation of expression by individuals during social interaction. So despite the acknowledged centrality of multi-modal expressiveness in communication, there has been limited research on the vocal communication of regulated expressions of emotion.

The expression of regulated emotion, although little understood (Gross, 1998), is a regular feature of everyday life (Morris & Reilly, 1987). Some consider variations in the expression of the same categorical emotion to be the result of "pull effects" (Johnstone & Scherer, 2000; Scherer, 1985) or "display rules" (Ekman & Friesen, 1969; Kirouac & Hess, 1999), terms that describe the constraints governing socially acceptable expressions of emotion, e.g. having to stifle a yawn in a meeting, or hide a smile in a serious situation. Others, such as Fridlund (1991) in his ecological behavioral theory, argue that variations themselves constitute different social intentions and acts. In either case, people are continuously managing and regulating their emotional displays in the rapid ebb and flow of social exchange. These modulated expressions, rather than the prototypical displays, are the "normal" forms of emotional expression that need to be explored. Although a number of studies have explored display rules (Ceschi & Scherer, 2003; Levenson, 1994; Levenson, 2002), virtually nothing seems to be known about the display rules for vocalizations.

Smiles are ideal for exploring the effects of regulation on the vocal communication of affect, since they offer a rich source of natural modulation within interaction. Yet despite having been well researched as visual displays, much less investigation has been conducted into the vocal expression of smiles. Smiles can express a large variety of meanings, ranging from embarrassment to amusement, triumph, bitterness and even anger. Despite this, smiles are often just distinguished using the criterion of the activation of the orbicularis oculi muscles (i.e. the presence of "crows feet" wrinkles around the eyes), differentiating Duchenne smiles (DS) from non-Duchenne smiles (NDS). Often dubbed as "genuine" or "felt" smiles, DS have attracted much debate concerning how indicative they are of positive affect (e.g. LeFrance & Hecht, 1999; Zaalberg et al., 2004) and whether they are simply a more intense version of NDS (Messinger et al., 1997), which have been considered by some to be more indicative of feigned enjoyment (e.g. Ekman & Friesen, 1988). There are, however, many more subtle types of smiles – Ekman (2001, p. 127) claims that his Facial Action Coding System (FACS) can distinguish more than 50 different smiles, and at least some of these have been shown to involve different facial acts such as suppression and control (Keltner, 1995; Keltner & Buswell, 1997).

The existing findings about the effects of smiles on speech (Aubergé & Cathiard, 2003; Tartter, 1980; Tartter & Braun, 1994) have yet to be explored in conjunction with what we already know about the different social effects of smiles (Ekman & Friesen, 1988; Fridlund, 1991; LaFrance, & Hecht, 1999). Studies on how smiling affects vocalizations have typically focused on the acoustical effects of a mechanical smile gesture (e.g. Tartter & Braun, 1994) or amused smiles (Aubergé & Cathiard, 2003) and have not yet considered the vocal effects of other smiles or indeed suppressed smiles (SS). This suppression may be as evident in the voice as in the face, or even more so, given the speaker's greater awareness of facial actions over bodily and vocal displays (Ekman & Friesen, 1974).

In light of the literature, DS, NDS, and SS could either be considered distinctive categories (each with distinct social intentions) or on a dimensional scale of "smile intensity" (with DS being most "smiley" then NDS, and finally SS). What remains to be determined is whether differences between smiles (that have arisen either as a consequence of motivational requisites or as a result of affect intensity) have an influence on their vocal expressivity and auditory availability. The present study explores the distinction between naturally occurring DS, NDS, and SS, and whether these have implications for vocal accessibility. One of the major issues when conducting research into vocal communication is how to control for verbal content. Previous studies have overcome this issue by using acted portrayals or synthesized speech, which has resulted in a dearth of information on natural expressions (Juslin & Laukka, 2003). In order to improve on this methodology, the present study utilizes a novel interview technique designed to induce varying facial expressions whilst the speakers repetitively utter the same words. Ensuring utterances are standardized not only controls the verbal content, but also provides a platform from which to study both the encoding and decoding components of the communicative process (as called for by Juslin & Laukka, 2003).

2. Method

The present study investigated the acoustical basis for the discrimination of Duchenne Smiles (DS), Non-Duchenne Smiles (NDS), and Suppressed Smiles (SS) from No Smiles (NS). The study was conducted in three main stages: (1) inducing smiles in speakers to obtain the auditory stimuli, (2) coding and extracting appropriate utterances, and (3) the testing of perceptual discrimination.

2.1. Stage 1: Obtaining the Auditory Stimuli

2.1.1. Speakers

Eight native English adult speakers (three male, five female, aged 18 to 40 years old), with Southern English accents, participated in the recorded interviews. Two of the speakers wore glasses, the frames of which were of a suitable size and shape so not to interfere with coding of the orbicularis oculi muscles (B. Waller, qualified Facial Action Coding System coder, personal communication, October 2007). Speakers were recruited within the Psychology Department at the University of Portsmouth, and consisted of the first eight people to volunteer participation.

2.1.2. Procedure

One-to-one interviews were held with speakers in one interview room using a JVC Compact VHS video camera fixed on a tripod. The set-up of the interview room remained the same for all speakers; the distance between the speaker and the microphone (over a meter) was sufficiently large that any effects on the recording resulting from minor movements on the part of the speaker, would have been minimal. It was explained to speakers that participants in a future auditory discrimination task would be played a recording of their voice but would not see their face. Speakers were not told until after the interview that the focus of the future study was hearing different types of smiles. This information was temporarily withheld in an attempt to limit speakers' consciousness of their facial expressions. It was explained to the speakers that in order to conduct auditory analyses on the voice recordings, it was necessary for them to always reply to questions with the same phrase: "I do in the summer".

Speakers were asked to utter the words "I do in the summer" three times ("for the purposes of the recording") before the first interview question. This was asked in an attempt to obtain some "neutral" utterances in case the speaker found the whole interview amusing and so smiled throughout. A standard set of interview questions were read in the same order for each speaker. Questions (N=17) were designed to be neutral in some instances and induce feelings of amusement and potentially mild embarrassment in others. Examples include: "Do you ever sunbathe?", "Do you ever leave the house without a brolly?", and "Do you go skinny-dipping?"

(NB: “brolly” is a colloquial term for “umbrella” in England, and “skinny-dipping” is a coy term for “nude bathing”).

2.2. Stage 2: Coding the Auditory Stimuli

2.2.1. Coders

Two female coders viewed the videotapes, one of whom was a qualified Facial Action Coding System (FACS) coder (Ekman et al., 2002).

2.2.2. Equipment

Adobe Premier (version 6.0) was the editing software used to extract the speaker responses from the interviews and to view the clips for coding. Sound Forge (version 6.0) was the acoustical software used to extract the chosen utterances and convert them from a stereo format into a mono-wave format (in order to be compatible with PRAAT- the program used in the discrimination task).

2.2.3. Coding scheme

Clips were coded into four categories (DS, NDS, SS, and NS). The codes (as based on FACS) were as follows:

Duchenne Smile. DS involve both the raising of the lip corners (lip corner puller: AU12) and a contraction of the orbicularis oculi (eye corners) muscles (cheek raiser: AU6). The lip corners were coded as raised in a smile formation, only if they lasted longer than a phoneme and the movement had increased in intensity from the speakers’ “neutral” state. The contraction of the orbicularis oculi muscles (AU6), was the only identifying feature used to distinguish Duchenne from non-Duchenne smiles.

Non-Duchenne Smile. NDS involve the raising of the lip corners (AU12), with no eye muscle contraction. If a smile occurred with such a tiny indication of the Duchenne marker that it was barely visible, then it was coded as NDS with an asterisk (to indicate there is some minimal evidence of the Duchenne marker).

Suppressed Smile. SS involve "smile controls" entailing any movement of the lower face (apart from that required for enunciation) that counteracts the movement of the lip corner raise and cheek raise, such as: "lip corners depress" (AU15); "lip press" (AU24); "tongue show" (AU19); "jaw drop" (AU26); "mouth stretch" (AU27); "jaw thrust" (AU29); and "cheek puff" (AU34). Although the SS category encompasses a variety of different expressive movements, for the purposes of this study the movements were grouped together.

No Smile. Any action that does not involve smiling or smile controls constitutes NS. Again, this category may encompass a variety of different expressions.

Mixed categories. Due to the dynamic nature of the face it was likely that clips would contain a variety of expressions (despite their short nature). With this in mind the same clip could be coded into a mixture of the above categories, but with an indication of where the expressions occurred in each clip (beginning/end), or placing emphasis on the dominant expression. Coders attempted to categorize the clips into the four main categories (above) wherever possible, and used this category for clips where a decision could not be reached. Clips that were predominantly mixed categories were not included in the auditory perception task.

Additional. Extracts of the interview were deemed inappropriate for further use if they contained additional noises, such as movement of a chair.

2.2.4. Data Reduction

Following independent coding, coders met to discuss their choices and reach an agreement on which clips to use in the final study. Only clips that attained 100% agreement after discussion were entered in the main study of perceptual testing. An attempt was made to obtain two samples of each smile type from every speaker. However this was impossible for speaker “RA” due to the nature of her smiling; she contributed one NDS and three SS to the final corpus of clips. Each clip was converted into mono sounds (from stereo) and tapered by fading the beginning and end of each clip over a few milliseconds immediately before and after the words were spoken to create ‘smoother’ sounds; this process was to remove any sudden alarming ‘clicks’ of noise that may have occurred when each recording started and finished. This process (using Sound Forge) also served to remove laughter, which followed two clips (and was considered as confounding). The actual verbalization “I do in the summer” was not tapered (all words were present in all clips), so it is possible that residue of laughter (e.g. in breathing) remained within the spoken words for these two clips.

2.3. Stage 3: Auditory Perception Task

2.3.1. Design

The corpus of audio clips was played to each listener in a randomized order, and comprised 64 utterances obtained from eight speakers (with each speaker contributing eight utterances). Four different categorical expressions (DS, NDS, SS, and NS) were obtained from each speaker, with two examples of each (except in the case of speaker RA). Listeners had to code each clip in to one of three response categories (Open Smile, Suppressed Smile, No Smile).

2.3.2. Participants

Eleven native English-speaking listeners participated in the discrimination task (five female, six male; mean age = 34.1 years, *SD* = 12.5). All listeners volunteered their time and were recruited from the Department of Psychology, University of Portsmouth.

2.3.3. Equipment

The acoustical software PRAAT (version 4.1.1) was used to run the experiment on a computer monitor. Utterances were played through circumaural headphones and all apparatus settings remained the same for all listeners.

2.3.4. Procedure

Prior to the discrimination task, listeners were told what the experiment would entail and the meanings of the three multiple-choice response categories. An “open smile” was described as one in which the lips are drawn back, the teeth are showing and the person appears happy; a “suppressed smile” was described as one in which the speaker is trying to hide their smile by pulling their lips in or down as they speak, perhaps through embarrassment or because they want to disguise the fact they find something amusing; “no smile” was described as when the muscles around the face are relaxed and the person is speaking normally. Given the ongoing debate surrounding Duchenne Smiles (i.e. whether they are indicative of genuine smiles or simply a more intense smile), it was decided to use the term “open smile” to encapsulate both Duchenne and non-Duchenne smile types. This lay term enabled the use of simplified instructions to listeners, without having to make inferences of underlying meaning or resolve the Duchenne debate. Ekman (2001) also highlights the problem of asking people to distinguish between “genuine” and “false” smiles due to a general lack of understanding of the varieties of smile and the facial movements involved (p. 150). This approach additionally does not compromise the ability to assess how well listeners can distinguish the different smile types, which is retained in the analysis, as explained below.

The discrimination experiment was scripted in PRAAT, specifying that listeners would hear each clip once and clips would be presented in a random order. On hearing each clip listeners would have to click on one of a set of labeled rectangles (“Open Smile”, “Suppressed Smile”, “No Smile”). The next utterance would not be presented until the listener had responded. Three breaks were scheduled at regular intervals during the experiment (one after every 16 utterances).

2.4. Analysis

In analyzing the results we used a discrimination index to take into account response bias. This approach considers both the number of times a listener correctly classifies a stimulus (‘hits’) and the number of times a listener incorrectly uses that same classification for different stimuli (‘false alarms’). Due to the small number of trials, it was unsuitable to use “d-prime” (*d’*), the standard measure in Signal Detection Theory. In a review of discrimination indices, Swets (1986) recommends Yule’s Q as an appropriate measure for smaller data sets. Although not widely used in acoustics research, Yule’s Q is used within the context of Signal Detection Theory (e.g. van Puijenbroek et al., 2002). It has also been used as an association index between events and behaviors (e.g. Yoder & Feurer, 2000), and as an index in experiments in perception and memory (e.g. Hayman & Tulving, 1989; Levitin, 1994). To its advantage, this measure has, as explained below, a well-defined sampling distribution.

Listeners' abilities to discriminate DS from NS, NDS from NS, and SS from NS were tested by comparing their response choices to these pairs of stimuli. All response choices to each stimulus category under assessment were grouped into two-by-two contingency tables in order to calculate Yule's *Q* statistic. The response category of SS was grouped with OS (to compare OS+SS vs. NS responses). Because *Q*-values (like *d'*) become infinite if there is a zero in any of the cells in the contingency table, any zeros were replaced by one for the purpose of the analysis.

Using the respective cell values of each contingency table, Yule's *Q* was calculated on the number of hits (*h*), false alarms (*f*), misses (*m*), and correct rejections (*c*) for each participant for each contrasted pair of stimuli [$(hc - fm) / (hc + fm)$]. *Q* varies from –1.00 through 0 to 1.00, where a score of "1.00" implies perfect discrimination, a score of "0.00" implies no discrimination, and negative values imply that, although the stimuli were discriminated, they were consistently misidentified. The sampling distribution of Yule's *Q* is equivalent to that of Kendall's *S* statistic (Kendall, 1970) in the case of a dichotomy of ties on both variables. For the case of a two-by-two ordered contingency table $S = hc - fm$, and the standard deviation of *S* is given by the square root of $(t1 \times t2 \times u1 \times u2) / (n - 1)$, where $t1$, $t2$, $u1$, and $u2$ are the marginal totals derived from the contingency table (Jonckheere, 1970). The sampling distribution of *S* rapidly approximates the normal distribution, so that is possible to use the formula $z = (S -1) /$ (standard deviation of *S*) to derive *z* scores. *Z* scores were derived for individuals on each comparison, and to determine the statistical reliability of the overall ability of listeners, these were accumulated across individuals to derive an overall *z* score for each pair of stimuli (i.e. listeners' *z* scores were added together and divided by the square root of the number of listeners).

Additional analyses were made of the acoustical characteristics of the study clips, and compared to the respective discrimination scores for each clip, as well as the amount of times clips were categorized by listeners as NS. Acoustical analyses were run in PRAAT (version 4.1.1). Given that the verbal content was controlled for across clips (and therefore comprised similar formant trajectories), and there was less control over exactly where and for how long in each utterance the categorized facial expression occurred, it was decided to use aggregated acoustical measures, computed over the entire duration of each utterance (as opposed to selecting certain vowels for analysis). As such, it is possible that some "noise" may have been introduced into the findings presented here. These aggregated measures describing the overall utterance are therefore only indicative of the ways expressions may manifest vocally, and of the cues that listeners may have utilized during the experiment.

Pitch contours were manually assessed for octave jumps; one octave jump was identified, and the clip was rectified and saved as a new object in PRAAT for analysis. Mean pitch (fundamental frequency) was extracted from each utterance based on an autocorrelation method using an acoustic periodicity detection (Boersma, 1993). A time step of 0.01 seconds was used (e.g. 100 pitch values analyzed per second). Pitch range was calculated through extracting the maximum and minimum pitch of each utterance, identified via parabolic interpolation around each point of the pitch contour. Formant analysis was based on linear predictive coding (LPC), sampling each entire utterance and computed with an algorithm by Burg (see Press et al., 1992), with a visual check made for outliers. Distances between the first three formants were computed based on the aggregated mean values (e.g. F2/F3 difference = mean F3 - mean F2). Relative distances between successive formants was chosen as a measure in order to aid comparability across speakers, and as an indication of formant dispersion (distance between successive formants), which has been shown to correlate with shortening of the vocal tract (Fitch, 1997);

also a function of smiling. Clip duration was obtained by sampling the entire clip (including the 'tapered' start and finish, which was similar across clips).

3. Results and Discussion

3.1. How Listeners Assigned Utterances to Categories

Listeners' usage of the three response categories (Open Smile, Suppressed Smile, and No Smile) was reliably associated with the different stimulus categories, χ^2 (6, N = 704) = 79.8, $p <$ .001 (see Table 1). If the four stimulus categories are considered as falling along a dimension of "smileyness" (with DS being the most smiley, followed by NDS, then SS, and lastly NS), the use of the responses can also be seen to vary along this dimension. The "No Smile" response category is used *less* frequently as the stimuli categories become more "smiley", and the "Open Smile" response category is used *more* frequently as the stimuli categories become more "smiley". The "Suppressed Smile" response category is used most in response to SS and used progressively less to the other three stimulus categories as their "smileyness" decreases.

As can be seen in Table 1, the "No Smile" response category was utilized most frequently (N = 312 or 44.3%). This is interesting given that the corpus of utterances only comprised one quarter NS with the remaining three-quarters being some form of smile (SS, NDS, and DS). It is possible that listeners had a strong 'response bias' towards using the "No Smile" category when unsure, or else were simply not hearing vocal indicators for smiles in these instances.

3.2. Individual Listeners' Discrimination Indices

Listeners varied considerably in their ability to discriminate the various smile types from NS. As can be seen in Table 2, the majority of listeners showed a good discrimination between

DS and NS, with eight of these discriminations being statistically significant. Listeners "EC" and "MU" both obtained negative values for some of their indices, indicating that they were mislabeling the stimuli. Listener MU in particular scored consistently low in each of the three contrasts, other listeners varied between contrasts (e.g. "BW" obtained a high *Q* value for her ability to discriminate NS from DS, but a very low discrimination score for the NS vs. SS contrast). Listener "KN" in particular, stands out as good at all three contrasts, achieving statistical significance in 'DS versus NS' and 'NDS versus NS' comparisons. Although listeners obtained some relatively high indices of discrimination for NDS and SS utterances, when taken on an individual basis, only one of these (Listener KN) was reliably better than chance.

Overall, listeners could successfully discriminate each of the smile types from NS (DS: mean $Q = .71$, $SD = .18$, $z = 7.75$, $p < .0001$; NDS: mean $Q = .25$, $SD = .28$, $z = 2.34$, $p = .019$; SS: mean $Q = .30$, $SD = .22$, $z = 2.89$, $p = .0039$). Listeners achieved significantly better scores for the 'DS versus NS' comparison than they did for the 'NDS versus NS' comparison ($z = 3.82$, $p = .0001$) or the 'SS versus NS' comparison ($z = 3.43$, $p = .0006$). As a whole, listeners scored no more for the 'SS versus NS' comparison than they did for the 'NDS versus NS' comparison ($z = -0.39$, *ns*). So listeners can discriminate SS from NS; they can also discriminate NDS from NS, and are even better at discriminating DS from NS. Listeners do make mistakes however, and these are explored in the following section.

3.3. Acoustical Analyses

A series of acoustical measures have been associated empirically with smiles (Frick, 1985) and others can be derived (in principle) from what is known about vocal tract formation and its associated correlates (Ladefoged, 1975). Vocal correlates of emotional expressions have been investigated in a large number of studies (Juslin & Laukka, 2003; Scherer, 2003). It has frequently been demonstrated that portrayals of emotion influence global descriptors such as mean fundamental frequency (pitch), pitch range, and formant frequencies. In the present study no statistically significant systematic differences in the acoustical measures (aggregated across utterances) were found between the various expression types. In the following analyses, the relationship between acoustic descriptors and listeners' abilities to successfully discriminate expressions are explored. For this, acoustical characteristics are related to an index of discrimination derived for each utterance. The indices compared the listeners' responses (NS vs. SS + OS) to each individual SS, NDS, and DS utterance ($N = 6$ for each speaker) with the combined responses to the NS stimuli ($N = 2$ for each speaker). It should be borne in mind that the findings presented here, relating acoustical characteristics to listeners' abilities to discriminate, use correlations and are therefore tentative in nature; causality can not be assumed but it is hoped that hypotheses may be generated for future research.

3.3.1. Do Differences in Mean Pitch aid Discrimination Accuracy?

Existing research, typically using actors to simulate expressions, has reported that smiles and expressions of happiness are linked to a general raise in pitch (Juslin & Laukka, 2003; Frick, 1985; Ohala, 1984). However, it has also been demonstrated that pitch is not necessary for smile discrimination as listeners can still identify smiles in whispered speech (Tartter & Braun, 1994). With the exception of one speaker (JG), $r(6) = .79$, $p < .01$, there were no significant correlations between mean pitch and discrimination accuracy for each of the smile types. This was the case for individual speakers as well as when all speakers were grouped together. However, mean pitch did appear to influence listeners' decision-making in the discrimination task, with increasing pitch correlating with fewer “No Smile” responses, r (64) = -.35, $p < .05$. Speakers, on the other hand, did not systematically change their mean pitch during smiling, hence the lack of correlations between mean pitch and discrimination accuracy. This finding is in concordance with Tartter and Braun (1994), who also found pitch increase to occur inconsistently during smiled speech. It can be concluded that although listeners may have been utilizing mean pitch as a cue to help identify smiles, the speakers did not utilize mean pitch (at least, not in the same way) as a means of expressing smiles vocally. This discrepancy between listeners’ expectations and speakers’ actual usage of acoustical cues inevitably results in listeners making some mistakes.

The acoustical cues that listeners utilize may be derived from prototypical ideals about what smiles are expected or believed to sound like. Listeners may have been forced to utilize these prototypical ideals because they were unfamiliar with the speakers. If the acoustical patterning of vocal affect is expressed in an individual-specific way (as these results seem to suggest, given that no systematic differences in acoustical measures were found between smile types across speakers) then the way speakers express themselves will not always parallel the ideals of listeners. Varying the degree of familiarity between speakers and listeners may be a suitable future line of enquiry to test this hypothesis.

3.3.2. Do Differences in Pitch Range aid Discrimination Accuracy?

Pitch range gives an indication of how monotone the pitch within each utterance appears, and has previously been shown to vary strongly with the degree of emotional arousal in acted emotional portrayals (Bänziger & Scherer, 2005). If the smile types were seen to be on a dimensional scale of intensity (as opposed to individual categories) it could be predicted that DS would have the larger ranges in pitch, then NDS, SS, and least of all NS . However, in the present study, no significant correlations were found between the pitch range of the various smile types (DS, NDS, and SS) and how well they were discriminated from NS. Listeners did not appear to utilize pitch range as a cue to help them identify smiles (there is no correlation between pitch range and the number of "No Smile" responses listeners gave), and no consistent patterns of pitch range were found across the various smile types.

In future, it may be worthwhile to compare pitch contours (i.e. how pitch changes through time) between the different smile types. Although beyond the scope of this study, some preliminary qualitative analyses, involving visual inspection and comparison of the pitch contours, suggest that there might be some very subtle indicative patterning for individual smile types that global measures (such as mean pitch and pitch range) will not capture.

3.3.3. Do Differences in Intensity and Clip Duration aid Discrimination Accuracy?

As a precaution, further analyses were conducted to check whether clips of higher intensity and longer duration provide more information on which listeners can base their judgments. Some variation in clip intensity may be accounted for by the distance from speakers' mouths to the microphone, which was not measured; however, the information on this variable is important, as it indicates how listeners perceived clips during the experiment. NS were generally shorter, and less intense than smiles (Table 3), although there was no statistically significant difference between smile types on these acoustical variables. There was, however, a positive correlation between intensity and discrimination accuracy, r (64) = .30, p = .003. Furthermore increased intensity correlated with fewer "No Smile" responses, r (64) = -.24, p = .007, indicating that listeners identified more intense utterances as being more smiley sounding. These correlations may indicate a) that more intense utterances better portray the underlying expression or b) that listeners associate less intense utterances with less emotionality and therefore consider them more likely to reflect a "No Smile" expression. No correlation was found between clip duration and discrimination accuracy, or number of "No Smile" responses. This is in line with previous findings which have demonstrated that people can hear smiles even from monosyllabic nonsense words (Tartter, 1980; Tartter & Braun, 1994), indicating that smile discrimination is still possible in clips of short duration.

3.3.4. Comparison of the Perception and Formant Frequencies of Individual Speakers' Voices

Formant frequencies were analyzed as they provide an audible cue of vocal tract length (Fitch, 1997), and potentially, smiling. One interesting observation is that certain speakers sounded more "smiley" or "straight-faced" than others. Speakers "BG", "ML", and "FB" all received a high proportion of "No Smile" responses (NS responses = 77.3%, 75.0%, and 72.7% respectively), which is surprising given that only 25% of each speaker's utterances were actually NS. These speakers were not audibly projecting their smiles in their speech. Speaker "KH", on the other hand was more often categorized in the "Open Smile" category (OS responses = 58.0%) than any of the others, and hardly ever sounded as though she was not smiling (NS responses = 8.0%). There is an interesting relationship between the qualities in the speakers' voices and how "smiley" they were perceived; the more "straight-faced" a speaker was perceived, i.e. the more times a speaker is heard as "not smiling", the *larger* the difference between the second and third formant frequencies (F2 and F3) of their voice, r (64) = .46, $p <$ .01. Additionally, the more times a speaker is heard as "not smiling" the *smaller* the difference between F1 and F2, r (64) = -.37, $p < .01$. This finding supports the theory of prototypical ideals;

listeners appear to utilize differences between the formants to try to discriminate smiled from non-smiled speech, even though this did not necessarily parallel the acoustical correlates of actual smiles and non-smiles and hence resulted in errors of judgment.

3.3.5. Do the Differences between the Formant Frequencies aid Discrimination Accuracy?

There was no statistically significant relationship between the mean difference of the first and second formants (F1 and F2) and discrimination accuracy, and the differences between these formants did not systematically vary with smile type. However, as the difference between the F2 and F3 increases, the discrimination accuracy of listeners decreases; this correlation is significant for the discrimination of DS from NS, r (16) = -.56, $p < .01$, and NDS from NS, r (15) = -.53, $p < .01$, but not for the discrimination of SS from NS, r (17) = -.01, *ns*. These correlations suggest that the further apart F2 and F3, the more smiles sound like "No Smiles". The mean differences in F2 and F3 did gradually increase from DS utterances to NDS through SS to NS (mean difference = 1195Hz, 1234Hz, 1238Hz, 1239Hz, respectively), however these differences are not statistically significant, F (3, 60) = 1.01, *ns*. Listeners perceive voices with small differences between F2 and F3 as more smiley-sounding; and although the formants do not necessarily converge when speakers smile, when they do the smile is better discriminated.

This finding is intriguing since previous research suggests that lengthening of the vocal tract will lead to a decrease in formant dispersion (Fitch, 1997), and smiling should in effect shorten the vocal tract leading to increase in formant dispersion. In line with the notion of shortened vocal tract, larger differences between F1 and F2 seem to relate to more smiley sounding voices, even though the differences between these formants did not physically relate to the actual expression types. For F2 and F3 differences, a convergence related to more smiley sounding voices, but again no actual systematic differences between expression types were obtained. It is strange that the better discriminated smiles had characteristics which, from a mechanical perspective, one would assume more indicative of non-smiled speech. On the other hand, Aubergé and Cathiard (2003) have also found no significant differences in formant repartition for amused, mechanical, and neutral stimuli, and Tartter and Braun (1994) , although they did not report on F3, found that smiled speech increased F2 frequency (which they attribute to vocal tract shortening). It is possible then, that audible smiles in the present study were characterized by an increase in F2 relative to F3, producing a convergence of formants.

Moreover, Aubergé and Cathiard (2003) found that for some speakers, F3 was lower in amused utterances when compared to neutral utterances, providing additional complimentary evidence for the findings reported here. If, as some postulate, F2 is more closely related to phone identity (which remained constant across utterances) and vocal tract length (which may have changed), and F3 is more related to prosodic features (such as tones of amusement), a hypothesized drop in F3 relative to the speakers' normal enunciation, would elicit a convergence of these formants (and even more so if F2 also increases with an evident smile), producing more amused sounding speech. The listeners in the present research could therefore have been deducing smiles from emotional cues in the voices of speakers, even if speakers were not actually smiling. Emotional cues have been associated with the dynamic variation of larynx height (independent of lip movement), which also serves to influence vocal tract length (Xu & Chuenwattanapranithi, 2007). Further research is paramount to discern the relative complexities of F1, F2, and F3 relationships during smiled speech, or if indeed the correlations evident here are a product of some other confounding.

4. Conclusions

The present research has demonstrated that listeners can, with varying degrees of success, hear different types of smiles in the voices of strangers in the absence of visual cues. Listeners are very good at discriminating 'Duchenne Smiles' from 'No Smiles'. They can also, to a lesser degree, successfully discriminate 'Non-Duchenne Smiles' from 'No Smiles', and 'Suppressed Smiles' from 'No Smiles'. These findings support previous research findings demonstrating that smiles *can* be communicated vocally (De Gelder & Vroomen, 2000; Tartter & Braun, 1994).

Listeners do make mistakes however, and appear to relate certain acoustical characteristics (possibly: increased pitch, increased intensity, increased F1 and F2 dispersion, and decreased in F2 and F3 dispersion) with more smiley sounding voices. Listeners have (potentially misguided) preconceptions of what smiles should sound like. This possibility raises serious questions about the use of actors in research. Actors too, like the listeners in the present study, may have prototypical ideals regarding emotional vocal expressions, and if anything these ideals are reinforced in society through the use of actors (on television, stage, and radio). Results obtained from studies utilizing actors should not, therefore, be taken as evidence of how vocalizations of affect are expressed in *everyday* usage, but more as illustrative of how vocal affect expression is *represented* by society.

Smiles can have many different meanings (Eibl-Eibesfeldt, 1970) and relate to different underlying emotions. An interesting research issue is the extent to which smiles can be discriminated as a function of the degree of emotion experienced by the speaker. Aubergé and Cathiard (2003) showed that amused smiles carry much more information than mechanical smiles and the present results offer some support for this hypothesis. Duchenne smiles are often deemed "genuine" and attributed to positive affect (Zaalberg et al., 2004), which may explain why, in the present study, they were best discriminated from non-smiles, and to a significantly better degree than non-Duchenne smiles. It is possible that a genuine emotion of amusement may have been portrayed in voices with Duchenne smiles, that was absent from the other expression types. If speakers were producing non-Duchenne smiles purely to serve an appeasement or politeness function as opposed to expressing an underlying emotion, then they may have portrayed fewer vocal cues to aid listeners in the characterization of the smile.

Following this emotion-specific view of acoustical patterning (e.g. Johnstone & Scherer, 2000), it could be possible that the suppressed smiles that were best identified in the present study displayed characteristics of embarrassment (enabling listeners to identify the associated smile). Theoretically, it could also be hypothesized that those that were best identified contained cues that arose from speakers' intentions to express a suppressed smile, in line with Behavioral Ecology Theory (Fridlund, 1994). Alternatively, the best discriminated suppressed smiles may have contained cues that, paradoxically, arose from a determination *not* to smile, in line with Ekman's Non-Verbal Leakage Theory (Ekman et al., 1980) and the observations of Cowie and Cornelius (2003). In order to assess these possibilities, future research should include ratings of emotions from speakers, using a validated technique, and relate these ratings to the discrimination accuracy of the smile types by listeners. Taking this line of enquiry, future research may also assess whether the motivational and emotional requisites of smiles (such as genuineness, appeasement, suppression, embarrassment) are better expressed in the voice or face, or if a combination of expressive modalities portray more than any modality alone ever can.

The implications of this research are vast; for example synthetic voices have a host of applications (e.g. 'embodied conversational agents' Cassell et al., 2000; computer games; text-to-speech technology). According to Burdick (2003), although synthetic speech programs have made huge advances in recent years, effective *emotional* speech programs have yet to be developed. Although the present computerized voices are clear, they lack the emotional qualities which make human speech so meaningful and naturalistic. In many respects, the present research findings have demonstrated some possible acoustical correlates that may help make synthetic speech sound more “smiley”.

Beyond the practical implications of these findings, the present research has highlighted the importance of assessing aspects of both the encoding and the decoding of vocal expressions, since one does not always reflect the other. It is hoped that the approach undertaken in the present study will encourage future researchers to delve deeper into understanding how daily communicative acts of regulated emotion are expressed and perceived.

Acknowledgements

With special thanks to all the participants who lent their voices and ears to this study. Thank you also to Dr. Bridget Waller, Dr. Gwenda Simmons, Monja Knoll, Dr. Paul Morris, and Dr. Maria Uther for their time and expertise.

References

Aubergé , V., Cathiard, M., 2003. Can we hear the prosody of smile? Speech Comm. 40, 87-97.

Bänziger, T., Scherer, K.R., 2005. The role of intonation in emotional expressions. Speech Comm. 46, 252-67.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Proc. Institute of Phonetic Sci. 17, 97-110. University of Amsterdam.

Burdick, A., 2003. The mathematics of…artificial speech. Discover, 24, 1.

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. 2000. Embodied Conversational Agents. The MIT Press, USA.

Ceschi G., Scherer, K.R., 2003. Children's ability to control the facial expression of laughter and smiling: knowledge and behaviour. Cognition and Emotion, 17, 385-411.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Comm. 40, 5-32.

De Gelder, B., Vroomen, J., 2000. The perception of emotions be ear and by eye. Cognition and Emotion, 14, 289-311.

Douglas-Cowie, E., Cowie, R., Campbell, N., 2003. Speech and emotion. Speech Comm. 40, 1-3.

Eibl-Eibesfeldt, I., 1970. Ethology: the biology of behaviour. Holt, Rinehart and Winston, New York.

Ekman, P., 2001. Telling lies: clues to deceit in the marketplace, politics, and marriage. W.W. Norton & Company, New York.

Ekman, P., Friesen, W.V., 1969. The repertoire of nonverbal behavior: categories, origins, usage and coding. Semiotica, 1, 49-98.

Ekman, P., Friesen, W.V., 1974. Detecting deception from the body or face. J. Pers. Soc. Psychol. 29, 288-298.

Ekman, P., Friesen, W.V., 1988. Smiles when lying. J. Pers. Soc. Psychol. 54, 414-420.

Ekman, P., Friesen, W.V., Ancoli, S., 1980. Facial signs of emotional experience. J. Pers. Soc. Psychol. 39, 1125-1134.

Ekman, P., Friesen, W.V., Hager, J.C., 2002. FACS. Research Nexus, Salt Lake City.

Fernández-Dols, J.M., Ruiz-Belda, M.A., 1997. Spontaneous facial behavior during intense emotional episodes: Artistic truth and optical truth, in: Russell, J.A., Fernández-Dols, J.M. (Eds.), The psychology of facial expression, Cambridge University Press, Paris, pp.255-273.

Fitch, W.T., 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J. Acoust. Soc. Amer., 102, 1213-1222.

Frank, M.G., Ekman, P., Friesen, W.V., 1997. Behavioural markers and recognizability of the smile of enjoyment, in: Ekman, P., Rosenburg E. (Eds.), What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, New York, pp. 217-242.

Frick, R.W., 1985. Communicating emotion: the role of prosodic features. Psychol. Bull., 97, 412-29.

Fridlund, A.J., 1991. Sociality of solitary smiling: potentiation by an implicit audience. J. Pers. Soc. Psychol. 60, 229-240.

Fridlund, A.J., 1994. Human facial expression: an evolutionary view, Acedemic Press, San Diego.

Gross, J.J., 1998. Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression and physiology. J. Pers. Soc. Psychol. 74, 224-237.

Hayman, C.G., Tulving, E., 1989. Contingent dissociation between recognition and fragment completion: the method of triangulation. J. Experimental Psychology: Learning, Memory, and Cognition, 15, 228-240.

Johnstone, T., Scherer, K.R., 2000. Vocal communication of emotion, in: Lewis, M., Haviland-Jones, J.M. (Eds.), Handbook of emotions, second ed. The Guilford Press, New York, pp. 220-235.

Jonckheere, A.R., 1970. Design and analysis; review. Techniques for ordered contingency tables, in: Riemersma, J.B.J., van der Meer, H.C. (Eds.), Formulation and assessment of statistical models for experimental psychology. Proc. of the NUFFIC international summer session in science at ‘Het Oude Hof’, NATO, The Hague, pp. 190-215.

Juslin, P.N., Laukka, P., 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion, 1, 381-412.

Juslin, P.N. Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull., 129, 770-814.

Keltner, D., 1995. Signs of appeasement: evidence for the distinct displays of embarrassment, amusement, and shame. J. Pers. Soc. Psychol., 68, 441-454.

Keltner, D., Buswell, B.N., 1997. Embarrassment: its distinct form and appeasement functions. Psychol. Bull., 122, 250-270.

Keltner, D., Ekman, P., 2000. Facial expression of emotion, In: Lewis M., Haviland-Jones, J.M. (Eds.), Handbook of emotions, second ed. The Guilford Press, New York, pp. 236-249.

Kendall, M.G., 1970. Rank correlation methods, fourth ed. Griffin, London.

Kirouac, G., Hess, U., 1999. Group membership and the decoding behaviour, in: Philippot, P., Feldman, R.S., Coats, E.J. (Eds.), The social context of nonverbal behaviour, Cambridge University Press, Cambridge, pp. 182-210.

Ladd, D.R., Silverman, K.E.A., Tolkmitt, F., Bergmann, G., Scherer, K.R., 1985. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. J. Acoust. Soc. Amer., 78, 435-444.

Ladefoged, P., 1975. A course in phonetics, Harcourt Brace Jovanovich, Inc., New York.

LaFrance, M., Hecht, M.A., 1999. Option or obligation to smile: the effects of power and gender on facial expression, in: Philippot, P., Feldman, R.S., Coats, E.J. (Eds.), The social context of nonverbal behavior, Cambridge University Press, Cambridge, pp. 45-70.

Levenson, R.W., 1994. Emotional control: variation and consequences, in: Ekman P., Davidson, R.J. (Eds.), The nature of emotion: fundamental questions, Oxford University Press, New York, pp. 273-279.

Levenson, R.W., 2002. Blood, sweat and fears: the autonomic organisation of emotion, in: Ekman, P., Davidson, R.J., Campos, J.J., de Waal, F.B.M. (Eds.), A New York Academy of Sciences conference: emotions inside out: 130 years after Darwin's the expression of the emotions in man and animals, The Rockefeller University, pp. 12.

Levitin, D.J., 1994. Absolute memory for musical pitch: evidence from the production of learned melodies. Percept. Psychophys. 56, 414-423.

Lieberman, P., Michaels, S.B., 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J. Acoust. Soc. Amer. 34, 922-927.

Messinger, D.S., Fogal, A., Dickson, K.L., 1997. A dynamic systems approach to infant facial action, in: Russel, J.A., Fernández-Dols, J.M. (Eds.), The psychology of facial expression, Cambridge University Press, Cambridge, pp. 205-226.

Morris, W.N., Reilly, N.P., 1987. Toward the self-regulation of mood: theory and research. Motivation and Emotion. 11, 215-249.

Ohala, J.J., 1984. An ethological perspective on common cross-language utilization of F0 of voice. Phonetica, 41, 1-16.

Press, W.H., Teukolsky, S.A., Vetterling W.T., Flannery, B.P., 1992, Numerical Recipes in C: the art of scientific computing, second ed, Cambridge University Press.

Scherer, K.R., 1985. Vocal affect signalling: a comparative approach. Adv. Study Behav. 15, 189-244.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Comm. 40, 227-256.

Scherer, K.R., Banse, R., Wallbott, H.G., 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross Cult. Psychol. 32, 76-92.

Scherer, K.R., Wallbott, H.G., 1994. Evidence for universality and cultural variation of differential emotion response patterning. J. Pers. Soc. Psychol. 66, 310-328.

Scherer, K.R., Wallbott, H.G., Summerfield, A.B., 1986. Experiencing emotion: a cross-cultural study. Cambridge University Press, Cambridge.

Swets, J.A., 1986. Indices of discrimination or diagnostic accuracy: their ROCS and implied models. Psychol. Bull. 99, 100-117.

Tartter, V.C., 1980. Happy talk: perceptual and acoustic effects of smiling on speech. Percept. Psychophys. 27, 24-27.

Tartter, V.C., Braun, D., 1994. Hearing smiles and frowns in normal and whisper registers. J. Acoust. Soc. Amer. 96, 2101-2107.

Trevarthen, C., Malloch, S., 2000. The dance of wellbeing: defining the musical therapeutic effect. Nordic J. Music Therapy, 9, 3-17.

van Puijenbroek, E.P., Bate, A., Leufkens, H.G., Lindquist, M., Orre, R., Egberts, A.C., 2002. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol. Drug Saf. 11, 3-10.

Xu, Y., Chuenwattanapranithi, S., 2007. Perceiving anger and joy in speech through the size code. 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 2105-2108.

Yoder, P.J., Feurer, I.D., 2000. Quantifying the magnitude of sequential association between events or behaviors, in: Thompson, T., Felce, D., Symons, F.J. Behavioral observation: technology and applications in developmental disabilities, Paul H Brooks Publishing, Baltimore, MD, US, pp. 317-333.

Zaalberg, R., Manstead, A.S.R., Fischer, A.H., 2004. Relations between emotions, display rules, social motives, and facial behaviour. Cognition and Emotion, 18, 183-207.

Table 1

Listeners' usage of response categories

	Listeners' response choice		
Stimulus category	Open Smile	Suppressed Smile	No Smile
Duchenne Smile	78 (11.1)	62 (8.8)	36 (5.1)
Non-Duchenne Smile	40 (5.7)	45 (6.4)	80 (11.4)
Suppressed Smile	35 (5.0)	66 (9.4)	86 (12.2)
No Smile	25 (3.6)	41 (5.8)	110 (15.6)
Total	178 (25.3)	214 (30.4)	312 (44.3)

Note. Each cell is a simple count (and %) of how often each response category was collectively used by listeners in response to each stimulus category.

Table 2

Discrimination indices (Q) and associated z scores for listeners' judgments of Duchenne Smile (DS), Non-Duchenne Smile (NDS), and Suppressed Smile (SS) vs. No Smile (NS) stimuli.

	DS vs. NS (a)		NDS vs. NS (b)		SS vs. NS (c)		*z* score differences[a]		
Listener	*Q*	*z*	*Q*	*z*	*Q*	*z*	a-b	a-c	b-c
EB	.81	2.78*	.19	0.48	.52	1.54	2.31*	1.24	-1.07
SL	.68	2.20*	.35	0.97	.36	1.02	1.24	1.18	-0.05
JO	.81	2.89*	.48	1.33	.57	1.72	1.56	1.17	-0.39
BW	.84	2.39*	.22	0.56	.05	0.13	1.82	2.26*	0.44
EC	.62	1.58	-.31	-0.89	-.08	-0.23	2.48*	1.82	-0.66
DB	.96	3.87*	.45	1.22	.35	0.95	2.66*	2.93*	0.27
KN	.86	3.12*	.71	2.27*	.62	1.91	0.85	1.20	0.36
SA	.63	1.81	.07	0.16	.29	0.82	1.65	0.99	-0.66
MU	.29	0.73	-.05	-0.16	.18	0.46	0.89	0.26	-0.63
JP	.69	2.23*	.35	0.99	.12	0.33	1.25	1.91	0.66
PM	.67	2.08*	.33	0.85	.35	0.95	1.23	1.13	-0.09
Mean *Q*	.71		.25		.30				
SD	.18		.28		.22				
Overall *z*		7.75*		2.34*		2.89*	3.82*	3.43*	-0.39

[a]*z* score differences compare listeners' discrimination success within one stimuli pair to another.

**p* < .05.

Table 3

Duration and intensity data by smile type

Smile type	Duration (s)	Intensity (dB)
Duchenne Smile	0.88	52.53
Non-Duchenne Smile	0.90	50.48
Suppressed Smile	0.87	50.58
No Smile	0.83	49.94