A CROSS-CULTURAL INVESTIGATION OF EMOTION INFERENCESS FROM VOICE AND SPEECH: IMPLICATIONS FOR SPEECH TECHNOLOGY

Klaus R. Scherer
University of Geneva, Switzerland

ABSTRACT

Recent years have seen increasing efforts to improve speech technology tools such as speaker verification, speech recognition, and speech synthesis by taking voice and speech variations due to speaker emotion or attitudes into account. Given the global marketing of speech technology products, it is of vital importance to establish whether the vocal changes produced by emotional and attitudinal factors are universal or vary over cultures and/or languages. The answer to this question determines whether similar algorithms can be used to factor out (or produce) emotional variations across all languages and cultures. This contribution describes the first large-scale effort to obtain empirical data on this issue by studying emotion recognition from voice in nine countries on three different continents.

1. OVERVIEW OF THE ISSUE

The important role of emotion in shaping human voice production has been known since antiquity (see [1] for a review of early research). Yet, due to the relative neglect of the issue of emotion-induced voice and speech changes in speech research, little progress has been made in the past 25 years. It has been only rather recently that the important role of emotion and attitude dependent voice and speech variations has found the interest of phoneticians and engineers working on speech synthesis, speech recognition, and speaker verification [2,3]. Currently, a number of efforts are underway to understand the effects of emotion on voice and speech and to examine possibilities to adapt speech technology algorithms accordingly [4-6]. However, one central issue remains largely unexplored: the question of the universality vs. the cultural and/or linguistic relativity of the emotion effects on vocal production. This is obviously of major importance for the development and marketing of speech technology products that have been enhanced to allow the factoring out of emotional or attitudinal variability in recognition and verification or to produce appropriate acoustic features in synthesis. Obviously, little customization of the respective algorithms would be necessary if emotion effects on the voice were universal whereas culturally or linguistically relative effects would require special adaptations for specific languages or countries. It is of major practical interest, then, to determine whether there are systematic and differentiated effects of different emotions on voice and speech characteristics, whether these are universal or culturally/linguistically relative, and whether intercultural recognition of emotion cues in the voice is possible.

The apparent predominance of the activation dimension in vocal emotion expression has often led critics to suggest that judges’ ability to infer emotion from the voice might be limited to basing inference systematically on perceived activation. In consequence, one might expect that if this single, overriding dimension is controlled for, there should be little evidence for judges’ ability to recognize emotion from purely vocal, non-verbal cues. However, evidence from a recent study by Banse & Scherer [7] that used a larger than customary set of acoustic variables, as well as more controlled elicitation techniques, points to the existence of both activation and valence cues to emotion in the voice. Twelve professional actors were asked to portray fourteen different emotions using two standard, meaningless sentences. Their portrayals were based on scenarios provided for each emotion. The actors were asked to use Stanislavski techniques, i.e. attempt to immerse themselves into the scenario and feel the emotion in the process of portraying it. The resulting speech samples were presented to expert judges to eliminate unsuccessful or unnatural expressions. In the next step naive

Figure 1: Beta weights and multiple R for selected acoustic parameters predicting emotion judgments in multiple regressions (adapted from [7]).
judges were used to assess the accuracy with which the different speech samples could be recognized with respect to the expressed emotion. 224 portrayals that were recognized with a sufficient level of accuracy were chosen for acoustic analysis. The results of these analyses suggest the existence of emotion-specific vocal profiles that differentiate the different emotions not only on a basic arousal or activation dimension but also with respect to qualitative differences.

Figure 2: Accuracy percentages for emotion classification attained by human judges, jackknifing, and discriminant analysis (adapted from [7]).

These results explain why studies investigating the ability of listeners to recognize different emotions from a wide variety of standardized vocal stimuli achieve a level of accuracy that largely exceeds what would be expected by chance [8]. The assumption is, of course, that the judges' inference is in fact based on the acoustic cues that determine the vocal profiles for the different emotions. Figure 1, which plots the beta weights from multiple regressions of vocal features on emotion judgment, shows that this is indeed the case: A sizeable proportion of the variance in the judgments can be explained by the acoustic cues that are prominently involved in differentiating vocal emotion profiles.

This interpretation is bolstered by a comparison of the confusion matrices produced by human judges with those found for statistical techniques of classification on the basis of predictor variables, namely jack-knifing and discriminant analysis. Figure 2 shows the respective accuracy percentages for these three types of classification. Not only is the pattern of accuracy coefficients across emotions (the diagonals in the confusion matrices as shown in Figure 2) highly comparable (with a few interesting exceptions) but the off-diagonal errors are also very similar across judges and classification algorithms (see [7] for details). This provides further evidence for the assumption that judges base their inference of emotion in the voice on acoustic profiles that are characteristic for specific emotions.

Furthermore, many aspects of the profiles identified by Banse & Scherer were predicted by Scherer's Component Process model of emotion [9] that assumes a universal, psychobiological mechanism (push effects) for the vocal expression of emotion (even though it also allows for sociocultural variations – pull effects). This suggests that many of the emotion effects on the voice should be universal. Unfortunately, there is no study to date that examined these effects across speakers from different languages and cultures. In the meantime, it may be useful to examine whether judges from different countries can identify vocal expressions from another language/culture.

Whereas the perception of emotion from facial expression has been extensively studied cross-culturally, little is known about the ability of judges from different cultures, speaking different languages, to infer emotion from voice and speech encoded in another language by members of other cultures. This contribution summarizes the results of a series of recognition studies conducted by Scherer, Banse, and Wallbott [10] in nine different countries in Europe, the United States, and Asia (using vocal emotion portrayals with content-free sentences as produced by professional German actors).

2. METHODS

2.1 Emotional Voice Samples

Four German professional actors (two male and two female) with many years of experience in radio, TV, and stage work portrayed the emotions of fear, joy, sadness, anger, disgust on the basis of appropriate emotion situation scenarios (as well as producing neutral speech samples). The speech material used consisted of two artificial sentences that been designed by a phonetician, in view of the intended use in cross-linguistic research, to include different phonetic elements from different Indo-European languages (/haet sandik prongnu ventsi/ and /fi goettlaich schongkill gostaer/). The portrayals were recorded in the studios of a West German radio station [11].

A series of judgment studies was conducted to determine whether the intended emotions could be reliably recognized by lay judges. In a first study with all 80 emotion portrayals (using 29 judges) 30 particularly well recognized stimuli were selected (5 fear, 5 disgust, 6 joy, 7 sadness, 7 anger). These were
presented to a second group of 20 judges. Figure 1 shows the
accuracy percentages obtained (compared to the percentage
expected by chance, taking the frequency of category usage by
the judges into account; see [11] for details).

Figure 3: Accuracy of recognition of 30 stimuli by 20 German
listeners

Digital analyses of these stimuli were performed to determine
how a number of major acoustic cues were affected by the
portrayals of the different emotions. The results are shown in
Figure 4. The values shown are difference scores to the neutral
renderings, standardized by dividing through the standard
deviation across all renderings (including neutral). Since the
values for male and female actors were generally similar gender
differences are not shown.

Figure 4: Acoustic profiles for four emotions

2.2 Cross-cultural recognition study

Given the low recognition rate for disgust, this emotion was
eliminated in the intercultural study. Furthermore, contrary to
the earlier judgment studies, 7 neutral expression stimuli were
included in the stimulus set to be judged. Collaborators in nine
countries on three continents participated in the study: D =
Germany, CH = Switzerland, GB = Great Britain, NL =
Netherlands, US = United States, I = Italy, F = France, E = Spain,
IND = Indonesia. They presented the 30 stimuli to groups of 5
judges per session, presenting the stimuli via cassette recorder
and yoked headphones (identical equipment had been shipped to
all nine sites). Judges were asked to circle one of the five
emotion categories that described the expressed emotion best (or
two if they thought that an emotion blend was expressed). The
data were returned to Geneva for analysis (see [10] for further
details on the procedure)

3. RESULTS

The data, graphed in Figure 3, show an overall accuracy
percentage of 66% across all emotions and countries. Joy was
inferred with much lower accuracy (42%) than the other
emotions (around 70%). While accuracy was in all cases
substantially better than chance, there were sizable differences
across countries ranging from a high of 74% in Germany to 52%
in Indonesia. The languages of the countries with the highest
accuracy rates are of Germanic origin (Dutch and English),
followed by Romanic languages (Italian, French, and Spanish).
The lowest recognition rate was obtained for the only country
studied that does not belong to the Indo-European language
family, Indonesia.

Figure 5: Percentage of correct responses by type of emotion
and country of judges (adapted from [10]).

However, interestingly enough, the profile correlations across
emotion-specific hit rates show a high degree of similarity with
the patterns in the other countries (see Table 1).
Table 1: Correlations of the accuracy profiles over emotions between countries. (adapted from [10]).

Furthermore, the patterns of errors are highly comparable as shown by the profile correlations across errors listed in Table 2.

Table 2: Correlations between countries over error patterns in the confusion matrices. Legend as in Table 1.

4. CONCLUSIONS

These data provide some support for the claim that vocal emotion expression may be at least in large part driven by universal psychobiological mechanisms since judges from different cultures, speaking different languages, recognize the expressed emotions with much better than chance accuracy. However, it seems that as soon as vocal expressions other than pure nonlinguistic affect bursts are used, segmental and suprasegmental aspects of language affect encoding and decoding of emotion. Since we used content-free utterances composed of phonological units from many different Indo-European languages, effects must be either due to segmental information (such as phoneme specific fundamental frequency, articulation differences, formant structure, or the like) or to suprasegmental parameters, such as prosodic cues (intonation, rhythm, timing). Clearly, further research using encoders and decoders from different languages and cultures is urgently needed to establish in greater detail which language or culture-specific expression rules speech technology needs to incorporate into the development of products to be marketed on a global scale.

5. REFERENCES