A MACROSCOPIC ANALYSIS OF AN EMOTIONAL SPEECH CORPUS

J.E.H. Noad[1], *S.P. Whiteside*[1] *and P.D. Green*[2]
[1]Department of Human Communication Sciences and [2]Department of Computer Science
University of Sheffield
Sheffield, S10 2TN, England.
J.E.Noad@shef.ac.uk, S.Whiteside@shef.ac.uk, P.Green@dcs.shef.ac.uk

ABSTRACT

Macroscopic analysis of a corpus of emotional Standard Southern British speech signals has been performed to measure any changes in average fundamental frequency, speech rate, energy and first formant frequency. Seven acted emotional states were recorded and analysed for one male and one female speaker. Differences between neutral and emotional speech were found which agree with changes others have mentioned in the literature. Only the emotion sadness was found to be consistently and obviously different from neutral, while high activity emotions (such as elation and hot anger) could be distinguished from sadness. Additional measures are being developed which will further discriminate the emotions from one another. Results obtained to date are being evaluated for use in a speech synthesiser system.

1. INTRODUCTION

The study of emotional speech has become increasingly important in a number of areas. Speech synthesiser systems are now of a sufficiently high quality that they are capable of attempting to portray more than just one characterless voice [1],[2].

In the Automatic Speech Recognition community, there is a growing awareness that if the performance drop between human and machine recognition [3] is to be reduced, it will be necessary for recognisers to do more than consider variations in speaking style as noise to be averaged out.

Other work is attempting to assess the emotional state of a speaker from their speech parameters [4] so that a system may adjust its behaviour as appropriately.

Of the small amount of work on vocal emotion in the literature most has concentrated on features at the phone and syllable level. Very few empirical results have been published. This is because of the small size and number of corpora that are available.

The emotions shown in table 1 were chosen to make up the corpus. These are a sub-set of those being studied by Scherer et. al. [5],[6]. This set was chosen because they would be of most benefit to users of speech synthesiser systems. The emotions are a mixture of primitive primary emotions and common secondary emotions. They also cover a range of valence, strength and activity [7]. Primary emotions are expected to cause similar effects across languages because of their biological origins.

Emotion	Mnemonic
Cold Anger	CA
Elation	EL
Hot Anger	HA
Happiness	HP
Interest	IN
Sadness	SA
Neutral	NE
Neutral (extended)	NE2

Table 1: Emotions recorded and their mnemonics

2. METHOD

Recordings were made of acted read speech. The texts consisted of five emotionally neutral passages each comprising four or five sentences. A number of short sentences, isolated words and words in a constant context were also recorded at the same time. To account for normal variations in speech a second separate extended session of neutral speech (NE2) was recorded. This is made up of recordings of all of the same texts as the other sessions along with additional material while maintaining the proportions of data types in the NE data. For each emotion there is approximately two minutes of speech per speaker and the NE2 data lasts four minutes per speaker.

An example of the data recorded is passage ***P1***:
"Last year I went to a wine tasting evening. The event took place in a majestic room with cut glass chandeliers hanging from the ceiling. There were at least ten wine bottles on each of the dozen tables with several dozen elegant wine glasses. However I am sure the intention was only to have a good mouthful of each and not a whole bottle as one very drunken person thought."

Recordings were made using a single table-top microphone in a quiet room using Sony DAT equipment. The recordings were then transferred onto a computer using an OROS AU21 card and a sampling rate of 20kHz.

Two actors were recorded for the corpus. Speaker RCT is a 27 year old female and RP is a 32 year old male. Both speakers are non-smokers, have no speech, language, or hearing difficulties and have Standard Southern British (SSB) accents. RCT has performed in several amateur productions over the last three years and RP has been a professional actor for eleven years.

Several sets of measurements were made from the corpus. Analysis of the signals was performed using programs from the Speech Filing System (SFS) [8], in combination with programs written by the authors.

2.1. Fundamental Frequency

The fundamental frequency (F0) contours were calculated on a frame-by-frame basis using the cepstral and auto-correlation methods, as well as from the time separation of pitch pulses. These three data sets were combined so as to reduce the impact that the inevitable erroneous points have. From this information a distribution of F0 values for the emotive speech is obtained. The mean F0 and standard deviation are then calculated.

2.2. Speed of Speaking

The total duration of each passage of speech was measured. This time included the inter-sentence silences which in themselves contribute to the emotional effect. These times were totalled together for each emotion. To make the data more easily comparable with the other measurements made the durations were converted to an overall speaking rate by taking the reciprocal. The term 'speech rate' is normally applied to the rate of syllable production - this is a localised measure whereas the speed of speaking is global.

2.3. Energy

The Root Mean Square (RMS) amplitudes of the speech waveforms were calculated. Because all of the data were recorded in one session with no adjustments to the recording equipment it was possible to compare the signals without requiring the presence of a known calibration signal.

2.4. Formants

The 'high-quality' covariance LPC analysis program in SFS was used to estimate formant positions. Formant frequencies only from the frames marked as voiced were used to build up a distribution of formant frequencies. From this information a mean and standard deviation were determined.

3. RESULTS

Figure 1 shows the F0 distributions obtained for speaker RP's portrayal of sadness, neutral and hot anger. Results for the other emotions are not plotted because the graph would become overly cluttered. Sadness has a single narrow high peak at around 100Hz, neutral somewhat broader peak while hot anger is broader still. The area under each curve is related to the number of frames of speech observed, which is directly related to the speaking rate. The curves are comparable because they are obtained from readings of the same texts. The distributions are asymmetric because of the more rapid decrease in F0 at the start of a phrase than towards the end

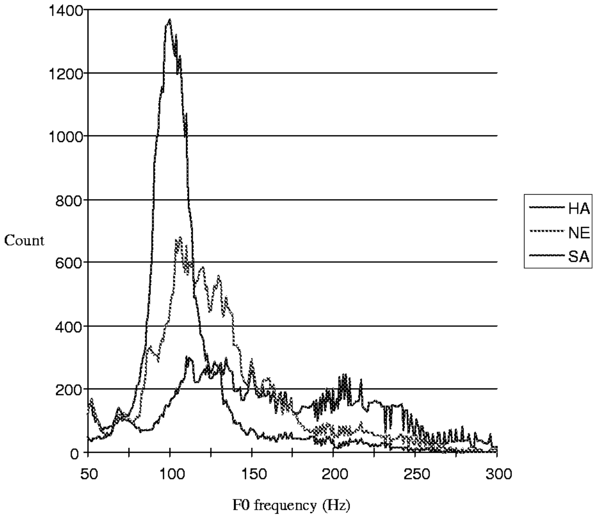

Figure 1: F0 distributions of RP's hot anger, sadness and neutral

In the figures 2 to 7 the sets of results have been plotted on a normalised scale to make comparison easier. The average of each measurement over a speaker's two neutral portrayals was defined to have the value 1.0. The measurements for each emotion were then plotted relative to this. Because any changes are relative to a speaker's neutral baseline inter-speaker comparison is made easier.

Figure 2 shows the results obtained for mean F0. The average F0 for RCT's neutral speech was 192Hz, and 142Hz for RP's. Figure 3 shows the standard deviation from the mean F0. RCT's average neutral F0 standard deviation was 61Hz and RP's was 62Hz. The speaking rate results are shown in figure 4. The Root Mean Square (RMS) values of the signals are shown in figure 5. Figure 6 shows the relative first formant frequency. For speaker RCT the mean neutral F1 frequency is

467Hz and for RP it is 469Hz. Figure 7 shows the standard deviation of the mean of F1. RCT's average neutral F1 standard deviation was 90Hz and RP's 79Hz.

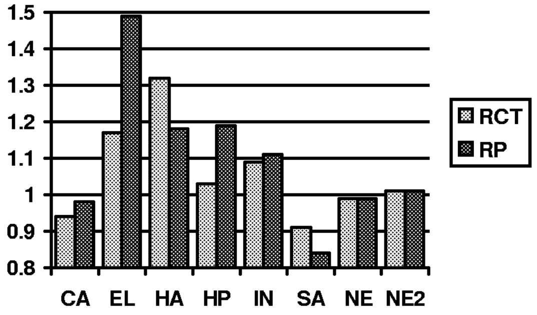

Figure 2: Relative mean F0 frequency

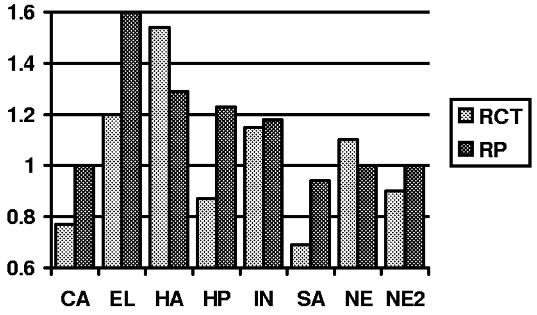

Figure 3: Relative standard deviation of F0 distribution

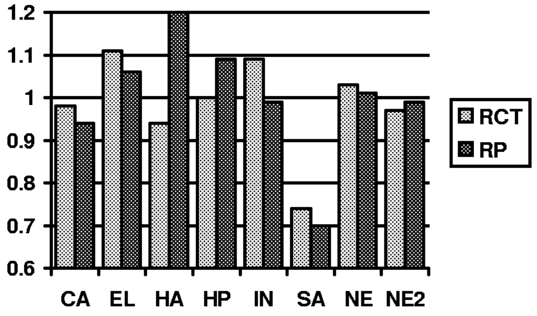

Figure 4: Relative speed of speaking

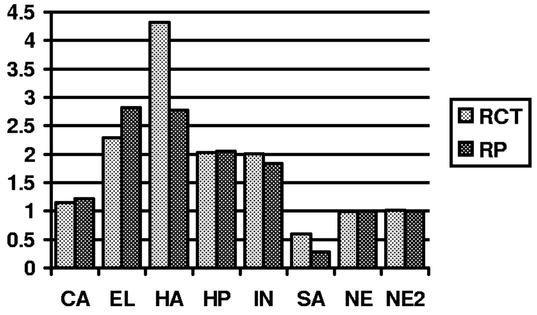

Figure 5: Relative RMS energy

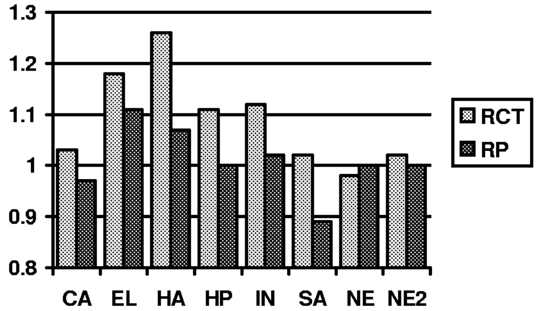

Figure 6: Relative mean F1 frequency

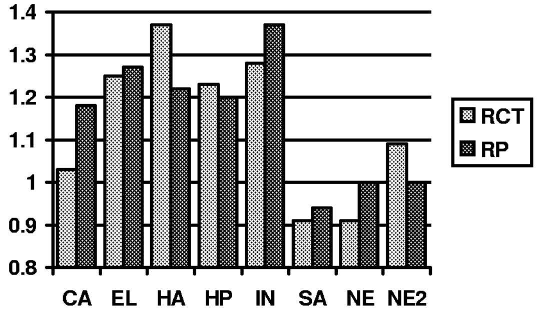

Figure 7: Relative standard deviation of mean F1

4. DISCUSSION

Comparison of the two neutral recordings for both speakers shows them to be remarkably similar. The neutral results were always the most similar pair of results both between emotions and between speakers with the exception of F1 standard deviations.

The average F0 increases for all of the emotions except cold anger and sadness. This replicates what has been reported elsewhere [7],[9].

There is little obvious and consistent change in speech rate except for sadness which shows a large decrease. It is expected that speech rate is related to activity level and to a certain extent also to emotion strength. RCT's hot anger is therefore unusual because despite this it has a lower speech rate than her neutral speech.

The RMS energy shows the widest and most consistent variation. Emotions with high activity (elation and hot anger) have increased signal energies. Conversely the low activity emotion (sadness) has a low signal energy. The major disadvantage of using RMS energy as a measure of emotion is that it is only meaningful when a signal can be calibrated against a neutral signal recorded under identical conditions. Because humans have a historical knowledge of a wide range speakers, emotions and an understanding of the words being spoken to draw on, it is possible to use signal energy as a secondary emotional cue.

The first formant is expected to increase in frequency when the mouth is more open. This was found to be the case for hot anger by [9]. Formant frequencies were observed to increase by Tartter and Braun [10] when a speaker was smiling, and decreased when a speaker was frowning, even if they were not attempting to convey emotion. Consequently one would expect that increased formant frequencies would be found in happiness and elation where the speaker will normally be smiling, as well as in hot anger where the mouth is more open. This does appear to be the case from figure 6.

Actors also add extra cues [11] such as sighs, lip-smacks and loud inhalation to help convey emotion. RCT makes particular use of deliberately loud inhalation in elated and happy passages.

The variation among the emotions may be due to a number of causes. The most important are the different sexes of the speakers, the difference in acting experience, the fact that all speakers are unique and how idiosyncratic each emotion is. Despite this the general trend across the emotions for both RCT and RP is similar.

These measures have been applied to large sections of emotional speech and changes found from neutral speech. While this produces good results for primary emotions, secondary emotions are less distinguishable. This is likely to be because the emotion is conveyed continuously in more subtle ways, or during critical segments of speech.

5. CONCLUSIONS AND FUTURE WORK

The measures used in this paper show that sadness is easy to differentiate analytically from neutral speech and other emotions. The other emotions are more difficult to characterise although hot anger and elation are found at the opposite end of the scale to sadness.

Perception tests with the corpus, re-synthesised and synthesiser-generated speech are under way. Further long term measures are being developed which it is hoped will improve on the results given here. Examining the effectiveness of transforming neutral into emotive speech while maintaining listeners high recognition rates is the next step in establishing the validity of the measures in this paper.

To clarify the general significance of the measures it will be necessary to record more speakers. The additional difficulty in comparing speakers is that some emotions have broader stereotypes than others. It may also be that different sexes portray the same emotion in different ways.

6. ACKNOWLEDGEMENTS

Our special thanks go to our actors, without whom this work would not have been possible. This work has been funded by the EC TIDE project TP-1174 VAESS, EPSRC and Barnsley District General Hospital NHS Trust.

7. REFERENCES

[1] J.E. Cahn, "Generating Expression in Synthesized Speech", Master's Thesis, MIT, 1989.

[2] I.R. Murray, J.L. Arnott and A.F. Newell, *"HAMLET - Simulating Emotion in Synthetic Speech",* Proc. Speech'88, The 7th FASE Symposium, pp. 1217-1223, Edinburgh, 1988.

[3] R.P. Lippmann, *"Speech Perception by Humans and Machines",* Proc. ESCA-NATO workshop on the Auditory Basis of Speech Perception, Keele, UK, July 1996.

[4] T. Moriyama, H. Saito and S. Ozawa, *"Evaluation of the Relationship Between Emotional Concepts and Emotional Parameters on Speech",* Proc. ICASSP'97, pp. 1431-1434, Munich, 1997.

[5] K.R. Scherer, "Vocal affect expression as symptom, symbol and appeal", in "Non-verbal vocal communication: Comparative and developmental approaches", H. Papousek, U. Jürgens and M. Papousek (Eds.), Cambridge University Press, 1992.

[6] R. Banse and K.R. Scherer, "Acoustic profiles in vocal emotion expression", *J. Personality and Social Psychology,* Vol. 70(3), pp. 614-636, 1996.

[7] I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *JASA*, Vol. 93(2), pp. 1097-1108, 1993.

[8] M.A. Huckvale, "SFS for Users", Department of Phonetics, UCL. ftp://pitch.phon.ucl.ac.uk/pub/sfs, 1996.

[9] C.E. Williams and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates", *JASA*, Vol. 52(4), pp. 1238-1250, 1972.

[10] V.C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers", *JASA*, Vol. 96(4), pp. 2101-2107, 1994.

[11] B. Granström and L. Nord, *"Ways Of Exploring Speaker Characteristics and Speaking Styles",* Proc. ICPhS'91, pp. 278-281, Aix-en-Provence, 1991.