Prosodic Expressions of Emotions and Attitudes in Communicative Feedback

Gustaf Lindblad University of Gothenburg Gothenburg, Sweden
gustaf.lindblad@gu.se

Jens Allwood SCCIIL (SSKKII) University of Gothenburg Gothenburg, Sweden
jens.allwood@gu.se

Abstract

This study investigates the communication of affective-epistemic states (AES) by means of prosody in vocal verbal feedback. The study was conceived as a pilot study to test certain methodological queries about investigating prosody as a part of multimodal communication, and as part of feedback in particular. We find that our method, with some slight adjustments, seems adequate to answer several interesting questions about prosodic features of feedback.

Keywords: prosody, emotion, attitude, communicative feedback

1 Introduction

We communicate emotions and attitudes using a few different means, perhaps the most well known and most studied being facial expressions. In vocal communication we also modulate different aspects of our speech, in particular voice quality and prosody, to convey our internal states.

Like facial expressions, speech modulations can both be voluntary and involuntary, as well as both innate and learned. In all cases, they need to be shared to a certain extent among their users to function as markers of internal states. In other words, there needs to exist some systematic mapping between expression and internal state that guides the interpretation of the expression. This mapping has not been subject of as much study for prosody as is the case for facial expressions.

The differences between emotions, attitudes and several other internal mental states are not always clear. Basically, the grounds for differentiating between these mental states are the experience and behaviour of the agent. Anger, happiness and fear are typically described as emotions, whereas being interested, sceptical or condescending are typically designated as attitudes. Similar internal states such as being hungry or tired are not generally regarded as either emotions or attitudes, but are in many ways similar to emotions and attitudes. All such mental states can be construed as having the function of making an agent more likely to behave in a particular way. For these reasons, we chose to use the inclusive term 'affective-epistemic state' (AES), to refer to any such internal mental state that can manifest itself both in an experience and a behaviour. This article is focused on the vocal expression of such affective-epistemic states.

Communicative feedback can be defined as unobtrusive vocal and bodily expressions informing an interlocutor about the feedback giver's ability and willingness to (i) continue the interaction, to (ii) perceive and (iii) understand what is communicated, and (iv) in other ways attitudinally and emotionally react (e.g. Allwood, 1988; Allwood et al, 1992). Examples of vocal verbal feedback expressions in Swedish are words such as *m* ('m'), *ja* ('yes'), *nä* ('no'), and *okej* ('okay'), phrases such as *jag förstår* ('I see'), and repetition of what the interlocutor just said. Feedback often reflects the speaker's attitude or emotion with regard to the topic, interlocutor or context in general. Since vocal verbal feedback often consists of short one-word utterances of a limited number of words, vocal verbal feedback is a good candidate for also studying the communicative functions of prosody.

There are two basic features of speech that are modulated to convey affective-epistemic states: voice qualities and prosody. It is not possible to disassociate the two completely, as they can be partly dependent on each other. Voice qualities can be such thing as a raspy or nasal voice, and prosody correlates to intonation and stress, and often translates to what in common speech is called the tone of voice, e.g. a sharp or sarcastic tone.

Prosody is typically measured through three aspects of the vocal signal: pitch, intensity and duration. The pitch is usually identified with the fundamental frequency of the voice (F0) and measured in Hz, intensity is the volume of the voice measured in dB, and the duration is simply measured in milliseconds. The pitch and intensity varies along the duration of an utterance and can be visualized as a curve.

Pitch and intensity of an utterance can be measured, for example, by maximum, minimum or mean value, or by the shape of the curve, such as rising, falling or flat. The intensity of any utterance will always feature a rise in the beginning and a fall in the end as a natural part of the sound.

2 Method

The affective-epistemic states that we chose to use for the recordings were selected with the intention of getting a substantially different array of differently sounding samples. This list is not intended to be thought of as exhaustive or to reflect any specific opinion or statement on behalf of the authors. The affective-epistemic states have been translated into English, with the original Swedish word in brackets.

determined (bestämd)	surprised (överraskad)
factual (konstaterande)	interrogative (frågande)
hurried (stressad)	bored (uttråkad)
happy (glad)	uncertain (osäker)
irritated (irriterad)	neutral (neutral)

Table 1. List of the affective-epistemic states that the speakers were instructed to produce.

We recorded two different persons who were instructed to produce the five most basic feedback words in Swedish ('ja', 'nej', 'm', 'okej', and 'jo') in a way that they felt captured the different affective-epistemic states. Both persons were male, one in his sixties, and the other in his forties. They produced every feedback word three times in each affective-epistemic state (in total yielding a library of 300 samples).

The samples were recording using a studio grade condenser microphone (ADK A-51, fixed cardioid) in close proximity to the mouth of the speaker, with a bit depth of 24 and a sample rate of 48 kHz, in all giving a linear and clear (high signal-to-noise ratio) recording of the signal.

We selected one sample of each affective-epistemic state produced by either speaker, resulting in 20 samples in total, which we played back to an audience of 25 first year students in cognitive science. They were each given a form with 20 numbered lines and asked to write down what emotion or attitude they intuitively felt was being expressed by each utterance. The samples were played in random order with regards to the affective-epistemic state, but not in terms of speaker. That is, first the ten samples of the first speaker were played back in random order, and then the ten samples of the second speaker were played in random order.

The prosodic features of the samples were analysed using Praat (ref). Duration was measured in 10ths of milliseconds. Intensity was measured as the mean intensity between the start- and stop-point of the utterance. Peak intensity was also measured but not used for this study. Pitch was measured in several ways: the shape of the f0-curve was categorized into one of eight categories (Boholm & Lindblad, 2011). The mean pitch of the utterance was measured, as well as the averages of the highest and lowest 30 ms portions. The difference between the highest and lowest pitch of the utterance is calculated using the following formula ((hi pitch)-(lo pitch) / (hi pitch)), which gives a value between 0 and 1, where higher value means greater difference. The label we have given to this value is "pitch difference".

Category	Description
Flat	The pitch curve describes a more or less flat f0, neither rising nor falling, with fluctuations smaller than 5%
Rise	The curve is increasing throughout
Complex-rise	The curve has an overall trend of increase, but contains smaller anomalies or fluctuations
Fall	The curve is decreasing throughout
Complex-fall	The curve has an overall trend of decrease, but con-

	tains smaller anomalies or fluctuations
Fall-rise	The curve has a distinct u-shape
Rise-fall	The curve has a distinct arced shape
Complex	The curve describes a more varied shape and does not fall in any category

Table 2. Categories of pitch curve shapes.

There are valid concerns that experimental speech has a low ecological validity for drawing conclusions about the intrinsic qualities of speech, and that ideally natural speech should be used in research such as this. We share this concern, but we find that we cannot get good enough quality recordings of natural speech to make reliable measurements. When measuring the fine qualities of prosody, it is very important to have a good control of the signal so you know what you are actually measuring: most importantly, the signal needs to be isolated from other sources of sound, and the subject needs to have a fixed distance and angle to the microphone. If these conditions are not met, the pitch measures often become distorted or void, and the intensity measures cannot be compared to each other.

The recordings in our study are experimental and the actors emulate the emotions, but the reactions of our panels are genuine. We believe that this approach gives us a suitable balance between ecological validity and clear data.

3 Results

Twenty-five respondents heard twenty samples each and were asked to write down their interpretation of the AES in the sample in Swedish (free choice), resulting in 500 answers. After correction of spelling errors and grouping synonyms and derivations together (e.g. if one respondent had written 'happiness' and another 'happy' these would be sorted in the same category, i.e. 'happy'), we found that the twenty most common reported affective-epistemic state's in English translation, were the following:

Occurr-	**AESences**

Table 3. Occurrences of most commonly reported AES's.

These are in total 302 of all 500 answers, all other responses occurred four times or less, 32 were blank, 44 were of a more inventive and un-categorizable nature, such as "condescendingly sympathetic" or "you are right, but I don't agree with you".

There is no one-to-one mapping of the answers of our respondents to the instructions of our actors, but there is considerable overlap. The difference in agreement between the respondents for individual samples ranges from almost unanimous for certain samples, to almost no agreement for others. We restrict ourselves to three typical examples, as there is too little data to be conclusive in any way.

One of the samples what was recorded with an 'interrogative' (questioning) AES got the following responses: 9 pensive, 7 interrogative, 2 happy, 2 agreeing, 1 blaming, 1 bored, 1 hesitant, 1 hopeful, 1 blank. We can see that even though there are some quite differing answers, the two most common answers have some similarity in meaning.

Another sample that was recorded with the AES 'happy' got the following responses: 11 happy, 8 positive, 1 enthusiastic, 1 exalted, 1 interested, 1 inviting, 1 sure, 1 determined. This shows quite a high degree of agreement between the respondents.

An example of a sample with very low agreement is one that was recorded with the AES 'hurried'. The responses were: 5 determined, 4 stubborn, 3 harsh, 2 afraid, 1 commanding, 1 formal, 1 negative, 1 offended, 1 sure, 1 unsettled, 2 blank and 3 incomprehensible. This sample had a duration of only 230 ms, which is likely to have contributed to this sample being hard to interpret.

The samples that got the least agreement among the respondents on average show higher intensity and pitch, and wavering intensity- and pitch-curves. On contrast, the samples with the highest listener agreement have lower intensity and pitch, as well as more even intensity- and pitch-curves.

It should also be noted that among the samples with the highest listener agreement, several had quite distinct and audible non-prosodic voice qualities, such as creaking voice or audible breath sounds. This indicates that such features of the voice signal can have similar functions to prosody in indicating and displaying AES's.

For every category we calculated an average for the different parameters, based on every sample that was reported as belonging to that category. This means that every sample was counted as one instance every time it was reported as being a specific AES. E.g. if sample x was reported to be an instance of 'happy' by two different respondents and sample y was reported by one respondent, the average for any parameter of 'happy´ would be $((x_p + x_p + y_p)/3)$. This table presents four of these prosodic parameters.

Some interesting patterns emerge when the different categories are grouped together based on their averages on three key parameters, i.e. duration, intensity and pitch. For each parameter we split the range of the resulting values into three equal parts, e.g. if the range of the values on a particular parameter was between 1-30, all values between 1-10 would be designated as a low value, 11-20 medium and 21-30 high.

Term	Dur	Mean pitch	Pitch diff	Mean int
Hesitant	0,75	153	0,38	66
Determined	0,27	136	0,32	73
Surprised	0,54	200	0,63	76
Thoughtful	0,77	146	0,6	66
Uncertain	0,91	160	0,47	66
Happy	0,39	171	0,61	70
Interrogative	0,44	156	0,59	69
Hurried	0,24	140	0,27	76
Positive	0,34	164	0,52	73
Agreeing	0,31	144	0,52	67
Certain	0,31	134	0,37	73
Harsh	0,21	139	0,29	72
Neutral	0,31	131	0,44	70
Tired	0,46	124	0,26	67
Reluctant	0,69	140	0,29	66
Dejected	0,6	-	-	65
Interested	0,4	193	0,65	70
Stubborn	0,3	167	0,36	74
Bored	0,72	147	0,4	67
Sceptical	0,98	144	0,5	66

Table 4. Averages of the main prosodic features of the most commonly reported AES's.

	Low dur	Med dur	Hi dur
High int	stubborn hurried	surprised	
Med int	neutral certain determined harsh	positive happy asking	
Low int	agreeing	tired	sceptical thoughtful bored hesitant uncertain reluctant

Table 5. Grouping of the most commonly reported AES's in terms of duration and intensity.

	Low pitch	Med pitch	Hi pitch
High int	hurried		surprised stubborn
Med int	neutral sure harsh determined	asking	positive happy
Low int	reluctant tired	thoughtful bored hesitant uncertain agreeing	sceptical

Table 6. Grouping of the most commonly reported AES's in terms of pitch and intensity.

	Low pitch	Med pitch	Hi pitch
High dur	reluctant	thoughtful bored hesitant uncertain	sceptical
Med dur	tired	asking	surprised positive happy
Low dur	hurried neutral sure harsh determined	content agreeing	stubborn

Table 7. Grouping of the most commonly reported AES's in terms of pitch and duration.

There are three groups of labels that are not differentiated from each other in these dimensions, and they are 1) positive, happy; 2) neutral, sure, decided, harsh; 3) thoughtful, bored, hesitant, unsure.

In the case of the first group, it is hardly surprising to see that 'positive' and 'happy' fall close together. Looking more closely at their typical patterns, we can also find that both typically have a rise-fall pitch curve, and that both have a high pitch difference. Not very much set them apart in these data. We do find that 'happy' has somewhat longer duration, higher pitch and larger pitch difference, but lower intensity. The values for intensity are a little counterintuitive, as the word 'happy' would suggest a more intense affective state than 'positive', and the other three variables also indicate this. But emotional intensity does not equal vocal intensity in our data.

Term	Dur	Mean pitch	Pitch diff	Mean int
Happy	0,39	171	0,61	70
Positive	0,34	164	0,52	73

Table 8. Average values of the main prosodic features of 'happy' and 'positive'.

The second group has some cohesion between the labels, at least 'certain' and 'decided' seem to have some semantic similarity. Looking more closely at the data for this group, we also find that for three out of four values, 'certain' is closer to 'neutral' while 'decided' is closer to 'harsh', and that 'neutral' and 'harsh' are at the opposite ends of the spectra. The exception is intensity, where 'harsh' has a lower value than both 'decided' and 'certain'. But the differences are very small. 'Neutral' seems to be predominantly characterized by a fall-rise pitch curve, while none of the others have any typical pitch contour.

Term	Dur	Mean pitch	Pitch diff	Mean int
Neutral	0,31	131	0,44	70
Certain	0,31	134	0,37	73
Determined	0,27	136	0,32	73
Harsh	0,21	139	0,29	72

Table 9. Average values of the main prosodic features of 'neutral', 'certain', 'determined' and 'harsh'.

In the third group, 'thoughtful', 'hesitant' and 'uncertain' have some semantical similarity, but 'bored' seems to be a completely different thing. It should be noted that bored was only reported five times, whereas the others were reported more than 20 times each. All four typically have a rising pitch.

Term	Dur	Mean pitch	Pitch diff	Mean int
Bored	0,72	147	0,4	67
Hesitant	0,75	153	0,38	66
Thoughtful	0,77	146	0,6	66
Uncertain	0,91	160	0,47	66

Table 8. Average values of the main prosodic features of 'bored', 'hesitant', 'thoughtful' and 'uncertain'.

We also find clear indications that duration and intensity show a roughly linear correlation; shorter durations are correlated with higher intensity and vice versa. Interestingly we also find a clear pattern that medium duration expressions are correlated with higher pitch and larger pitch differences on average.

4 Problems

Since this work was carried out as an extended pilot study to test the methodology, we will report some problems that should be avoided in future applications of this method.

The samples played back to the respondents were not of only one particular feedback word; rather it was random which AES was coupled with which feedback word. The idea behind this was that there should no systematic influence from the semantics of the word on the interpretation. The drawback of this is that the prosodic measures are not as comparable between different instances as they would be if the same word were used for all cases. Different sounds have different intrinsic qualities in terms of pitch and intensity, and different words have different durations. The latter is more notable in the case of the word 'okej' which has a longer duration than the others, which are more similar to each other in this regard. However, these differences between the different words are much smaller than the differences in focus in our results, and can be provisionally disregarded. By selecting to focus on only one word at a time, this problem is avoided. With enough data these intrinsic differences of the words could also be compensated for.

In two of the twenty samples the pitch was not detectable, because of a creaky voice quality. This might have had an influence on some of the composite pitch values. While creaking of the voice can be a very interesting signal with regards to communication of AES's, it does not fall within the scope of this investigation. In future studies, the problem will be handled by making sure that all samples can be analysed in all dimensions beforehand.

The small number of samples produced by only two different speakers and the relatively small number of respondents means that there is very little chance of making any extended statistical inferences, or calculating significance or doing variance analysis. This was an expected problem, as this is a pilot study.

5 Discussion

The fact that the respondents were in next to complete agreement about the AES of certain samples, while there was very little agreement for others, begs the question if there are certain qualities of these samples that produce these results. Further research to establish whether certain prosodic qualities are more easily identified with specific AES's, and the inverse for those that are difficult to identify, is under way. We are also interested in identifying if there are specific AES's that are easier to identify using vocal signal alone, and whether others are more reliant on other bodily expressions.

With regard to any specific measures presented here, we generally concede that we have to little data to draw any firm conclusions yet. What we have presented are preliminary findings, which at can be taken as indications of the kind of results that we hope to present later, and as indications of general directions of what these results might show. Even so, we are encouraged that many of the results that we do see are in agreement with our preconceptions of how these AES's are expressed in Swedish. Many of the categories seem to be distinguishable in terms of their prosodic qualities. The fact that some seem to cluster together can be seen as a call for more research into these specific AES's, using both our present methodology as well as other kinds. A possible hypothesis might be that these AES's are not distinguished very clearly in terms of prosody, but might instead rely more on facial expressions or other bodily expressions.

Acknowledgements

The research that has led to this work has been supported by the NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social Sciences (NOS_HS) and the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287(SSPNet).

References

Allwood, J. (1988) Om det svenska systemet för språklig återkoppling. In: P Linell, V. Adelswärd & P. A. Pettersson (ed.) *Svenskans beskrivning* 16, vol. 1. Linköping: Tema kommunikation, Linköpings universitet.

Allwood, J., Nivre, J. & Ahlsén, E. (1992) On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1), 1-26.

Boholm, M., Lindblad, G. (2011) Head movements and prosody in multimodal feedback. *NEALT Proceedings Series: 3rd Nordic Symposium on Multimodal Communication*, pp. 25-32.