CROSS-LANGUAGE VOCALISATION OF EMOTION: METHODOLOGICAL ISSUES

Alison Tickle
Department of Speech, University of Newcastle, UK

ABSTRACT

A major question in work on phonetic correlates of emotion is to what extent vocalisation of emotion is due to psycho-biological response mechanisms and is therefore quasi-universal and to what extent it is due to social convention. Cross-language research gives an angle in on this question. But methodologically this is a very difficult area and few studies have been conducted. This paper identifies methodological issues and describes a study which has been designed to address some of these problems. Nonsense utterances and quasi-universally recognised facial expressions of emotions are used. These help deal with translation, ethical problems in data collection, the trade-off between artificiality of data and consistency and the masking of verbal utterances whilst allowing any influence exerted by specific vowel qualities to be highlighted.

1. INTRODUCTION

The long-held, very common intercultural belief in the existence of innate aspects of human emotion was supported in 1872 by Darwin’s [1] observations of similar vocal and facial expressions of specific emotions by people of diverse cultures. He also found that phylogenetic continuity was evident in his comparative observation of animals, thereby suggesting biological and therefore evolutionary foundations of emotion: most 20th century theories of emotion also include the influence of and/or effect upon cognition, including consciousness, when considering human experience of emotion.

Much scientific thought in the first half of the 20th century regarded emotional expression as culturally determined. Research now tends to point towards both psycho-biological response mechanisms and cultural influence and a major challenge now lies in finding more evidence to support this claim and in attempting to distinguish these two influences. One area in which this research can be applied is in the analysis and treatment of emotional disorder.

This paper discusses methodological difficulties involved in investigating the phonetic correlates of psycho-biological response mechanisms (PBRMs) and suggests one possible way forward, describing a methodology for use in cross-cultural, including cross-language, comparison, which is being used in an ongoing study comparing English and Japanese. We should, however, bear in mind that universality is a likely but not definite indication of innateness.

Further research into the vocal channel will also allow more comparison between response mechanisms along different channels. It may well be possible to connect acoustic and auditory parameters reflecting vocal indicators to parameters used to measure indicators of emotion in other channels (face and body). This will require consideration of the parts played by the somatic nervous system (SNS) and the autonomic nervous system (ANS) in the expression of emotion along different channels.

Research into the facial expression of emotion, comparing cross-cultural recognition - [2, 3] are two major studies in this field - has suggested that there are universal tendencies in the recognition of particular facial configurations as expressing discrete categories of emotion, especially the so-called “basic” emotions. Emotions studied include happiness, sadness, anger, fear, contempt and disgust. An extensive cross-cultural study, - International Study on Emotion Antecedents and Reactions (ISEAR), described in [4] - by a large group of researchers, involving self reporting of emotional experience of happiness, sadness, anger and fear in more than 30 countries, including self-assessment of the part played by the voice, has also found “a very strong universality in emotional experience, both in terms of antecedent situations and in terms of physiological, expressive, and subjective feeling reactions”. [4] (page 3).

Research into cross-cultural recognition of vocally expressed emotion is very sparse, possibly due to methodological problems of data collection and vocal parameters being more difficult to measure. See [5] for reviews. Frick notes that of the very few cross-cultural studies made, some found that cross-cultural recognition was as accurate as monolingual recognition, whilst others found that cross-culturally recognized was adversely affected. [6,7] review monolingual studies.

2. METHODOLOGICAL ISSUES

Methodological problems have been recognised in monolingual research into the vocalisation of emotion and these difficulties will often be exacerbated in cross-language research. In monolingual studies, there is considerable lack of agreement regarding how emotion is conceptualised; which emotions are studied; which methodology is used to collect the data - whether or not this is made explicit; the type of recognition test; whether or not an auditory analysis and/or an acoustic analysis is carried out and which auditory and/or acoustic parameters are analysed.

In cross-language studies, a complicating factor is the translation of emotion labels. Monolingual cross-cultural groups and even different individuals or the same individual at different points in time, may attach different connotations to the same emotion label. It is also possible that there is a western cultural bias as to which emotions are regarded as basic and further research in this area may well be helpful.
It is recognised that there are ethical problems where emotions are encoded by inducing speakers to feel an emotion. The same stimulus may also induce different emotions in different individuals. The induction of emotion may be even less tolerated in some cultures than others.

Natural, spontaneous speech will fully encode the influence of social convention upon emotion vocalisation, although there are problems of consistency. Finding examples of natural vocalisations of particular emotions can be difficult and some or all emotions may be more or less likely to be expressed in different cultures. It can sometimes also be problematic, especially with pre-recorded data, to verify the emotion the speaker intended to express. Where emotions are encoded through induction or natural speech the verbal channel needs to be masked to isolate emotion cues in the vocal channel alone. Various masking procedures have been used, all of which pose further problems. Speech may be played backwards, which distorts intonation contours. A low-pass filter may be used but this removes higher frequencies which may be important for the expression of emotion. Random splicing removes pauses and reorders the remaining recording but this removes temporal features and distorts intonation. The use of a throat microphone and laryngograph is not completely effective in masking the verbal channel and does not pick up supra-laryngeal activity.

Emotions may be vocalised in a language subjects in a recognition experiment do not understand. Using this method, recognition by native and non-native speakers of the particular language cannot be compared. Certain sounds present in the verbal channel of a language may also perhaps carry affect significance for speakers of another language which does not contain these sounds.

To avoid problems involved in masking the verbal channel and to give greater consistency, simulated data has often been used in monolingual studies, where emotions can be vocalised using semantically neutral utterances. It can be difficult to find genuinely semantically neutral sentences which can convincingly be said with different emotions. However, distinct acoustic profiles for 14 emotions were found in a study [8] using simulated vocalisations.

Vowel quality may also influence the emotion recognised. For example, Bezooijen [9] (page 57) found that “extra (para- or extralinguistic) lip rounding is easier to detect in rounded vowels, whereas extra lip spreading is easier to detect in unrounded vowels. She comments that Laver [10] suggested the reverse relationship between vowel quality and perception of lip-rounding. To investigate this further by controlling for vowel quality in semantically neutral sentences across different languages would be at best extremely difficult.

It is suggested [8] that more detailed analysis of recognition data and including a wider range of acoustic parameters may also lead to more comprehensive results. Distinct acoustic correlates divide emotions along a dimension of activation, distinguishing high arousal and low arousal vocalisations, have emerged from many studies but many authors now suggest that a multidimensional approach including, for example valence (pleasant/unpleasant; positive/negative), may be more useful in the search for acoustic and auditory correlates which distinguish discrete emotions. Studies over the past 15 years also suggest that more discrete categories may emerge from the acoustic analysis of voice quality.

Recent and continuing advances in manipulation of recorded speech, acoustic phonetic research including digital speech signal analysis and automated speech analysis systems may also be useful in this regard, as well as possibly giving us a better understanding of which muscle action patterns are associated with which acoustic parameters.

Bezooijen [9] (page 57) suggests that for acoustic parameters of emotion to be considered “meaningful and appropriate for use as standardised tools for describing emotional expressions” correlations should exist between perceptual and acoustic parameters.

3. ADDRESSING ISSUES IN AN ONGOING STUDY

3.1. Aims and hypotheses
The key aim of the study described here is to contribute to the debate on the roles of possibly innate PBRMs and social convention in emotion vocalisation. The production procedure therefore attempts to elicit valid emotion vocalisations by native English and native Japanese speakers. With the aid of a recognition test, auditory and acoustic profiles, with regard to discrete emotions and emotion dimensions, will be compared.

It is hypothesised that vocalisations of basic emotions by native speakers of each group will be recognised, beyond chance expectancy, by both groups, although subjects may be more accurate in their recognition of emotions expressed by speakers of their own language. It is also hypothesised that groups will vocalise basic emotions with some common auditory and acoustic characteristics.

3.2. Subjects
Subjects included native speakers of Japanese and native speakers of English. 8 subjects for each language participated in the production experiment and 30 native speakers of each language will take part in the recognition experiment. All are university students within an 18-35 age range. English and Japanese were chosen because they are unrelated languages, although it should be noted that Japan has been increasingly influenced by Western culture over the past 50 years. Although pitch-accent in Japanese may be a complicating factor, neither English nor Japanese is a tone language - analysis of prosodic correlates of vocal expression of emotion in tone languages may be easier to deal with when more cross-language data on non-tone languages has been gathered. There was also ease of access to native speakers of both Japanese and English. Subjects had had little contact with native speakers of any other language but their own. This meant that only Japanese speakers who had recently arrived in England were tested.

Previous studies which have investigated the possible influence of gender upon the vocalisation and recognition of emotion have found no significant influence. The scope of the study described here includes only analysis of female emotion vocalisation so only female subjects were included in the production groups. However, both male and female subjects will participate in the recognition test. In acoustic analysis,
normalisation is a problem when comparing male and female voices; Japanese females are reputedly less likely than Japanese males to hide emotion due to social stigma attached to emotion expression which is regarded as childish or feminine. Therefore if vocalisations of only one sex were to be analysed, females were more appropriate in a study searching for PBRMs since Japanese males may be more likely to control these mechanisms due to social convention. The fact that there is less research into the female voice generally is another reason for opting for female subjects in the production test.

Three subjects from each group were selected as the most reliable vocalisers of emotion, using the reliability test explained below. The edited data sample for decoding includes: 3 subjects x 2 languages x 6 nonsense utterances x 5 emotions = 180 vocalisations

### 3.3. A procedure for collecting vocalisations of emotion in 2 languages

The main goals of the procedure were to obtain valid emotion vocalisations by English and Japanese subjects whilst avoiding problems of translation, being as segmentally consistent as possible and avoiding influence from the verbal channel. Subjects were asked to simulate emotions. The fact that quasi-universal tendencies have been found in the recognition of simulated facial expressions is encouraging.

Reliability tests were performed by 4 Japanese raters on vocalisations by Japanese subjects, which had been collected by two similar but distinct procedures. There is space here to describe only one: this procedure yielded generally higher reliability ratings and 3 reliable subjects from the original 8 tested. 4 English raters also found the same three Japanese subjects reliable. Since according to the ISEAR study mentioned above, Japanese subjects self-reported very little vocal expression of emotion, it is expected that at least the same number of reliable English subjects will be found.

#### 3.3.1. Procedure and rationale

| 1. QUESTIONNAIRE on 8 facial expression photographs |
| 2. practise NONSENSE UTTERANCES |
| 3. recorded INTERVIEW |
| (i) practise nonsense utterances |
| (ii) display FACIAL EXPRESSION PHOTOGRAPHS for each emotion in turn: |
| (a) INTERJECTIONS |
| (b) Subject gives examples of emotion vocalisations |
| (iii) GAME - main focus is on researcher |
| (a) display all facial expression photographs |
| (b) instructions for game including suggested stimuli |
| (c) researcher turns around so as not to face subject |
| (d) Subject chooses emotions and vocalises them using nonsense utterances |
| (e) Researcher tries to guess emotions vocalised by subject |

Table 1 Summary of production procedure

In this ongoing study, native English and native Japanese speakers encoded emotions onto 6 nonsense utterances which are phonotactically possible in both English and Japanese. Table 1 gives a summary of the encoding procedure. These vocalisations are decoded by native speakers of both languages.

This study includes 4 emotions - happiness, sadness, anger and fear - for which universal tendencies have been found in previous research on emotion experience and recognition of facial expression. Photographs of facial expressions [2, 11] are used which have been recognised quasi-universally as expressing the same emotion, within the limits imposed by translation. To attempt to focus subjects’ attention on the photographs and emotions, they complete a questionnaire eliciting open and forced judgements of emotions expressed and asking for suggestions for interjections they, or people in the photographs, might make if they were feeling like this. These emotion labels may be cross-referenced with responses in the recognition test.

A neutral category, is used as a base against which to measure emotion - Japanese participants especially commented that this neutral category can have rather negative but never positive connotations, which could have implications for valency measurement. To balance this, “calm”, which tends to have positive connotations, may be a useful alternative base in future studies as it suggests a lack of visceral perturbation which we would associate with emotion.

There are 2 sets of 3 nonsense utterances, one set being longer than the other to test if utterance length influences recognition accuracy. For each set of 3, consonants remain the same and the vowel quality changes. The 3 vowels are [aiu]. This will allow any influence due to these vowel qualities to be highlighted (see section 2). Using nonsense utterances in a cross-language study also has other advantages: there is no information content from a verbal channel but the vocal channel is not interfered with as it is in other techniques; problems of translation are avoided; more or less the same sounds are produced by all the speakers so there is less likelihood of segmental interference in auditory and acoustic analysis.

Vocalisation of emotion in a normal social setting will tend to be controlled by the different social conventions of each group. This procedure attempts to relax this control by various methods.

Firstly, the subject is distanced as much as possible from association suggested by the verbal channel: nonsense utterances rather than verbal utterances are used; photographs rather than word labels are used for emotion categories. Secondly interjections, which may be primitively derived affect vocalisations, [12] are optionally vocalised. In an attempt to disinhibit the subject, their task is to teach the researcher how to imitate their vocalisations in order to prepare the researcher to recognise the emotions the subject was to vocalise in the game.

The game is played to elicit vocalisations: this is meant to be a disinhibiting task with a real goal - that of vocalising an emotion which the researcher could recognise. In the game the researcher turns her back in an attempt to further disinhibit the subject, especially as the subject is offered the stimulus of imitating the facial expressions on the photographs as they vocalise the emotions. Another reason for the researcher not being able to see the subject is so that the subject can only give
vocal cues to the emotion expressed. Optional stimuli include thinking of a situation in which the subject had felt or would feel the emotion, simply trying to feel the emotion, thinking of appropriate interjections, concentrating on the photographs and copying the facial expressions or any other technique the subject wished to use.

After the interview, all subjects reported that copying the facial expressions was definitely the most useful stimulus. All subjects also reported experiencing, during the experiment, at least some of the emotions they vocalised - this may be evidence of PBRMs being triggered and may be partly explained by the interpersonal feedback hypothesis and/or the facial feedback hypothesis [13]. The James-Lange theory also suggested that “afferent feedback from disturbed organs produces the feeling aspect of emotion.” [14] In this regard, the initial emotional state of the subject may have been important, as well as whether or not they were consciously or unconsciously disposed to be affected.

Since the subject is free to use any or none of the suggested stimuli to help them vocalise emotion, this method does not have the ethical problems associated with procedures. Future experiments may perhaps have a signalling device which allows the subject to indicate when they have experienced an emotion, its intensity and whether this was pleasant or unpleasant.

### 3.3.3. Reliability test

A reliability test is conducted to reduce the amount of data presented to decoders. Data from the 3 most reliable subjects for each language will be used in the recognition experiment.

Long nonsense utterances including [a] vowel are manually edited, randomly ordered by computer and numbered. The first vocalisations of each emotion by each subject in the game section are used.

There are 4 raters for each language - for each vocalisation in turn, each rater makes a forced judgement as to whether the speaker sounds happy, sad, angry or neutral. Each rater then rates the recognisability of the emotion for each utterance on a scale of 1-5: 1.insufficient, 2.almost sufficient, 3.sufficient, 4.good, 5.excellent. Only data from speakers who, taking an average over all utterances and all raters, score a 3 or higher is used in the recognition test.

### 3.4. Recognition test

30 native Japanese speakers and 30 native English speakers will make an open judgement test which will be conducted including a 1-5 rating for degree of arousal and degree of pleasantness. Subjects are then asked to make a forced judgement for each vocalisation as to whether H,S,A,F or N is expressed. They then rate for arousal and pleasantness using the rating above. A confusion matrix is constructed to highlight patterns of misjudgements by Japanese and English speakers, which are then compared.

### 3.5. Auditory and acoustic analyses

Auditory analysis of the data will include parameters such as those used by Bezooijen [9]. Acoustic analysis considers a range of parameters including frequency variables, intensity variables, speech rate and spectral energy distribution. Parameters which are musically orientated will also be analysed by music students because of the possible connections along acoustic and auditory parameters between expression of emotion in speech and music. There is little empirical research in this area although there has been some research into acoustic signals. [12].

### 4.SUMMARY

This paper has explored difficult methodological issues surrounding research into the influence of PBRM and social convention upon the auditory and acoustic correlates of emotion vocalisation. One way in which we can weave around some of these issues in a cross-cultural study has been proposed. However, extensive cross-cultural, including cross-language studies need to be conducted before any conclusions may be drawn. Comparing results of studies which have used different methodologies to address the same question as well as a multi-channel and interdisciplinary approach may help to support or question findings and lead to the gradual emergence of a fuller picture of human experience of emotion.

### REFERENCES