HEARING SMILES - PERCEPTUAL, ACOUSTIC AND PRODUCTION ASPECTS OF LABIAL SPREADING

Julie Robson and Janet MackenzieBeck
Queen Margaret University College

ABSTRACT

This paper discusses the role of labial spreading in the perception of happiness and investigates whether untrained judges perceive labial spread samples as more smiled than neutral samples.

The main experiment is a paired forced-response design perceptual experiment comparing neutral and labial spread voice quality samples. This is supplemented with articulatory measurements, which provide evidence of the differences in articulatory settings.

The study's findings are discussed with reference to current theories of affect behaviour.

1. EXPRESSION OF EMOTION

The expression and perception of emotion is a subject area which has fascinated scientists for many years. In his 1872 work Darwin [1] acknowledges the multi-modal nature of emotion, recognising the significance of the voice, as well as facial expressions and other aspects of non-verbal communication. Following this, researchers in disparate fields, including psychology, physiology, psychiatry, linguistics and speech science have studied emotions from their individual subject-perspectives.

The multi-modal approach is currently enjoying renewed attention. The work of Klaus Scherer and his colleagues, in particular, takes an inter-disciplinary psychology-phonetics approach to the subject. For example, Banse and Scherer [2] highlight the need for a convergence of the study of facial and vocal emotion expression. They recognise that many aspects of emotional expression are likely to have multi-modality. For example, the innervation of particular facial muscles may produce a facial expression, but will also affect the acoustic characteristics of the speaker's voice.

This multi-modality is extended in the Component Process Theory [3, 4]. In this theory emotion involves a 'temporary synchronization of all major subsystems of organismic functioning' throughout the duration of the emotional state.

Another approach is Ekman and Friesen's Non-verbal leakage approach [5]. This proposes, rather than the various modalities working together during emotional states, some elements may be suppressed, causing 'leakage' in other elements. For example, during deception, a facial expression may be manipulated suggesting one emotion, but the true emotion may be revealed through another modality, say, voice quality or posture.

Facial expression is thought to be one of the most important elements in the expression of emotion. Ekman and Oster [6] claim that 'In humans the face seems to be a richer and more dependable source of information about emotion than any other expressive modality'. In fact, many of the studies involving the perception of emotions from vocal cues present high levels of recognition, in some cases higher than those obtained from (at least static) visual stimuli.

Sixty years of research in this area shows that listeners are good at inferring affective states and speaker attitudes from vocal expression [2]. A typical recognition rate of at least four or five times that attributable to chance would be expected for a study involving six emotions [8, 9, 2]. Most studies have concentrated on the effects of emotion on pitch and intensity measures. Johnstone et al [9] and Banse and Scherer [2] constructed predictive acoustic profiles for a wide-ranging set of emotions.

One emotion which is often cited in the literature as being recognised relatively well from both visible and vocal cues is happiness or joy. From their studies of facial expression, Ekman et al [10] found happiness to have better recognition rates than any of the other emotions, as did Boucher and Ekman [11] with Bassili [12] placing it second after surprise. In vocal experiments van Bezooijen [7] found happiness to be recognized second best, after anger.

Most acoustic studies of happiness have concentrated on changes in F0, F0 range, contour and intensity measures [13]. A happy speaker will typically display an extended F0 range reflecting the animation and stimulation of the state. The mean F0 and intensity will both tend to be increased.

Bassili [12] states that 'because facial muscles are fixed in a certain spatial arrangement, the deformations of the elastic surface of the face to which they give rise during facial expression may be very informative in the recognition of emotions'. This statement, used to describe facial expressions, is somewhat reminiscent of Abercrombie's [14] phonetic description of an articulatory setting: 'tendency for the vocal apparatus to be subjected to a particular long-term muscular adjustment' or Laver's [15] description of voice quality settings. Where Bassili describes an 'upwards displacement of each side of the cheeks and mouth' for the facial expression accompanying happiness, Laver's 'horizontal expansion of the inter-labial space' is equivalent. Both refer to smiling, or labial spreading.

Tartter [16] and Tartter and Braun [17] looked specifically at the effects of smiling on speech, thereby combining facial expression with vocal cues to emotion. They recorded one set of speech samples from the usual speaking voices of their speakers, then gave the following instruction: 'stretch your lips into a smile but do not try to sound happy or inject happiness into your speech'. This instruction was intended to generate test data which differed from the neutral data only in the facial

expression of the speaker. The samples were presented to naive listeners in a forced-response task in which they were asked to identify which of each pair of samples sounded happier. Their results showed that listeners were able to identify the smiled speech for four out of the six speakers.

Such evidence suggests that listeners have an ability to recognize the acoustic consequence of the facial expression, smiling, as a vocal cue. However, these studies are limited: firstly, they involved only a small number of speakers and listeners. The main potential problem, however, is in the instruction to the speakers. They are asked to form a smile: an emotionally meaningful gesture and therefore, may find it difficult to disassociate the other vocal effects. Moreover, the consequence of producing a smile without the accompanying injection of happiness could be that the speakers hyper-correct for emotion, introducing an element of unwanted 'sadness' or 'boredom' into their speech.

The experiment which follows has the following aims: to discover whether neutral speech and labial spreading can be distinguished aurally; to discover whether naive listeners would perceive labial spreading as smiling; to discover whether speech with neutral affect will be perceived as smiled given the presence of labial spreading. Also, a preliminary attempt at an integration of the perceptual results with acoustic measurements and with measurements of the lip position is included.

2. METHOD

2.1 Methodological Design

The determining factor in the design of the perceptual part of the experiment was the aim of identifying whether there is a perceivable difference between the two sample groups - neutral and labial spreading. A matched-pair forced-response design was chosen. This design involved the listeners being exposed to pairs of matched samples and asked to make a choice for every pair. With this design there is a 50% chance of randomly selecting the target sample so very high levels of concurrence are required before conclusions may be drawn.

The advantages of this method are that listeners are more likely to be able to recognise what may be very subtle differences between the two samples in a pair, and that the level of training required is minimised.

2.2 Speakers

The speakers were eleven members of staff from the Department of Speech and Language Sciences at Queen Margaret University College, Edinburgh. Ages ranged from mid-twenties to mid-fifties. Whilst the speakers originate from a variety of linguistic communities, they have all worked in educational institutions and lived in the Edinburgh area for several years. Their accents, therefore, tend towards Standardised forms such as Standard Southern English or Standard Scottish English rather than more distinctive regional varieties.

The speakers were trained prior to the recordings in the production of non-neutral voice quality settings from Laver's [15] framework. For the neutral samples they spoke using their habitual voice quality and for the labial spreading they aimed for scalar degree three to four, where scalar degree four corresponds to the long term lip position comparable with that of cardinal vowel two /e/.

2.3 Recording procedure

Recordings took place in the soundproof booth in the Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh. The speakers wore a lapel microphone 25cm from their mouths which was connected to a Sony DAT recorder.

The speakers were prompted by target recordings of the required voice quality setting, which were played to them through headphones immediately before each recording.

Simultaneous video recordings were made using High U-Matic video at 25 frames per second.

2.4 Speech samples

In total there were 66 speech samples obtained by the eleven speakers reading three different sentences using neutral and labial spreading. The sentences were devised to include differing levels of the segments which have been suggested as being more susceptible in the perception of labial voice qualities [15]. The three sentences were: Will Anna tell the dirty lie to Milly? (sentence A), Will Asha sell the thirty lie to Missy? (sentence B) and Will Sasha sell the thirsty lice to Sissy? (sentence C).

Samples of each sentence were transferred onto audio tape in randomly ordered pairs of neutral and labial spread for each of the eleven speakers. There was a gap of one second between the two samples in each pair and four seconds between pairs. Three different tapes were produced so that any order effect could be kept to a minimum. On the first tape the order of the sentences was ABC, for tape two it was CAB and tape three was BCA.

2.5 Listeners

There were fifteen listeners, twelve female and three male, none of whom were phonetically or linguistically trained. They were between the ages of 22 and 50, were native speakers of English and with no known speech or hearing problems.

2.6 Procedure for perceptual task

The listening session took place in a quiet room. The tapes were played using a Marantz tape recorder. Listeners were split into three equal groups of five. Each group was assigned one of the three tapes.

The samples were presented to the listeners in a forced response type experiment. For each pair the listeners were asked to choose 'which sounds more like the speaker is smiling?'. This choice was then marked onto a simple response sheet. Speakers were asked to make a choice for every pair of samples even if the difference was perceived to be marginal or virtually indistinguishable.

In total the subjects heard 33 pairs comprising the three sentences spoken by the 11 speakers. Subjects were asked to try to make their judgments on the first time of hearing the samples. Although repeat hearings were permitted, this occurred infrequently.

2.7 Video measurements

Articulatory measurements were made from the video using a manual tracing method. The measurements reported in this paper are of the maximum horizontal and vertical dimensions of labial aperture across the three sentences for each speaker and setting. These measurements are intended to provide a crude indication as to the nature of the change in articulation due to the labial voice quality setting change and also help establish that this change was realised physically. More detailed articulatory measurements from this data-set will be discussed further in future publications.

2.8 Acoustic analysis

The acoustic analysis involved measuring values for F1, F2 and F3 during the stable portion of the vowel segments constant to the three sentences. These are /a/ from 'Anna', 'Asha' and 'Sasha', /ɛ/ from 'tell' and 'sell'; /ə/ from 'the' /ɜ/ from 'dirty', 'thirty' and 'thirsty' /ɪ/ and /i/ from 'Milly', 'Missy' and 'Sissy'.

3. RESULTS

3.1 Results of perceptual experiment

Table 1 shows the number of labial spread samples (S) and the number of neutral (N) samples identified as 'smiled' for each of the three sentences, and for each speaker.

Speaker	Sentence A		Sentence B		Sentence C	
	'S'	'N'	'S'	'N'	'S'	'N'
AW	13	2	12	3	15	0
LE	10	5	9	4	13	2
AM	11	4	15	0	15	0
FG	15	0	15	0	15	0
SM	14	1	14	1	13	2
JS	14	1	15	0	15	0
JB	10	5	13	2	10	4
WH	14	1	15	0	15	0
ED	15	0	14	1	14	1
NH	15	0	13	2	13	2
TH	13	2	15	0	6	9
TOTALS	**144**	**21**	**150**	**15**	**144**	**20**

Table 1. Rates of identification as 'smiled' for spread (S) and neutral (N) labial voice quality samples for the three sentences.

Overall the spread samples were identified as 'smiled' in 90% of cases. This can broken down as follows: 85% for sentence A, 94% for sentence B and 87% for sentence C. There was only one pair of samples for which the neutral sample was more often identified as spread. This was speaker TH, sentence C. The other pairs for this speaker provided typical results. There is no obvious explanation for this anomalous result although one suggestion, arising from the video data, will be discussed in section 4.1.

3.2 Results of video measurements

Table 2 shows the percentage differences in horizontal and vertical maximum dimensions of labial aperture between the neutral and the spread labial voice quality settings. A negative percentage value indicates a decrease from neutral to spread, with a positive value indicating a corresponding increase.

The measurements of maximum horizontal aperture for labial spreading are greater, for all eleven speakers, than for the neutral setting. This, as expected, reflects the main articulatory change: horizontal expansion. The changes in the vertical dimension are more varied. Seven of the speakers display decreases in the maximum vertical dimension of labial aperture. This indicates vertical constriction which, accompanying horizontal expansion, is described by Laver [15] as a commonly occurring articulatory configuration. Two speakers display no apparent change in the vertical dimension, with two speakers showing an increase from neutral to spread. This would indicate vertical expansion, which is less commonly combined with horizontal expansion.

Speaker	Horizontal maximum	Vertical maximum
AW	18	0
LE	10	-14
AM	23	10
FG	17	-13
SM	15.2	-18
JS	19	-20
JB	10	-17
WH	11.5	-27
ED	27	0
NH	8.7	-20
TH	10.9	17

Table 2. Percentage differences in labial dimensions between neutral and spread labial voice quality setting.

3.3 Acoustic results

For most speakers and in most vowel contexts the values of F2 and F3 were higher for the spread labial voice quality setting than the formant values for the corresponding neutral setting conditions, as expected. However, the results were not conclusive. Further work, to be published at a later date, will investigate the formants further using dynamic formant tracking rather than such limited static measurements.

4. DISCUSSION

4.1 Discussion of results

Despite the fact that, during debriefing, the listeners reported difficulties in distinguishing the smiled from the non-smiled samples, the perceptual results show a clear pattern. The recognition of the labial spread samples as smiled is extremely high at 90% indicating the listeners aptitude at identifying lip spreading as a vocal cue to the affective state associated with the facial expression, smiling.

The method of obtaining the data for this study, using the Laver [15] voice quality framework, allowed for the virtual isolation of this one element of emotional expression. The instructions to the trained speakers were anatomically defined with no reference to any intended emotion. The intention was that the pairs of samples presented to the listeners differed only in labial voice quality. There is a possibility that the very act of smiling (or even labial spreading) may produce an emotional

response which in turn will cue other physiological correlates, but, this aside, emotional content is minimised. This can be seen as a major advantage over studies such as Tartter [16] and Tartter and Braun [17] because emotionally loaded terminology is avoided.

It is important to avoid inflating the importance of the labial spreading facial expression in the identification of emotion in this experiment. The forced-response design of this experiment is such that the paired samples are presented in succession and the listeners judges which pair sounded most as if the speaker was smiling. The high levels of agreement observed from the results show that listeners could tell a difference under these conditions but it is unlikely that they would identify the labial spread samples as smiled so consistently if presented in isolation.

Looking at the results for the three different sentences it seems that the labial spread samples were identified as smiled slightly more often for sentence B than for sentences A and C. However, the difference between sentences B and C can be attributed to the anomalous result of TH (highlighted in section 3.1). The expected result would have been the highest recognition rate for sentence C with the highest incidence of 'susceptible' segments, (i.e. those segments for which the perceptual effects of labial spreading are most prominent) followed by B and with A having the lowest recognition rate.

It is interesting to note that one of the speakers for whom both horizontal and vertical expansion is displayed is TH, for whom the anomalous perceptual result was obtained. It could be that the vertical expansion altered the acoustic quality of sentence C in particular, where the high incidence of 'susceptible' segments might influence perceptual judgments. No firm assumptions can be made, especially given that the other speaker displaying this combination of horizontal and vertical expansion failed to follow the same pattern for her perceptual results.

4.2 Theoretical implications

This finding that information about affective state seems to be communicable through the vocal channel fits Scherer's multi-modal approach. In this case, the facial muscular configuration associated with happiness, smiling, seems to have a vocal correlate which is itself capable of conveying affective information.

The articulatory evidence from the video data provides the evidence that the labial spreading occurred. As the mouth and lower part of the face are often considered to be the dominant areas of the face in the expression of happiness [12], and are also amongst the fundamental organs of speech, it is not surprising that changes in the configuration of the lower part of the face (containing visual cues to emotion) will produce corresponding changes in the vocal channel.

The non-leakage approach is not supported so well by this experiment. The theory requires a level of independence between the modalities. It cannot support labial voice quality and facial expression as separate modalities as they are inter-dependent. The vocal aspect of labial spreading can be considered as a symptom of the facial expression, smiling. It would be impossible to suppress one whilst emphasising the other (either consciously or sub-consciously). This approach is more successful in explaining behaviour where there is a simple dichotomy, for example, between verbal and non-verbal channels.

5. CONCLUSIONS

The findings of this study show that lip spreading, characterised by horizontal labial expansion, and sometimes combined with vertical constriction can be distinguished from the neutral voice quality and will be perceived as smiling - the facial expression which accompanies happiness.

This helps to support the theory that the communication and perception of affective states involves multi-modal processes.

REFERENCES

[1] Darwin, C. 1872. *The expression of emotion in man and animals*. New York/London: Appleton.

[2] Banse, R. and Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614-636.

[3] Scherer, K R. 1984. On the nature and function of emotion: a component process approach in Scherer, K. R. and Ekman, P. (eds) *Approaches to emotion*, Hillsdale NJ: Erlbaum.

[4] Scherer, K. R. 1986. Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, 99, 143-165

[5] Ekman, P. and Friesen, W. V. 1969. Non-verbal leakage and cues to deception. *Psychiatry*, 32, 88-106.

[6] Ekman, P. and Oster, H. 1979. Facial expression of emotion. *Annual Review of Emotion*, 30, 527-545.

[7] van Bezooijen, R. 1984. *The characteristics and recognizability of vocal expression of emotions*, Drodrecht: Foris.

[8] Frick, R. W. 1985. Communicating emotion: the role of prosodic features. *Psychological Bulletin*, 97, 412-429.

[9] Johnstone, I. T., Banse, R. and Scherer, K.R. 1995. Acoustic profiles from prototypical vocal expressions of emotion. *Proc. of XIIIth International Congress of Phonetic Sciences, Stockholm*, 4, 2-5.

[10] Ekman, P., Friesen, W. V. and Ellsworth, P. C. 1972. *Emotion in the human face: Guidelines for research and an integration of findings*, New York: Pergamon Press.

[11] Boucher, J. D. and Ekman, P. 1975. Facial areas and emotional information. *Journal of Communication*, 25, 2, 21-29.

[12] Bassili, J. N. 1979. Emotion recognition: the role of facial movement and the relative importance of the upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37, 2049-2058.

[13] Williams, C. E. and Stevens, K. N. 1981. Emotions and speech: some acoustic correlates. *Journal of the Acoustical Society of America*, 52, 1238-1250.

[14] Abercrombie, D. 1967. *Elements of general phonetics*, Edinburgh: Edinburgh University Press.

[15] Laver, J. 1980. *The phonetic description of voice quality*, Cambridge: Cambridge University Press.

[16] Tartter, V. C. 1980. Happy talk: The perceptual and acoustic effects of smiling on speech. *Perception and Psychophysics* 27, 24-27.

[17] Tartter, V. C. and Braun, D. 1994. Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96, 4, 2101-2107.