Pitch Contour Parameterisation based on Linear Stylisation for Emotion Recognition

Vidhyasaharan Sethu, Eliathamby Ambikairajah, Julien Epps

School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
National Information Communication Techonology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia
vidhyasaharan@gmail.com, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

The pitch contour contains information that characterises the emotion being expressed by speech, and consequently features extracted from pitch form an integral part of many automatic emotion recognition systems. While pitch contours may have many small variations and hence are difficult to represent compactly, it may be possible to parameterise them by approximating the contour for each voiced segment by a straight line. This paper looks at such a parameterisation method in the context of emotion recognition. Listening tests were performed to subjectively determine if the linearly stylised contours were able to sufficiently capture information pertaining to emotions expressed in speech. Furthermore these parameters were used as features for an automatic 5-class emotion classification system. The use of the proposed parameters rather than pitch statistics resulted in a relative increase in accuracy of about 20%.

Index Terms: pitch stylisation, emotion classification, prosodic features, speech synthesis

1. Introduction

Expressing and recognising emotions is an integral part of human communication and humans are able to do so through a variety of means including speech. This ability to recognise emotions from speech is robust towards different speakers and humans are able to do so successfully in many cases even if it is the first time they are exposed that speaker. This suggests the existence of speaker-independent patterns in speech parameters that are characteristic of the emotion being conveyed. These characteristic patterns may exist in many levels, ranging from prosodic and acoustic patterns to patterns in word and language usage and form the basis of all automatic emotion recognition systems. Among these, prosody patterns are often treated as universal cues for emotion in emotional speech synthesis literature [1-3]. While most studies agree on the importance of global prosodic parameters such as F0 level, F0 range, loudness and rate of speech; F0 contours are taken into account less frequently even though they have been shown to play an important role in emotion recognition [4], [5].

The focus of the work reported in this paper is on parameterisation of F0 (used interchangeably with the term 'pitch' in this paper) contours and the use of these parameters as features for automatic emotion classification. Linear stylisation of F0 contours [6-8] is commonly carried out to make them simpler to analyse, but have the additional advantage of making them more compact. Approximating the pitch contour in each voiced segment by a straight line enables the representation of that contour by three parameters. This is different from typical F0 contour stylisations [6], [7] since each voiced segment is approximated by a single linear segment rather than a piecewise linear approximation. A subjective comparison of speech synthesised using the approximate F0 contour, speech synthesised using the actual estimated F0 contour and the actual speech sample allowed us to determine if linear approximation of F0 segments is representative of emotions. This was then validated by using the linear parameters as features for an automatic classification system as described in section 5.

2. Linear Approximation

The RAPT algorithm for pitch estimation [9] was used to estimate pitch contours from speech. A separate voicing activity detector (VAD) was used to identify voiced segments prior to linear curve fitting of the pitch contours in these segments as in [8].

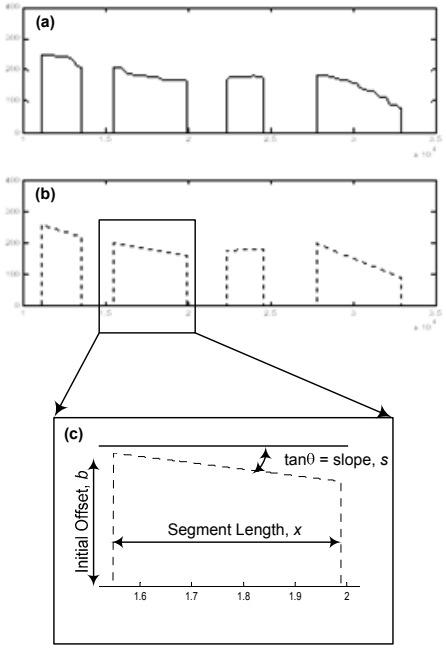

Figure 1: *(a) Estimated F0 contour (b) Linear approximation of F0 contour (c) Linear model parameters - b, s and x*

The linear approximation in each segment is represented by the slope of the line (s), the initial offset (b) and the length

of the segment (x) as shown in figure 1. Thus the pitch contour of any speech sample can be represented by $3N$ parameters, where N is the number of voiced segments in the utterance.

3. Speech Synthesis

In the work reported in this paper, the purpose of speech synthesis is to enable subjective comparisons of speech samples that use the estimated pitch contour and speech whose pitch contours have been replaced by linear approximations. Given the sole focus on pitch, a synthesis method based on a non-stationary AM-FM type representation of speech which is very close to the sinusoidal representation [10] was used.

$$s(t) = \sum_{k=1}^{N} A(kf(t), t) \sin\left(\int_0^t kf(\tau) d\tau \right) \quad (1)$$

where $f(t)$ is the F0 contour, N is the number of harmonics and $A(f,t)$ is an estimate of the spectral magnitude as a function frequency and time.

The representation of speech as a sum of harmonic sinusoids as given in equation (1) is directly dependent on the pitch contour, and allows for synthesis with both the estimated contour and its linear approximation. The spectrogram of the speech signal was used to determine the amplitude of the sinusoids for all the synthesis reported in this paper. However, other estimates such as the LPC spectrum or more complex forms reported in [11] may also be used.

4. Subjective Evaluation

The LDC Emotional Prosody speech corpus [12] was used in all investigations reported in this paper. It consists of speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers. The entire database consists of 7 actors expressing 15 emotions for around 10 utterances each. From this, data from five emotions, namely anger, sadness, happiness, boredom and neutral (no emotion) were used in all experiments. The speech data was sampled at 22.05 kHz and used at the same rate in all experiments reported in this paper. Two listening tests were performed to determine if the linear approximations to pitch contour segments retained sufficient information about the emotion being expressed by speech. Both tests were taken by the same eleven untrained listeners. Synthetic speech used in both listening tests was produced from spectrograms estimated from speech samples taken from the LDC database along with either the actual pitch contours estimated from these samples or linear approximations of the estimated contours.

4.1. Accuracy of linear approximation

The first test compared speech re-synthesised using the linear approximations with speech re-synthesised using pitch contours estimated from the original samples. The eleven untrained listeners were given two utterances, which they could listen to as many times as they needed to, and asked to give a non-fractional score between 1 and 5 depending on how close the two utterances were to each other. The scores were described as follows.

- 5 – Utterances are indistinguishable
- 4 – Utterance sound very similar
- 3 – Utterances sound moderately similar
- 2 – Utterances have very little similarity
- 1 – Utterances are completely dissimilar

The listeners were also asked to consider only how close the two utterances were to each other and to not take into account any other factors such as intelligibility, quality, clarity of emotional expression, etc. Each listener rated 30 comparisons, of which 15 were control where both utterances were identical (both were speech re-synthesised using the estimated pitch contour). For the other 15 comparisons one utterance was speech re-synthesised using linear approximations to pitch contours and the other utterance was speech re-synthesised using estimated pitch contours. Re-synthesised speech using estimated pitch contours was used instead of actual speech so as to negate the effect of some quality loss due to the re-synthesis method, which is independent of approximations to the pitch contour. The utterances for the 15 control and 15 comparisons were chosen to produce 3 samples of the five emotions in each set but were otherwise selected randomly from the database. The scores given by each listener were normalised using the mean control score of that listener as given below.

$$\hat{S}_i = \frac{S_i \times 5}{C_i} \quad (2)$$

where, $\hat{S}_i$ is the adjusted score for the i^{th} listener, S_i is the actual score and C_i is the mean control score

The mean comparison scores for each of the five emotions and the overall mean comparison score are listed in Table 1. The high scores across all five emotions indicate that the use of linear approximations is more or less indistinguishable from the use of the estimated pitch contours.

Table 1. *Subjective comparison scores (Range 1-5, with 5 indicating two versions were indistinguishable)*

Emotion	Mean Score
Neutral	4.67
Anger	4.32
Sadness	4.84
Happiness	4.30
Boredom	4.87
Overall	**4.60**

4.2. Emotion classification - Human

In the second test, listeners were given a sample of speech, which they could listen to as many times as necessary, and asked to classify it as one of the five emotions (Neutral, Anger, Sadness, Happiness and Boredom). Each listener classified 45 utterances, comprising three versions each of three samples drawn from each of the five emotions. The first version was the actual speech sample from the database, the second version was speech re-synthesised using the estimated pitch contour and the third version was speech re-synthesised using linear approximations to the pitch contours. The 45 utterances were presented in random order to the listeners. The confusion matrices for the three versions are given in Tables 2-4.

Table 2. *Confusion matrix for original speech*

	Neutral	Anger	Sad	Happy	Bored
Neutral	**69.7 %**	3 %	9.1 %	0 %	18.2 %
Anger	3 %	**93.9 %**	0 %	3 %	0 %
Sad	12.1 %	3 %	**57.6 %**	0 %	27.3 %
Happy	39.4 %	3 %	9.1 %	**45.5 %**	3 %
Bored	33.3 %	0 %	15.2 %	0 %	**51.5 %**

Table 3. *Confusion matrix for re-synthesised speech using actual estimated pitch contour*

	Neutral	Anger	Sad	Happy	Bored
Neutral	**78.8 %**	3%	9.1 %	0 %	9.1 %
Anger	3 %	**78.8 %**	0 %	18.2 %	0 %
Sad	18.2 %	6.1 %	**51.5 %**	0 %	24.2 %
Happy	39.4 %	0 %	12.1 %	**39.4 %**	9.1 %
Bored	18.2 %	0 %	24.2 %	6.1 %	**51.5 %**

Table 4. *Confusion matrix for re-synthesised speech using linear approximations of pitch contours*

	Neutral	Anger	Sad	Happy	Bored
Neutral	**60.6 %**	3 %	18.2 %	3 %	15.2 %
Anger	9.1 %	**72.7 %**	0 %	18.2 %	0 %
Sad	24.2 %	6.1 %	**39.4 %**	0 %	30.3 %
Happy	39.4 %	9.1 %	12.1 %	**30.3 %**	9.1 %
Bored	21.2 %	0 %	15.2 %	0 %	**63.6 %**

From these confusion matrices, it can be seen that the class confusion patterns across the five emotions are more or less consistent for all three versions. However, anger is not identified as well in both re-synthesised versions as it is in the actual speech sample, even though it is still the most accurately recognised emotion in all three cases. The most likely reason for this drop in accuracy is that voice quality factors are not preserved very well by the re-synthesis method adopted in this investigation. There is a drop in accuracy for sadness as well, but it is not as significant as the drop for anger. Happiness is not very well recognised even in the first case, making it hard to infer anything from the results for it. The recognition rates for boredom and neutral are more or less consistent for all three cases. The recognition and confusion rates in the second and third cases are similar, indicating that the linear approximations to the pitch contours are able to capture a significant amount of the information that the pitch contours contain about the emotion being expressed. To summarise, (i) the loss of voice quality as a result of the synthesis method led to a drop in recognition rates; and (ii) similar recognition rates in the second and third case indicate that the linear approximations are able to preserve emotion-specific information in pitch contours to a large extent.

5. Automatic Classification System

5.1. The Front – End

Apart from the subjective listening tests, an automatic emotion classification system based on the linear approximations to the pitch contour was also constructed to help determine if they were able to capture emotion-specific information. As shown in figure 2, linear approximations to segments of the pitch contour of each utterance were determined. Each linear segment was represented by a three-dimensional vector comprising the slope of the linear fit (*s*), the initial offset (*b*) and the length of the segment (*x*) (refer figure 1). Thus the entire utterance was represented by a sequence of *N* 3-dimensional vectors, and served as the front-end for the classification system.

5.2. The Back – End

Since the front-end produces a sequence of vectors for every utterance, the system requires a back-end that can model such sequences, and therefore a hidden Markov model (HMM) based back-end was chosen. The number of states in the HMMs is determined by how much of the variations in the contours between segments must be modelled and by how many segments were present in the utterances. A 2-state HMM has sufficiently many states to model the variations in the initial and terminal sections of the contour without over-fitting and losing the ability to generalise. Preliminary experiments supported this choice. Each state was represented by a 4-mixture Gaussian mixture model. Previously, we used a modified feature warping technique as a means of speaker normalisation [13]. We apply it to the features in all experiments reported in this paper.

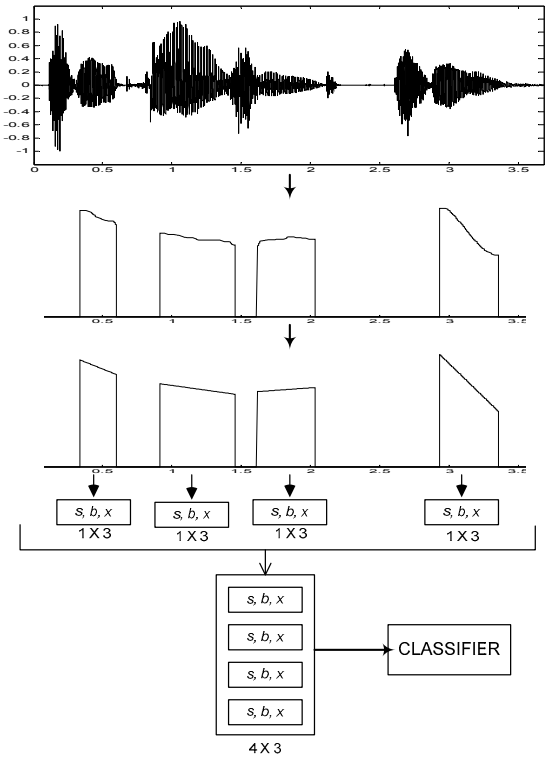

Figure 2: *An overview of the system*

5.3. Experimental Results

The automatic classification system was setup for a five-emotion (Neutral, Anger, Sadness, Happiness and Boredom) classification problem, implemented in a speaker-independent configuration. All experiments were repeated 7 times in a 'leave-one-out' manner, using data from each of the 7 speakers as the test set in turn, and the data from the other 6 speakers as the training set. The accuracies reported are the means of the seven trials.

Table 5. *Confusion matrix for the HMM based automatic emotion classification system*

	Neutral	Anger	Sad	Happy	Bored
Neutral	**51.1 %**	0.0 %	25.5 %	8.5 %	14.9 %
Anger	0.0 %	**79.2 %**	0.0 %	18.1 %	2.8 %
Sad	19.7 %	1.6 %	**47.5 %**	11.5 %	19.7 %
Happy	5.5 %	27.4 %	15.1 %	**50.7 %**	1.4 %
Bored	15.6 %	0.0 %	22.1 %	11.7 %	**50.6 %**

Comparing the accuracies reported in Table 5 to those reported in Table 2, it can be seen that the recognition rates for anger, sadness and boredom are similar, while the automatic classifier is not as good as humans in recognising neutral speech but is better than humans at recognising happiness. When comparing confusion rates, it can be seen that automatic classification and human classification are very different from each other, suggesting that the information contained in the pitch contours are used in different ways. While this is interesting, it suggests that direct comparisons of the two sets of accuracies must be done with a lot of care.

In order to determine the value of modelling the pitch contour rather than just the statistical distribution of the pitch values (without taking into consideration any temporal dependence), a Gaussian mixture model (GMM) based classification system that used pitch values as features was set up, and the classification accuracy of that system was compared with that of the system described above (Figure 2). Each utterance was classified by maximum-likelihood estimation, with the log-likelihoods of the utterance computed as the sum of the log-likelihoods of all the voiced frames in that utterance.

In the GMM-based system, the probability density functions of pitch values for each emotion were modelled by a Gaussian mixture model, capturing all the statistical information present in these values but no temporal information that may be contained in the shape of the pitch contours. On the other hand, the HMM-based system that models the feature sequences based on linear approximations to the pitch contour would model this temporal information along with the pitch values. Thus, the contribution of this temporal information towards recognising emotions can be determined by comparing the recognition rates of these two systems (Tables 5-6).

Table 6. *Confusion matrix for the GMM based automatic emotion classification system (No temporal pattern)*

	Neutral	Anger	Sad	Happy	Bored
Neutral	**47.2 %**	0.0 %	24.5 %	3.8 %	24.5 %
Anger	0.0 %	**75.7 %**	1.4 %	23.0 %	0.0 %
Sad	31.1 %	1.4 %	**50.0 %**	12.2 %	5.4 %
Happy	0.0 %	29.4 %	18.8 %	**47.1 %**	4.7 %
Bored	47.8 %	2.2 %	21.7 %	8.7 %	**19.6 %**

The comparison shows that the performance of the HMM-based system was much better than that of the GMM-based system. The overall accuracy of the HMM-based system was 56.4% as opposed to 46.6 % for the GMM-based one, lending support to the claim that temporal information contained in the shape of the pitch contour is useful in recognising emotions and that this information is preserved by the linear approximations to the pitch contours to a large extent.

Table 7. *Summary of Overall Accuracies*

Classification Test	Accuracy
Human – Actual Speech	63.6 %
Human – Re-synthesised with estimated F0	60.0 %
Human – Re-synthesised with approximate F0	53.3 %
Automatic – using temporal information	56.4 %
Automatic – without any temporal information	46.6 %

6. Conclusions

This paper has shown that the shape of pitch contours containing information about temporal variations in the pitch is useful in emotion recognition. The pitch contour is intrinsically segmented by voiced regions separated from each other by unvoiced sections where pitch does not exist. Linear approximations to the pitch contour in each segment are able to retain a lot of the information contained in the actual contours themselves, as shown by the listening tests. The linear approximations also allow for the pitch contours to be represented compactly with each segment being characterised by three parameters. This allows for these linear parameters to be used as features in an automatic classification framework. A comparison of such a system to another automatic classification system that is also based on pitch values but does not take into account the shape of the contour reveals that the additional information contained in the shape is significant and is retained by the linear approximations.

7. References

[1] Cahn, J. E., "The generation of affect in synthesized speech", in *Journal of the American Voice I/O Society,* 8:1-19, 1990

[2] Boula de Mareuil, P., Celerier, P., and Toen, J., "Generation of emotions by a morphing technique in English, French and Spanish", in *Proc. of Speech Prosody 2002,* pp. 187-190, 2002

[3] Murray, I. R., and Arnott, J. L., "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", in *Speech Communication,* vol. 16, issue 4, pp. 369-390, 1995

[4] Mozziconacci, S. L. J., and Hermes, D. J., "Role of intonation patterns in conveying emotion in speech", in *Proc. of the 14th International Conference of Phonetic Sciences,* pp. 2001-2004, 1999.

[5] Burkhardt, F., and Sandlmeier, W. F., "Verification of acoustical correlates of emotional speech using formant-synthesis", in *Proc. of ISCA Workshop on Speech and Emotion,* pp. 151-156, 2000.

[6] Mertens, P., and d'Alessandro, C., "Pitch contour stylization using a tonal perception model', in *Proc. Int. Congr. Phonetic Sciences* 13, 4, pp. 228-231, 1995

[7] Wang, D., and Narayanan, S., "Piecewise linear stylization of pitch via wavelet analysis", in *Proc. of INTERSPEECH-05,* pp. 3277-3280, 2005

[8] Ravuri, S., and Ellis, D. P. W., "Stylization of pitch with syllable-based linear segments", in *Proc. of ICASSP-08,* pp. 3985-2988, 2008

[9] Talkin, D., "A robust algorithm for pitch tracking (RAPT)" in Speech Coding and Synthesis, Elsevier, W. B. Kleijin, and K. K. Paliwal [eds], pp. 495-518, 1995.

[10] McAulay, R., and Quatieri, T., "Speech analysis/synthesis based on a sinusoidal representation", in *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 34, issue 4, pp. 744-754, 1986.

[11] Kawahara, H., Masuda-Katsuse, I., and Cheveigne, A. de, "Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of repetitive structure in sounds", in *Speech Communication,* vol. 27, issue 3-4, pp.187-207, 1999

[12] Emotional Prosody Speech corpus, Linguistic Data Consortium, University of Pennsylvania, PA, USA, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

[13] Sethu, V., Ambikairajah, E., and Epps, J., "Speaker normalisation for speech based emotion detection," in *Proc. 15th Int. Conf. Digital Signal Processing*, pp. 611-614, 2007.