

Emotions in speech: an interactional framework for clinical applications

Ani Toivanen¹ & Juhani Toivanen²

¹University of Oulu

²MediaTeam, University of Oulu & Academy of Finland

Abstract

The expression of emotion in human communicative interaction has been studied extensively in different theoretical paradigms (linguistics, phonetics, psychology). However, there appears to be a lack of research focusing on emotion expression from a genuinely interactional perspective, especially as far as the clinical applications of the research are concerned. In this paper, an interactional, clinically oriented framework for an analysis of emotion in speech is presented.

Introduction

Human social communication rests to a great extent on non-verbal signals, including the (non-lexical) expression of emotion through speech. Emotions play a significant role in social interaction, both displaying and regulating patterns of behavior and maintaining the homeostatic balance in the organism. In everyday communication, certain emotional states, for example, boredom and nervousness, are probably expressed mainly non-verbally since socio-cultural conventions demand that patently negative emotions be concealed (a face-saving strategy in conversation).

Today, the significance of emotions is largely acknowledged across scientific disciplines, and “Descartes’ error” (i.e. the view that emotions are “intruders in the bastion of reason”) is being corrected. The importance of emotions/affect is nowadays understood better, also from the viewpoint of rational decision-making (Damasio, 1994).

Basically, emotion in speech can be broken down to specific vocal cues. These cues can be investigated at the signal level and at the symbolic level. Such perceptual features of speech/voice vs. emotion/affect as “tense”, “lax”, “metallic” and “soft”, etc. can be traced back to a number of continuously variable acoustic/prosodic features of the speech signal (Laver, 1994). These features are f₀-related, intensity-related, temporal and spectral features of the signal, including, for example, average f₀

range, average RMS intensity, average speech/articulation rate and the proportion of spectral energy below 1,000 Hz. At the symbolic level, the distribution of tone types and focus structure in different syntactic patterns can convey emotional content.

The vocal parameters of emotion may be partially language-independent at the signal level. For example, according to the “universal frequency code” (Ohala, 1983), high pitch universally depicts supplication, uncertainty and defenseless, while low pitch generally conveys dominance, power and confidence. Similarly, high pitch is common when the speaker is fearful, such an emotion being typical of a “defenseless” state.

An implicit distinction is sometimes made between an emotion/affect and an attitude (or stance in modern terminology) as it is assumed that the expression of attitude is controlled by the cognitive system that underpins fluent speech in a normal communicative situation, while true emotional states are not necessarily subject to such constraints (the speech effects in real emotional situations may be biomechanically determined by reactions not fully controlled by the cognitive system). It is, then, possible that attitude and emotion are expressed in speech through at least partly different prosodic cues (which is the taking-off point for the symbolic/signal dichotomy outlined above). However, this question is not a straightforward one as the theoretical difference between emotion and attitude has not been fully established.

Emotions in speech

By now, a voluminous literature exists on the emotion/prosody interface, and it can be said that the acoustic/prosodic parameters of emotional expression in speech/voice are understood rather thoroughly (Scherer, 2003). The general view is that pitch (fundamental frequency, f₀) is perhaps the most important parameter of the vocal expression of emotion (both productively and perceptually); energy (intensity), duration and speaking rate are the other relevant parameters.

Somewhat surprisingly, although the emotion/vocal cue interface in speech has been investigated extensively, there is no widely accepted definition or taxonomy of emotion. Apparently, there is no standard psychological theory of emotion that could decide the issue once and for all: the number of basic (and secondary) emotions is still a moot point. Nevertheless, certain emotions are often considered to represent “basic emotions”: at least fear, anger, happiness, sadness, surprise and disgust are among the basic emotions (Cornelius, 1996).

Research on the vocal expression of emotion has been largely based on scripted non-interactive material; a typical scenario involves a group of actors simulating emotions while reading out an emotionally neutral sentence or text. There are now also databases containing natural emotional speech, but these corpora (necessarily) tend to contain blended/uncertain and mixed emotions rather than “pure” basic emotions (see Scherer, 2003, for a review).

Emotions in speech: clinical investigations

The vocal cues of affect have been investigated also in clinical settings, i.e. with a view to charting the acoustic/prosodic features of certain emotional states (or states of emotional disorders or mental disorders). For example, it is generally assumed that clinical depression manifests itself in speech in a way which is similar to sadness (a general, “non-morbid” emotional state). Thus, a decreased average f_0 , a decreased f_0 minimum, and a flattened f_0 range are common, along with decreased intensity and a lower rate of articulation (Scherer, 2000). Voiced high frequency spectral energy generally decreases. Intonationally, sadness/depression may typically be associated with downward directed f_0 contours.

Psychiatric interest in prosody has recently shed light on the interrelationship between schizophrenia and (deficient or aberrant) prosody. Several investigators have argued that schizophrenics recognize emotion in speech considerably worse than members of the normal population. Productively, the situation appears quite similar, i.e. schizophrenics cannot convey affect through vocal cues as consistently and effectively as normal subjects (Murphy & Cutting, 1990). In the investigation by Murphy & Cutting (1990), a group of schizophrenics were to express basic emotions (neutral, angry, sur-

prise, sad) while reading out a number of sentences. The raters (normal subjects) had significant difficulty recognizing the simulated emotions (as opposed to portrayals of the same emotions by a group representing the normal population).

In general, it has been found out that speech and communication problems typically precede the onset of psychosis; dysarthria and dysprosody appear to be common. Affective flattening is indeed a diagnostic component of psychosis (along with, for example, grossly disorganized speech), and anomalous prosody (e.g. a lack of any observable speech melody) may thus be an essential part of the dysprosody evident in psychosis (Golfarb & Bekker, 2009). Moreover, schizophrenics’ speech seems to contain more pauses and hesitation features than normal speech (Covington et al., 2005). Interestingly, although depressed persons’ speech also typically contains a decreased amount of speech per the speech situation, the distribution of pauses appears to be different from schizophrenic speech: schizophrenics typically pause in “wrong” (syntactically/semantically) unmotivated places, while the pausing is more logical and grammatical in depressed speech. Schizophrenic speech thus seems to reflect the erratic semantic structure of what is said (Clemmer, 1980).

It would be fascinating to think that certain prosodic features (or their absence) could be of help for the general practitioner when diagnosing mental disorders. Needless to say, such features could never be the only diagnostic tool but, in the best scenario, they would provide some assistive means for distinguishing between some alternative diagnostic possibilities.

Emotions in speech: an interactional clinical approach

In the following sections we outline a preliminary approach to investigating emotional speech and interaction within a clinical context. What follows is, at this stage, a proposal rather than a definitive research agenda.

Prosodic analysis: 4-Tone EVO

Our first proposal concerns the prosodic annotation procedure for a speech material produced in a (clinical) setting inducing emotionally laden speech. As is well known, ToBI labeling (Beckman & Ayers, 1993) is commonly used in the prosodic transcription of (British and Amer-

ican) English (and the system is used increasingly for the prosodic annotation of other languages, too), and good inter-transcriber consistency can be achieved as long as the voice quality analyzed represents normal (modal) phonation. Certain speech situations, however, seem to consistently produce voice qualities different from modal phonation, and the prosodic analysis of such speech data with traditional ToBI labeling may be problematic. Typical examples are breathy, creaky and harsh voice qualities. Pitch analysis algorithms, which are used to produce a record of the fundamental frequency (f0) contour of the utterance to aid the ToBI labeling, yield a messy or lacking f0 track on non-modal voice segments. Non-modal voice qualities may represent habitual speaking styles or idiosyncrasies of speakers but they are often prosodic characteristics of emotional discourse (sadness, anger, etc.). It is likely, for example, that the speech of a depressed subject is to a significant extent characterized by low f0 targets and creak. Therefore, some special (possibly emotion-specific) speech genres (observed and recorded in clinical settings) might be problematic for traditional ToBI labeling.

A potential modified system would be “4-Tone EVo” – a ToBI-based framework for transcribing the prosody of modal/non-modal voice in (emotional) English. As in the original ToBI system, intonation is transcribed as a sequence of pitch accents and boundary pitch movements (phrase accents and boundary tones). The original ToBI break index tier (with four strengths of boundaries) is also used. The fundamental difference between 4-Tone EVo and the original ToBI is that four main tones (H, L, h, l) are used instead of two (H, L). In 4-Tone EVo, H and L are high and low tones, respectively, as are “h” and “l”, but “h” is a high tone with non-modal phonation and “l” a low tone with non-modal phonation. Basically, “h” is H without a clear pitch representation in the record of f0 contour, and “l” is a similar variant of L.

Preliminary tests for (emotional) English prosodic annotation have been made using the model, and the results seem promising (Toivonen, 2006). To assess the usefulness of 4-Tone EVo, informal interviews with British exchange students (speakers of southern British English) were used (with permission obtained from the subjects). The speakers described, among other things, their reactions to certain personal dilemmas (the emotional overtone was, predictably, rather low-keyed).

The discussions were recorded in a sound-treated room; the speakers’ speech data was recorded directly to hard disk (44.1 kHz, 16 bit) using a high-quality microphone. The interaction was visually recorded with a high-quality digital video recorder directly facing the speaker. The speech data consisted of 574 orthographic words (82 utterances) produced by three female students (20-27 years old). Five Finnish students of linguistics/phonetics listened to the tapes and watched the video data; the subjects transcribed the data prosodically using 4-Tone EVo. The transcribers had been given a full training course in 4-Tone EVo style labeling. Each subject transcribed the material independently of one another.

As in the evaluation studies of the original ToBI, a pairwise analysis was used to evaluate the consistency of the transcribers: the label of each transcriber was compared against the labels of every other transcriber for the particular aspect of the utterance. The 574 words were transcribed by the five subjects; thus a total of 5740 (574x10 pairs of transcribers) transcriber-pair-words were produced. The following consistency rates were obtained: presence of pitch accent (73 %), choice of pitch accent (69 %), presence of phrase accent (82 %), presence of boundary tone (89 %), choice of phrase accent (78 %), choice of boundary tone (85 %), choice of break index (68 %).

The level of consistency achieved for 4-Tone EVo transcription was somewhat lower than that reported for the original ToBI system. However, the differences in the agreement levels seem quite insignificant bearing in mind that 4-Tone EVo uses four tones instead of two!

Gaze direction analysis

Our second proposal concerns the multimodality of a (clinical) situation, e.g. a patient interview, in which (emotional) speech is produced. It seems necessary to record the interactive situation as fully as possible, also visually. In a clinical situation, where the subject’s overall behavior is being (at least indirectly) assessed, it is essential that other modalities than speech be analyzed and annotated. Thus, as far as emotion expression and emotion evaluation in interaction are concerned, the coding of the visually observable behavior of the subject should be a standard procedure. We suggest that, after recording the discourse event with a video recorder, the gaze of the subject is annotated as follows. The gaze of the subject (patient) may

be directed towards the interlocutor (+directed gaze) or shifted away from the interlocutor (-directed gaze). The position of the subject relative to the interlocutor (interviewer, clinician) may be neutral (0-proxemics), closer to the interlocutor (+proxemics) or withdrawn from the interlocutor (-proxemics). Preliminary studies indicate that the inter-transcriber consistency even for the visual annotation is promising (Toivanen, 2006).

Post-analysis: meta-interview

Our third proposal concerns the interactionality and negotiability of a (clinical) situation yielding emotional speech. We suggest that, at some point, the subject is given an opportunity to evaluate and assess his/her emotional (speech) behavior. Therefore, we suggest that the interviewer (the clinician) will watch the video recording together with the subject (the patient) and discuss the events of the situation. The aim of the post-interview is to study whether the subject can accept and/or confirm the evaluations made by the clinician. An essential question would seem to be: are certain (assumed) manifestations of emotion/affect “genuine” emotional effects caused by the underlying mental state (mental disorder) of the subject, or are they effects of the interactional (clinical) situation reflecting the moment-by-moment developing communicative/attitudinal stances between the speakers? That is, to what extent is the speech situation, rather than the underlying mental state or mood of the subject, responsible for the emotional features observable in the situation? We believe that this kind of post-interview would enrich the clinical evaluation of the subject’s behavior. Especially after a treatment, it would be useful to chart the subject’s reactions to his/her recorded behavior in an interview situation: does he/she recognize certain elements of his/her behavior being due to his/her pre-treatment mental state/disorder?

Conclusion

The outlined approach to a clinical evaluation of an emotional speech situation reflects the Systemic Approach: emotions, along with other aspects of human behavior, serve to achieve intended behavioral and interactional goals in co-operation with the environment. Thus, emotions are always reactions also to the behavioral acts unfolding in the moment-by-moment face-to-face interaction (in real time). In addition, emotions often reflect the underlying long-term

affective state of the speaker (possibly including mental disorders in some subjects). An analysis of emotions in a speech situation must take these aspects into account, and a speech analyst doing research on clinical speech material should see and hear beyond “prosodemes” and given emotional labels when looking into the data.

References

- Beckman M.E. and Ayers G.M. (1993) Guidelines for ToBI Labeling. Linguistics Department, Ohio State University.
- Clemmer E.J. (1980) Psycholinguistic aspects of pauses and temporal patterns in schizophrenic speech. *Journal of Psycholinguistic Research* 9, 161-185.
- Cornelius R.R. (1996) *The science of emotion. Research and tradition in the psychology of emotion.* New Jersey: Prentice-Hall.
- Covington M, He C., Brown C., Naci L., McClain J., Fjorbak B., Semple J. and Brown J. (2005) Schizophrenia and the structure of language: the linguist’s view. *Schizophrenia Research* 77, 85-98.
- Damasio A. (1994) *Descartes’ error.* New York: Grosset/Putnam.
- Golfarb R. and Bekker N. (2009) Noun-verb ambiguity in chronic undifferentiated schizophrenia. *Journal of Communication Disorders* 42, 74-88.
- Laver J. (1994) *Principles of phonetics.* Cambridge: Cambridge University Press.
- Murphy D. and Cutting J. (1990) Prosodic comprehension and expression in schizophrenia. *Journal of Neurology, Neurosurgery and Psychiatry* 53, 727-730.
- Ohala J. (1983) Cross-language use of pitch: an ethological view. *Phonetica* 40, 1-18.
- Scherer K.R. (2000) Vocal communication of emotion. In Lewis M. and Haviland-Jones J. (eds.) *Handbook of Emotions*, 220-235. New York: The Guilford Press.
- Scherer K.R. (2003) Vocal communication of emotion: a review of research paradigms. *Speech Communication* 40, 227-256.
- Toivanen J. (2006) Evaluation study of “4-Tone EVo”: a multimodal transcription model for emotion in voice in spoken English. In Toivanen J. and Henrichsen P. (eds.) *Current Trends in Research on Spoken Language in the Nordic Countries*, 139-140. Oulu University & CMOL, Copenhagen Business School: Oulu University Press.