Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis

Marc Schröder[*], *Roddy Cowie*[+], *Ellen Douglas-Cowie*[+], *Machiel Westerdijk*[o] *& Stan Gielen*[o]

[*]DFKI, Saarbrücken; Institute of Phonetics, University of the Saarland, Germany
[+]Queen's University, Belfast, Northern Ireland
[o]University of Nijmegen, The Netherlands
schroed@dfki.de

Abstract

In a database of emotional speech, dimensional descriptions of emotional states have been correlated with acoustic variables. Many stable correlations have been found. The predictions made by linear regression widely agree with the literature. The numerical form of the description and the choice of acoustic variables studied are particularly well suited for future implementation in a speech synthesis system, possibly allowing for the expression of gradual emotional states.

1. Introduction

Emotion dimensions are a simplified description of basic properties of emotional states. The most frequently encountered emotion dimensions are activation (the degree of readiness to act), evaluation (in terms of positive and negative, of liking and disliking) and power (or dominance/submission)[1]. While they do not capture all the relevant aspects of an emotional state, they provide a taxonomy allowing simple distance measures between emotion categories, as well as a continuous framework for representing gradual, non-extreme emotional states.

In speech synthesis, attempts to express emotions have so far concentrated on a small number of discrete, extreme emotion categories, basically aiming to obtain maximally distinguishable prosodic profiles [14]. However, for applications, it would be desirable to be able to model weak emotions [2], as well as the gradual change of emotional tone over time. Both could be achieved with the help of emotion dimensions if emotions were appropriately represented and if acoustic correlates of emotion dimensions in speech could be found.

This article investigates a database of spontaneous emotional speech [7], searching for systematic correspondences between perceptually determined positions on emotion dimensions [3][4] and acoustic variables [5][6] relevant for speech synthesis.

2. Belfast Naturalistic Emotion Database

The Belfast Naturalistic Emotion Database contains audio-visual recordings of 100 English speakers exhibiting relatively spontaneous emotion [7]. The material contains TV recordings of chat shows and religious programs, as well as interviews recorded in a studio. In terms of scale and range of emotions, it appears to be the largest collection of natural emotional speech available. About 85% of the speakers included in the database are female.

The database has been perceptually annotated with respect to emotional content using the FEELTRACE tool [4], which has recently been made publicly available [8]. After a necessary training phase, subjects can locate the emotional tone of a clip in the 2-dimensional activation-evaluation space, continuously over time. In addition, each clip is labelled using the words of a Basic English Emotion Vocabulary [3]. The rich characterisation of these emotion words obtained in previous experiments [3] allows the addition of the power dimension, as associated with the emotion word, to each clip. Thus each clip is positioned on the three dimensions, with activation and evaluation changing over time during the clip and power remaining static.

The acoustic analysis of the database has been performed semi-automatically using the ASSESS system [6]. It generates a simplified core representation of the speech signal based mainly on the F0 and intensity contours. Key 'landmarks' are then identified, including peaks and troughs in the contours as well as boundaries of pauses and fricatives. Measuring the 'pieces' between these landmarks gives rise to a range of variables called 'piecewise'. They provide a rich description of the way contours (of pitch and intensity) behave over time. Variables, piecewise and others, are then summarised in an array of statistics (covering central tendency, spread and key centiles). Additional measures deal with properties of 'tunes' (i.e. segments of the pitch contour bounded at either end by a pause of 180 ms or more) as well as with spectral properties.

3. Acoustic variables relevant for speech synthesis

For speech synthesis, the most interesting acoustic variables are those which can be set in speech synthesis systems. The following variables have been identified for which corresponding ASSESS variables can be found.

For intonation,

- the global settings F0 mean and range,
- accent structure: number of F0 maxima / minima per time unit, duration, magnitude and steepness of F0 rises and falls.

For tempo,

- duration of pauses,
- articulation tempo.

[1] Other names for the same dimensions are frequently encountered, e.g. Arousal, Valence and Control. These dimensions were first proposed by Schlosberg [13].

For intensity[2],

- the global settings intensity mean and range,
- a simple measure of dynamics (the difference between mean intensity for intensity maxima and overall mean intensity).

For voice quality,

- spectral slope,
- approximations of the Hammarberg indices[3].

In total, 26 variables have been selected.

It is clear that not all interesting information can be provided by a semi-automatic analysis tool like ASSESS. For example, as no phone boundaries are detected, only approximate measures of articulation tempo are available, such as the length of 'tunes' (inter-pause stretches), the number of intensity peaks per second, or the number of fricative stretches per second. Other interesting information, such as the location of pauses with respect to sentence structure, or articulation precision, is not available at all, because it would require linguistic information ASSESS does not have.

4. Analysis

The dimensional ratings, on scales from –100 to 100, were aligned with the acoustic analyses of the individual 'tunes' (inter-pause stretches) within each clip. A total of ca. 5500 data points was thus obtained, ca. 4700 for female speakers and ca. 850 for male speakers.

Correlation and linear regression analyses of the data were performed using SPSS. The data was analysed separately for male and female speakers. The three emotion dimensions Activation (A), Evaluation (E) and Power (P) as well as their squares (A^2, E^2, P^2) were used as independent variables, predicting in turn each of the acoustic variables selected. The regression coefficients calculated by this analysis predict the value of the acoustic variable at a given point in the 3-dimensional emotion space.

The reason for incorporating the squared terms was that this allowed for the modelling of simple non-linear behaviour.

It is important to note that the regression coefficients for different independent variables are influenced by each other. As a consequence, the set of coefficients obtained is only optimal in precisely this combination. For example, if the input to be used for calculating acoustic variables specified activation and evaluation, but not power, then it would not be optimal to use the coefficients calculated here.

For the purpose of interpretation, only those correlations that were significant at the level of $p<.05$ or better were considered. As the number of data points available for female speech was about five times larger than for male speech, more significant correlations were found for female than for male speech.

[2] Intensity descriptions as fine-grained as for F0 would have been available from ASSES, but were not considered due to the fact that usually, speech synthesis systems do not allow the same degree of control over intensity as over F0.

[3] Hammarberg and her co-workers [9] showed that differences in voice quality were related to differences in the maximum intensity in each of three spectral bands. The measures used here reflect the differences specified by the Hammarberg group in the simplest possible way, ignoring coefficients and additional (pitch-related) variables.

5. Results

The most fundamental result is that nearly all acoustic variables show substantial correlations with the emotion dimensions. The null hypothesis that no systematic correlation between emotion dimensions and acoustic variables exists can thus be rejected.

The most numerous and strongest correlations were found for the activation dimension. Most acoustic variables correlate with activation, in the sense that expression of active emotion is accompanied by higher F0 mean and range, longer phrases, shorter pauses, larger and faster F0 rises and falls, increased intensity, and a flatter spectral slope.

Correlations with the evaluation dimension are less numerous and less strong, but systematic. Expression of negative emotion is accompanied by longer pauses, faster F0 falls, increased intensity, and more prominent intensity maxima.

As to correlations with the power dimension, higher power is accompanied by lower F0 mean; in addition, for female speakers, F0 rises and falls are less steep, and F0 falls have a smaller magnitude. While for female speakers, intensity is reduced, it is increased for male speakers.

Tables 1 and 2 show the regression coefficients, for female and male speakers respectively, for nine acoustic variables frequently used in emotional speech synthesis [14].

6. Discussion

The literature contains very few studies that use a database comparable in scale and naturalness. Nevertheless, it is useful to consider how our findings compare to published studies.

Direct comparisons are possible only with studies having examined the vocal correlates of emotion dimensions more or less explicitly. Pereira [12] has conducted such a study using two actors' portrayals of two sentences in a number of emotional states. Our findings are in agreement with hers in that activation positively correlates with F0 mean, F0 range and mean intensity, that correlations with evaluation are less strong, and that the power dimension is correlated positively to mean intensity for male speech. For female speech, however, we find a weak negative correlation of power to mean intensity rather than the positive correlation she finds. In addition, the positive correlations of power to F0 mean and range for male speech that Pereira reports cannot be confirmed from our data.

Banse & Scherer [1] analyse their data, among other things, in terms of "intense" emotions such as despair, hot anger, panic fear, and elation, for which the highest F0 mean (p. 624) and mean energy (p. 627) are found, which is in agreement with our findings when "intense" is interpreted as a high activation level. Banse & Scherer's observation that low coping potential (power) leads to a high energy proportion in low frequency bands (a steep negative spectral slope) cannot be found in our data. Rather, we find the opposite: Low power is accompanied by a flatter spectral slope for both female and male speakers.

The statement made in the literature about a frequency code type of voice use [11], according to which dominance (high power) is signalled through low F0, is supported by our data. The power dimension correlates negatively with F0 median for both female and male speakers.

Emotion	Activation	Evaluation	Power
neutral	0	0	0
sad	-8.5	-42.9	-55.3
angry	34.6	-34.9	-33.7
afraid	31.1	-27.1	-79.4
happy	28.9	39.8	12.5

Table 3. Positions on the three emotion dimensions for some emotion categories.

An interesting link to the larger literature on vocal emotion expression can be established through the positioning of emotion categories in the 3-dimensional emotion space studied here. Table 3 shows the positions for a few emotion categories out of the BEEVer vocabulary [3]. For activation and evaluation, these positions are mean FEELTRACE positions for clips which were assigned the given verbal emotion label by the same rater [3]; for power, the value was associated with the verbal emotion label in the BEEVer study [3]. For a given position in that space, the rules presented above set the acoustic variables to the values that were found optimal in linear regression analysis of the corpus. The acoustic properties provided by these rules for the position of a given emotion category can then be compared to the findings in literature about the acoustic properties of that emotion category. This allows the distinction between acoustic settings that are due to general features of emotion, as expressed in the emotion dimensions, and acoustic settings that are due to more specialised features of that emotion.

For example, anger is described in literature as showing a very much higher pitch average compared to neutral speech, much wider pitch range, abrupt pitch changes on stressed syllables, slightly faster speech rate, higher intensity, "chest tone" voice quality and tense articulation [8]. The position of 'angry' in the emotion dimensions (A=34.6, E=-34.9, P=-33.7) predicts higher F0 median[4]; wider F0 range; steeper F0 rises and falls; higher intensity; and a flatter spectral slope, i.e. more high frequency energy, corresponding to a tenser voice quality. Thus, much of what is presented as typical for anger can be predicted based on the three emotion dimensions.

The rules for expression of emotion according to emotion dimensions also provide the possibility to track acoustic similarities between emotions to shared emotional characteristics. Comparing anger to fear, for example, it can be seen that the two share acoustic properties as well as emotional properties. Acoustically, according to [8], fear is similar to anger in pitch average, pitch range, speech rate and articulation. Fear differs from anger in that pitch changes are not steeper than for neutral, the speech rate is even faster, the intensity is only normal, and voicing is irregular. On the three emotion dimensions, the two emotions are very close on the activation and evaluation dimension, but differ in power.

A stepwise application of the regression rules can now account for some of the acoustic similarities in terms of emotion dimensions: In a first step, the acoustic correlates of the activation and evaluation levels, nearly equal for anger and fear, are calculated. This leads to an increase of F0 median and range compared to neutral, steeper F0 rises and falls, increased intensity, and a flatter spectral slope. In a second step, the rules for the power dimension are applied using the different values for anger and fear. The power value for the latter being more extreme, the effects for fear are stronger than for anger. According to the regression rules, this leads to an even higher F0 median for fear compared to anger; a slightly higher F0 range for female voices; even steeper F0 rises and falls for female voices; and a lower intensity. Most of this is in accordance with the acoustic profiles as summarised in [8].

Similarly, the acoustic similarities of anger and happiness (higher pitch average, pitch range and intensity according to [8]) can be "explained" by the similarity in activation level, while some of the differences can be tracked to evaluation (slower and smaller F0 falls) and to power (slower F0 rises and falls).

It is important to bear in mind that the three emotion dimensions are not expected to capture all relevant characteristics of emotions. Emotion words such as 'worrying' or 'loving', for example, include highly specialised properties [3] which may have some acoustic effects. These would not be accounted for by the dimensional description employed here using only three dimensions.

7. Conclusion

This study has explored an emotional speech database in view of correlations between emotional characteristics in terms of emotion dimensions, obtained by listener judgement, and acoustic properties measured using a semi-automatic tool. The correlations have shown to be numerous, which in itself is an interesting result in a corpus as heterogeneous as the one used. In addition, the linear regression coefficients obtained seem to allow reasonable predictions of acoustic variable values, as judged by comparing to the literature.

A method for tracking acoustic similarities of emotions to similar emotional characteristics has been proposed, using stepwise application of the regression rules. It can be used to distinguish the effects of general emotional properties, captured through the emotion dimensions, from those arising from more specialised properties.

The rules obtained for the expression of emotions represented with emotion dimensions can now be implemented in a speech synthesis system. Contrarily to previous emotional speech synthesis systems, the resulting system will be able to express continuous grades of emotional meaning. While no claim is made that all relevant aspects of vocal expression of an emotion located at particular coordinates in the 3-dimensional emotion space are captured, the resulting vocal effects should be at least roughly compatible with such an emotion, as defined through situative or verbal context, or through different modalities. Obviously, it will be necessary to evaluate the results obtained in a synthesis system, e.g. in terms of improved naturalness or greater overall preference compared to neutral speech in the same emotional situation.

[4] Calculation example for female speech:
median F0 = 193.4 + 0.357*A + 0.00439*A^2 + 0.00245*E^2 - 0.215*P + 0.00096*P^2 = 222.3 Hz, compared to 193.4 Hz for neutral speech. See Table 1.

8. References

[1] Banse, R., & Scherer, K. R., Acoustic Profiles in Vocal Emotion Expression, *J. Pers. Soc. Psy., 70(3)*, 1996, p. 614-636.

[2] Cowie, R., Describing the Emotional States Expressed in Speech, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 11-18.

[3] Cowie, R., Douglas-Cowie, E., Appolloni, B., Taylor, J., Romano, A., & Fellenz, W., What a neural net needs to know about emotion words, in *Computational Intelligence and Applications* (N. Mastorakis, ed.), 1999, p. 109-114. World Scientific & Engineering Society Press.

[4] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M., 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 19-24.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz & J. Taylor, Emotion Recognition in Human-Computer Interaction. *IEEE Signal processing Magazine, 18 (1)*, p. 32-80, January 2001.

[6] Cowie, R., Sawey, M., & Douglas-Cowie, E., A new speech analysis system: ASSESS, *Proc. ICPhS 1995, Stockholm, 3*, p. 278-281.

[7] Douglas-Cowie, E., Cowie, R., & Schröder, M., A New Emotion Database: Considerations, Sources and Scope, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 39-44.

[8] Feeltrace download under http://143.117.150.84/~guest

[9] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. & Wedin, L., Perceptual and acoustic correlates of abnormal voice quality, *Acta Otolargngologica, 90*, 1980, p. 441-451.

[10] Murray, I. R., & Arnott, J. L., Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *J. Acoust. Soc. Am., 93(2)*, 1993, p. 1097-1108.

[11] Ohala, J. J., The frequency code underlies the sound-symbolic use of voice pitch, in *Sound symbolism* (L. Hinton, J. Nichols, & J. J. Ohala, eds.), 1994, p. 325-347. Cambridge: Cambridge University Press.

[12] Pereira, C., Dimensions of emotional meaning in speech, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 25-28.

[13] Schlosberg, H., Three Dimensions of Emotion, *Psychol. Rev. 61(2)*, p. 81-88, 1954.

[14] Schröder, M., Emotional Speech Synthesis: A Review, this volume.

Acoustic variable	**Unit**	**Linear Regression Coefficients**						
		Constant	**A**	**A²**	**E**	**E²**	**P**	**P²**
F0 median	Hz	193.4	0.357	0.00439		0.00245	-0.215	0.000957
F0 range	Hz	27.84	0.225	0.00186	-0.0420		-0.0472	
'tune' duration	sec	1.405	0.00418					
pause duration	sec	0.407	-0.00119			0.0000143		
med. steepness F0 rises	Hz/sec	74.62	0.338	0.00320			-0.231	
med. steepness F0 falls	Hz/sec	83.89	0.434	0.00440	-0.125		-0.228	
median intensity	cB	529.1	0.0981	0.00197	-0.0735	0.00183	-0.105	-0.00513
range intensity	cB	99.83	0.119		-0.0696	-0.00107	-0.0742	
spectral slope non-frics.	dB/oct	-7.532	0.0110	0.000184		0.0000659		-0.0000568

Table 1. Linear regression coefficients for acoustic variables predicted by emotion dimensions, for female speech.

Acoustic variable	**Unit**	**Linear Regression Coefficients**						
		Constant	**A**	**A²**	**E**	**E²**	**P**	**P²**
F0 median	Hz	140.3	0.452	0.00824		-0.00243	-0.194	
F0 range	Hz	21.57	0.248					
'tune' duration	sec	1.417	0.0107					
pause duration	sec	0.377	-0.00139		-0.00123	0.0000655		-0.0000332
med. steepness F0 rises	Hz/sec	60.73	0.469			0.00420		
med. steepness F0 falls	Hz/sec	60.99	0.383		-0.168	0.00336		
median intensity	cB	520.2	0.305		-0.231		0.185	
range intensity	cB	103.8				-0.00421		
spectral slope non-frics.	dB/oct	-7.923	0.00993	0.000206	0.000476	-0.000102	-0.00890	0.000221

Table 2. Linear regression coefficients for acoustic variables predicted by emotion dimensions, for male speech.