SOME ACOUSTIC CHARACTERISTICS OF EMOTION

Cécile Pereira
Catherine Watson

Speech, Hearing and Language Research Centre,
Macquarie University, Sydney, Australia
cpereira@elm.mq.edu.au
watson@srsuna.shlrc.mq.edu.au

ABSTRACT

This study presents an acoustic analysis of emotion. The material consisted of two semantically neutral utterances spoken by two actors, one male, one female, portraying three moods: anger, happiness and sadness; and a neutral tone. The duration, fundamental frequency (F0) and an estimate of the sound intensity (RMS) were analysed. The fundamental frequency parameter was the most revealing, showing differences between anger and happiness according to the shape of the contour, and between "cold" anger and "hot" anger on F0 mean. In addition, the study replicates previous findings showing hot anger and happiness having an F0 large range and high mean in contrast to the more subdued emotion of sadness, and the neutral voice.

1. INTRODUCTION

In a study on the perception of emotion by 40 normally hearing listeners, it was found that the different emotions of happiness, sadness, two forms of anger (hot and cold) and a neutral state were identified correctly 83% of the time. Most presentations of sadness were judged as intended (96% recognition rate), as was "hot" anger (86%) and happiness (85%), while "cold" anger (76%) was sometimes confused with neutrality (14%), and neutrality (78%) with sadness (21%) [1]. In order to identify the cues which listeners may be using to identify the different emotions, it was decided to conduct an acoustic study and a study of the prosodic structure and the intonation pattern of the corpus, as it was hypothesised that the signalling of emotion in speech is accomplished both through linguistic cues, e.g. the intonational contour of the utterance; and paralinguistic cues, e.g. F0 range.

Findings in the literature point to anger being characterized by an increase in fundamental frequency (F0) mean, F0 range, and F0 variability and mean intensity, with some evidence for an increase in high frequency energy. Downward directed F0 contours have also been reported, and an increase in rate of articulation [2]. Articulation rate is usually calculated by dividing the number of syllables by the overall utterance time, excluding pauses, but can also be reflected more sensitively by measuring the length of voiced segments [3].

Happiness also seems to be characterized by an increase in F0 mean, F0 range, F0 variability, and mean intensity, with some evidence for an increase in high-frequency energy [2].

The findings on sadness include a decrease in F0 mean, F0 range,
In a preliminary investigation, we looked at the duration and the mean and range of the F0 and RMS values across the entire utterances, as many other studies have done. However, we found that generally these measures were not distinguishing the emotions

and mean intensity. There is evidence that high frequency energy and rate of articulation decrease, and some evidence of downward directed F0 contours [2]. The F0 variability is small [3].

Scherer and Pittam have pointed out that increases in F0 mean, range and variability, increases in intensity mean and range, and high-frequency spectral energy indicate arousal [4][5]. Happiness and anger are considered high arousal emotions, while sadness and neutrality are considered low arousal ones. These authors write that in the literature "there is as yet little evidence for vocal differentiation of individual emotion states on similar levels of arousal" [5]. It was therefore of particular interest in our study to see if differences could be observed between anger and happiness, and sadness and neutrality.

2. METHOD

The corpus consisted of two semantically neutral utterances, spoken by two actors, one male and one female, and intended to express happiness, sadness, two forms of anger - one "cold" and one "hot" -, as well as a neutral tone to serve as a base for comparison. The utterances were: "I'm going to move house", and "Two thousand one hundred and ten". For their productions, the actors were guided by 10 framing stories which contextualised the utterances. These contextualising stories were based on the work of House (1989) for Swedish. There were in total 40 utterances studied: 2 utterances x 2 productions x 5 emotions x 2 speakers (these had been repeated 3 times in the perception study). All the data were labelled at the phonetic level, using ESPS/WAVES, and at the prosodic level.

The prosodic structure and the intonation pattern of the corpus utterances were examined and compared from two perspectives: that of the Hallidayan model [8], and that of ToBI, a proposed standard for labelling prosodic features of digital speech data bases in English [9].

For the acoustic analysis the F0, the RMS (an estimate of the sound intensity), and the duration of the utterances were calculated. The F0 values were calculated with ESPS/WAVES, using an algorithm developed by Secrest and Doddington [7]. The RMS was calculated from the amplitude of the speech signal using the root mean square of the amplitude (RMS). The RMS can be considered proportional to sound intensity [10]. A new RMS value was calculated for each contiguous 10 ms. of speech. The duration was calculated from the start and end times associated with the phonetic labels.

in our data clearly.

We then decided to examine the duration and the mean and range of the F0 and RMS values across all the non-schwa vowels, thus focussing on the segments which we thought most likely to carry the emotional information. For each vowel the corresponding portions of the F0 and RMS tracks were extracted, and average F0 and RMS values were obtained. We are assuming that these F0 and RMS mean values of the vowels represent the major trends which would be observed from the F0 and RMS tracks of the overall utterances. For each utterance the F0 range was calculated by subtracting the smallest F0 value found across all vowels from the highest value. The RMS range was calculated in a similar way. For the duration, the start and end times for each of the vowels under examination were extracted. In order to work out the importance of vowel and consonant length, we also worked out the ratio of all the vowels (including the unstressed ones) to the overall utterance duration.

The vowels selected were all stressed at least once, but generally more than once. For the utterance “I'm going to move house”, the vowels studied are, in chronological order, /ɑɪ/ /oʊ/ /u/ /ɑʊ/. For the utterance “Two thousand one hundred and ten”, they are /u/ /ɑʊ/ /ɑɪ/ / ʌ/ /ɛ/.

3. RESULTS

The results of the prosodic structure and the intonation pattern of the corpus utterances did not show any emotion specific differences which could be generalised, whether from the results of the Hallidayan analysis, or those of the ToBI analysis. We thus could not confirm our initial hypothesis that the signalling of emotion in speech is accomplished in part through linguistic cues, although, in acoustic terms, there was a pattern to the contours, as is described below.

The acoustic analysis was more productive, despite a certain amount of intra- and interspeaker variation. The results reported in this paper are those which apply to all speakers, utterances and productions, and are relative to neutrality.

The F0 analysis was the most revealing. For a display of the results of mean fundamental frequency values per vowel for each speaker and utterance, see Tables 1 to 4 at the end of this paper. These mean values and those of the RMS (not tabled) enabled us to observe the shape of the contour and which vowel carried the highest value. In summary, the following could be observed from the F0 analysis:

The F0 contour of neutrality, our reference state, is linguistically governed. For the sentence “I'm going to move house”, the contour starts with a rise to the vowel with the highest F0 value which is consistently that of “going”. That peak is followed by a continuous fall. For “Two thousand five hundred and ten”, the variety of contours reflect the possible different readings of that utterance, independently of emotion: either essentially one continuously falling contour, with thus an initial emphasis on *two*, and optionally a lesser one on *five*; or a falling contour ending with a small rise on the nucleus, *ten*.

Cold anger shows essentially a continuously descending contour, the first vowel carrying the highest F0 value; it has a larger F0 range than neutrality but a similar mean F0 value.

Hot anger also displays essentially a continuously descending contour, the first vowel carrying the highest F0 value; it too has a large F0 range but its mean F0 value is high relative to that of neutrality.

Happiness has an oscillating F0 contour three quarters of the time, with no particular vowel consistently having the highest F0 value; like both forms of anger it has a large F0 range, and like hot anger its mean F0 value relative to neutral is high.

Sadness has a tendency for a continuously descending contour, but may finish with a small terminal rise; its F0 range is similar to that of neutrality, while its mean F0 value is smaller than that of happiness and hot anger, but is marginally larger than that of cold anger and neutrality.

Overall, the results from the intensity analysis parallel the fundamental frequency findings and are therefore not reported. One point to note, however, is that sadness, which does not have the smallest F0 mean of all emotions, has the smallest RMS mean. This can be observed in the overall intensity figures. In terms of intensity, sadness is at least as low as neutral, but is usually the lowest .

Generally, measurement of duration findings cannot be generalised as there is an intra- and interspeaker variation, and they are therefore not reported here. However, for sadness, the figures indicate that the total of the duration mean of each stressed vowel tends to be small in comparison to the other emotions. This is not parallelled by a shorter length of total utterance duration, suggesting that for sadness the consonants are longer and the vowels shorter than in other emotions. This trend is confirmed by the value of the ratio of all vowels' duration to overall utterance duration, which for sadness tends to be small.

4. DISCUSSION

The findings of hot anger and happiness having an F0 and RMS large range and high mean in contrast to the more subdued emotion of sadness, and the neutral voice replicate the findings in the literature and confirm the distinction between high arousal and low arousal emotions. Of particular interest though was the evidence for differentiation of individual emotion on similar levels of arousal, that is the difference between anger and happiness according to the shape of the contour; and, less striking but nevertheless there, the differences between sadness and neutrality in their F0 contours, F0 and RMS means and the duration of the stressed vowel segments. Also of interest are the findings on cold anger, indicating a distinction with hot anger on F0 mean, and a distinction with neutrality on F0 range. Cold anger seems to share features of both the high arousal and the low arousal emotions.

While hot anger is found to share a number of features with happiness, cold anger with neutrality, and neutrality with sadness, it is worth restating that normally hearing listeners made the correct identification most of the time, although there were some confusions between cold anger and neutrality, and between neutrality and sadness. It may be that, most of the time, listeners can attune themselves to the variations reported here, such as different F0 contours, similar F0 means but different F0 ranges, and slight differences of contour, F0 mean, RMS mean, and duration of vowels relative to consonants. Other acoustic characteristics, such as the amount of high frequency energy or the voice quality, could also play a part in the identification of emotion, and will be the object of further research. Variation in all these acoustic parameters is important to identify as it has diverse and important applications, such as speech synthesis, the design of hearing aids and cochlear implants, or the auditory and communication training of hearing impaired people.

5. REFERENCES

1. Pereira, Cécile. 1996. Angry, happy, sad or plain neutral? The identification of vocal affect by hearing-aid users. In *Proceedings of the Sixth Australian International Conference on Speech Science and Technology,* Adelaide.

2. Pittam, Jeff, and Klaus Scherer. 1992. The encoding of affect: a review and directions for future research. In *Proceedings of the Fourth Australian International Conference on Speech Science and Technology,* Brisbane, 744–745.

3. Scherer, Klaus, Rainer Banse, Harald Wallbott, and Thomas Goldbeck. 1991. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15: 2, 123–148.

4. Scherer, Klaus. 1995. How emotion is expressed in speech and singing. In *Proceedings of ICPhS,* Stockolm.

5. Pittam, Jeffery, and Klaus Scherer. 1993. Vocal expression and communication of emotion. In *The handbook of emotion*, ed. M. Lewis, & J. Haviland, 185–197. New York: Guildford Press.

6. House, David. 1990. On the perception of mood in speech: implications for the hearing impaired. In *Working Papers* 36, 99-108, Lund University, Department of Linguistics.

7. Secrest, B.G., and G.R. Doddington. An integrated pitch tracking algorithm for speech systems. In *Proceedings ICASSP83*, 1325-1355.

8. Halliday, Michael A.K. 1967. Intonation and grammar in British English. The Hague: Mouton.

9. Beckman, Mary E., and G.M. Ayers. 1994. Guidelines for ToBI labelling guide, ver. 2.0. Manuscript, Ohio State University, Columbus, OH.

10. Clark, John, and Colin Yallop, Colin. 1990. An Introduction to Phonetics and Phonology. Oxford: Basil Blackwell.

	ɑɪ	oʊ	u	ɑʊ
cold anger	192	164	153	122
hot anger	263	232	291	207
happiness	249	228	242	199
neutrality	176	192	157	135
sadness	267	255	234	216

Table 1: Mean F0 of the stressed vowels of “I'm going to move house” (in order of occurrence) for the female speaker.

	ɑɪ	oʊ	u	ɑʊ
cold anger	135	115	83	82
hot anger	197	188	137	111
happiness	173	206	123	195
neutrality	93	105	85	81
sadness	108	110	98	97

Table 2: Mean F0 of the stressed vowels of “I'm going to move house” (in order of occurrence) for the male speaker.

	u	ɑʊ	ɑɪ	ʌ	ɛ
cold anger	248	123	165	149	105
hot anger	398	229	306	284	210
happiness	300	242	198	200	159
neutrality	196	168	163	155	152
sadness	251	195	205	190	183

Table 3: Mean F0 of the stressed vowels of “Two thousand five hundred and ten” (in order of occurrence) for the female speaker.

	u	ɑʊ	ɑɪ	ʌ	ɛ
cold anger	259	195	185	153	130
hot anger	244	171	183	160	140
happiness	224	195	169	164	209
neutrality	112	98	94	94	96
sadness	139	109	100	99	103

Table 4: Mean F0 of the stressed vowels of “Two thousand five hundred and ten” (in order of occurrence) for the male speaker.