EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

On the use of speech parameter contours for emotion recognition

Vidhyasaharan Sethu*, Eliathamby Ambikairajah and Julien Epps

Abstract

Many features have been proposed for speech-based emotion recognition, and a majority of them are frame based or statistics estimated from frame-based features. Temporal information is typically modelled on a per utterance basis, with either functionals of frame-based features or a suitable back-end. This paper investigates an approach that combines both, with the use of temporal contours of parameters extracted from a three-component model of speech production as features in an automatic emotion recognition system using a hidden Markov model (HMM)-based back-end. Consequently, the proposed system models information on a segment-by-segment scale is larger than a frame-based scale but smaller than utterance level modelling. Specifically, linear approximations to temporal contours of formant frequencies, glottal parameters and pitch are used to model short-term temporal information over individual segments of voiced speech. This is followed by the use of HMMs to model longer-term temporal information contained in sequences of voiced segments. Listening tests were conducted to validate the use of linear approximations in this context. Automatic emotion classification experiments were carried out on the Linguistic Data Consortium emotional prosody speech and transcripts corpus and the FAU Aibo corpus to validate the proposed approach.

Keywords: Emotion recognition; Paralinguistic information; Pitch contours; Formant contours; Glottal spectrum; Temporal information; LDC emotional prosody speech corpus

1. Introduction

Human speech is an acoustic waveform generated by the vocal apparatus, whose parameters are modulated by the speaker to convey information. The physical characteristics and the mental state of the speaker also determine how these parameters are affected and, consequently, how speech conveys the intended, and on occasion unintended, information. Even though knowledge about how these parameters characterise the information is not explicitly available, the human brain is able to decipher this information from the resulting speech signal, including the emotional state of the speaker.

Information about emotional state is expressed via speech through numerous cues, ranging from low-level acoustic ones to high-level linguistic content; several approaches to speech-based automatic emotion recognition, each taking advantage of a few of these cues, have been explored [1-9]. It would be impossible to list all of them; however, approaches that use linguistic cues [10,11] are not as common as those that make use of low-level acoustic and prosodic cues. The most commonly used acoustic and prosodic features tend to be those based on cepstral coefficients, pitch, intensity and speech rate.

The standard speech production model (source-filter model) [12], widely used in speech processing literature, is the model that underpins most low-level feature extraction algorithms. However, almost universally, a simplifying assumption is made about the shape of the glottal pulses. Specifically, the glottal model is assumed to be a two-pole low-pass system (typically with both poles at unity) whose effects are 'removed' at the pre-emphasis stage of feature extraction. Section 2 explores the estimation of glottal parameters without making such an assumption.

The standard speech production model is also a short-term spectral model, and almost all low-level features tend to be short-term frame-based features incapable of capturing long-term temporal information. This shortcoming is widely recognised, and a range of approaches have been developed to overcome it, including the use of delta features (and its variants) to capture frame-to-frame temporal variations, the use of hidden Markov models to model temporal

* Correspondence: v.sethu@unsw.edu.au
The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, New South Wales 2052, Australia

patterns, the use of neural networks with memory and the use of functionals of sequence of frame-based parameters estimated over long segments (turns) as features. Automatic emotion recognition systems commonly address this issue in one of two ways: either by using a suitable back-end to model temporal variations of short-term features [13,14], or by capturing this information with the front-end with techniques such as contour models or using functionals of sequences of short-term features extracted from speech segments/turns [7,9,15-20]. In this paper we explore a combination of both approaches. Specifically, temporal contour patterns of pitch, the first three formant frequencies and the gains of the vocal tract at these frequencies, along with those of the glottal parameters outlined in Section 2, are used to model temporal information roughly spanning durations of voice segment, followed by a back-end based on hidden Markov models (HMMs) to model the sequences of these temporal patterns. Short-term temporal information is captured by the front-end, and longer-term temporal information is modelled by the back-end. Section 5.2 reports the results of a listening test carried out to validate the use of linear approximations of glottal parameter contours, conducted in a similar manner to a previous listening test carried out to validate the use of linear approximations to pitch contours [21]. Finally, the proposed approach has the inherent advantage that the segmentation of the speech into turns is not required since the system carries out an implicit segment-by-segment modelling of voiced speech segments.

2. The glottal source parameters

2.1. Glottal flow models

In the well-established and commonly used speech production model [12], the glottal flow that serves as the input to the vocal tract is modelled as the response of a filter. The shape of this response (glottal pulse shape) has been associated with certain characteristics of speech that are subsumed under the cover term *voice quality* [22]. The importance of appropriate glottal models in the synthesis of natural sounding speech has been well established [22,23], and the modification of glottal voice quality factors has been shown to be significant for the synthesis of emotional (expressive) speech [24].

Approximating the glottal filter model by $G(z) = 1/(1 - az^{-1})^2$, as is commonly done and is equivalent to making the assumption that the shape of the glottal pulse is always the same, may not be the best approach to take in emotion recognition systems. The incorporation of a more detailed glottal source model may be desirable and while not common, glottal parameters have been used in an emotion classification framework and were shown to be useful in distinguishing between emotional category pairs that had statistically similar pitch values [25-27].

The most popular glottal models [28-31] are time domain models that vary in the specific parameters they incorporate but share a common framework, and the spectra of the glottal flow derivatives described by all of them can be stylised by three asymptotic lines with +6 dB/oct, −6 dB/oct and −12 dB/oct slopes [32] as shown in Figure 1. Such a stylised representation allows for a very compact characterisation of the glottal flow derivative magnitude spectrum since it can be uniquely identified based on three values, specifically two corner frequencies (F_g and F_c) and the magnitude at the first corner frequency (A_g). The compact representation also lends itself to use as a feature in a classification system.

2.2. Estimation of glottal spectral parameters

While it is obvious that the use of a glottal flow (or glottal flow derivative) model results in a more accurate modelling of the speech production system when compared with the fixed two-pole approximation that is commonly utilised, the estimation of the glottal flow signal from the speech signal is not a well-defined problem and lacks an analytical solution. However, numerous techniques have been proposed over the years that are based on the properties of the glottal flow signal [33-38]. The iterative adaptive inverse filtering (IAIF) method [33] was used to estimate the glottal flow derivative in the work reported in this paper. The IAIF method can be used pitch synchronously (with variable window lengths based on pitch) or asynchronously (with fixed windows). For pitch synchronous IAIF, the DYPSA algorithm [39] has been used to detect glottal closure instants and hence identify window boundaries.

Given an estimate of the glottal flow derivative, it is proposed that the best stylised fit to its magnitude spectral envelope (Figure 1), in terms of minimum mean squared error, can be estimated via brute force search of the three-dimensional parameter space. The choice of a brute force search was made purely due to the simplicity of the search algorithm even though it is not efficient. This approach was considered acceptable since the glottal parameters themselves are the focus of the work reported in this paper and not their estimation algorithms. For any given parameter set $\{F_g, A_g, F_c\}$, the stylised glottal flow derivative magnitude spectrum, $\tilde{G}(\cdot)$, was constructed and the mean squared error between the stylised and the estimated spectra, $\hat{G}(\cdot)$, was calculated as

$$\tilde{G}_{F_g,A_g,F_c}(\zeta) = \begin{cases} A_g\text{-}20(\zeta_g\text{-}\zeta), & \zeta<\zeta_g \\ A_g\text{-}20(\zeta\text{-}\zeta_g), & \zeta_g<\zeta<\zeta_c \\ A_g\text{-}20(\zeta_g\text{-}\zeta)\text{-}40(\zeta\text{-}\zeta_c), & \zeta_c<\zeta<\zeta_m \end{cases} \quad (1)$$

where $\zeta = \log_{10}(f)$, $\zeta_g = \log_{10}(F_g)$, $\zeta_c = \log_{10}(F_c)$; A_g is the magnitude at the corner frequency in dB; $\zeta_m = \log_{10}(F_s/2)$

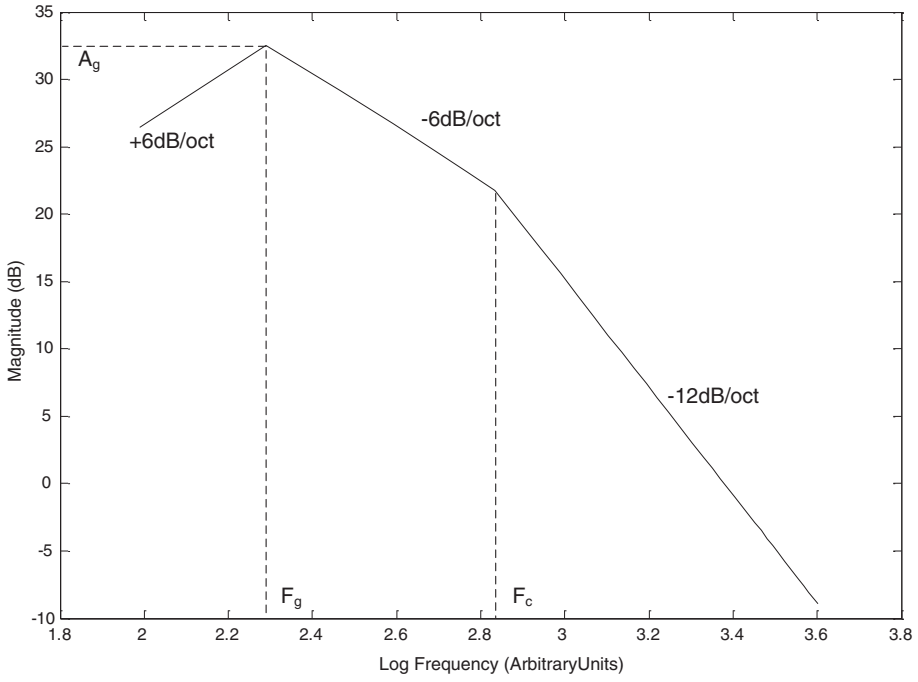

Figure 1 Stylised glottal flow derivative magnitude spectrum, after the work of Doval et al. [32].

and F_s is the sampling rate. In the experiments reported in this paper, all computations were carried out on discrete values of f, corresponding to the discrete Fourier transform coefficients computed from each frame of speech.

This mean squared error was computed for all possible combinations of F_g, A_g and F_c with the search space spanning all possible values of the three parameters with a resolution of 30 values for each parameter. The parameter set, $\{\hat{Fg}, \hat{Ag}, \hat{Fc}\}$, that gave the lowest mean squared error was then selected as the best fit:

$$\{\hat{Fg}, \hat{Ag}, \hat{Fc}\} = \arg \min_{Fg,Ag,Fc} \left\| \hat{G}(\zeta) - \tilde{G}_{Fg,Ag,Fc}(\zeta) \right\|_2 \quad (2)$$

where $\|\cdot\|_2$ denotes the l_2 norm.

An overview of the glottal spectral parameter estimation method is given in Figure 2.

Visual inspection of the glottal flow spectra for a few consecutive frames and their corresponding stylised spectra indicated that even though the glottal flow spectra for the frames were not identical and had estimation errors at different points in different frames, the stylised spectra were all similar (particularly with regards to the glottal formant). This indicates that the spectrum fitting process is robust, to a certain extent, to errors in the glottal flow derivative estimation process that result from incomplete removal of the formant structure, particularly in terms of identifying the glottal formant. However, the estimation of the corner frequency, F_c, is affected to a much larger degree by the errors and the estimated values were not very reliable. Therefore, F_c was ignored and only the glottal formant frequency and magnitudes, F_g and A_g, were used in the automatic classification results reported in this paper. These frequency domain parameters are related to the more commonly utilised time domain parameters such as the open quotient and speed quotient [32]. However, the frequency domain parameters can be obtained by fitting the linearly stylised spectrum as outlined, which is conceptually simpler than the time domain curve fitting methods required to estimate the time domain parameters.

3. Parameter contours

As previously mentioned, the study aims to capture short-term temporal information in the front-end prior to modelling longer-term information with the back-end . In this regard, the speech model parameter contours are estimated in the front-end. Parameter contours are representative of these variations over an entire utterance and are characterised by a much longer duration when compared with descriptions provided by deltas and shifted deltas. The most common and probably best studied is the pitch (F_0) contour, which is essentially pitch as a function of time, and its use in an automatic emotion recognition (AER) framework was the focus of a previous study [21]. Linear stylisation of F_0 contours [40,41] is commonly carried out, for a range of applications from speech coding to modelling dialogue acts and

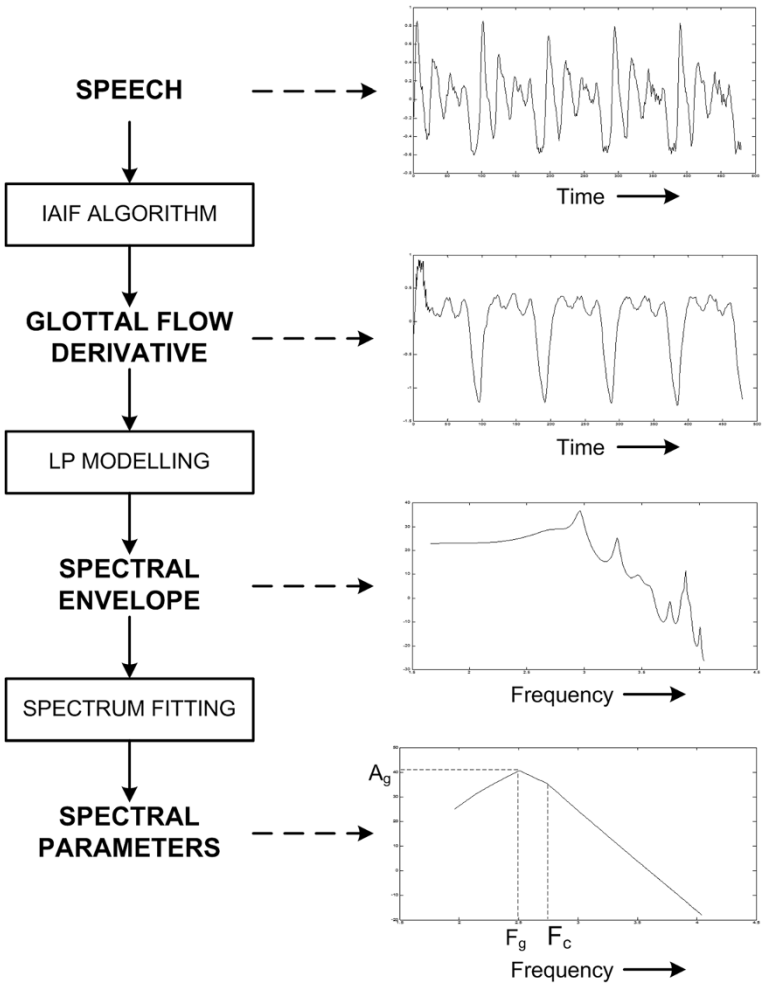

Figure 2 Overview of glottal spectral parameter estimation procedure.

speaker verification, to make them simpler to analyse, and also has the additional advantage of making their representation more compact than that of the original contour. The idea of such stylisation can be extended to the other parameters, such as the glottal formant and magnitude and formant frequencies and magnitude as well. Here we propose approximating the parameter contours in each voiced segment by a straight line that enables the representation of that contour using three parameters, namely, the slope of the line (s), the initial offset (b) and the length of the segment (τ), as shown in Figure 3. This is different from typical F_0 contour stylisation [40] since contours in each voiced segment are represented by single line segments rather than piecewise linear approximations. This is done to ensure that a single three-dimensional vector can describe an entire contour for each voiced segment. This is the simplest form of contour parameterisation, and if an AER system that makes use of such a representation outperforms a system that models the distribution of the parameter values (without taking into account any temporal information), then it is reasonable to suppose that the shape of the parameter contours contain emotion-specific information.

A small variation to the representation of the linear approximation to the contours involves the use of the value of the midpoint (m) of the contour instead of the initial values. This serves two purposes, namely: (1) In case of errors in the estimation of parameter values, particularly near the beginning of the voiced segment, the magnitude of the error in the value of the midpoint will be smaller than the error in the initial value; (2) If there are multiple errors in parameter estimation, the value of the slope can be dropped and only the value of the midpoint (which will be an estimate of the mean parameter value) can be used for a more robust but less detailed representation of the contour.

In general, any parameter that varies with time within an utterance can be linearly stylised for compact representation and ease of analysis. This paper focuses on the

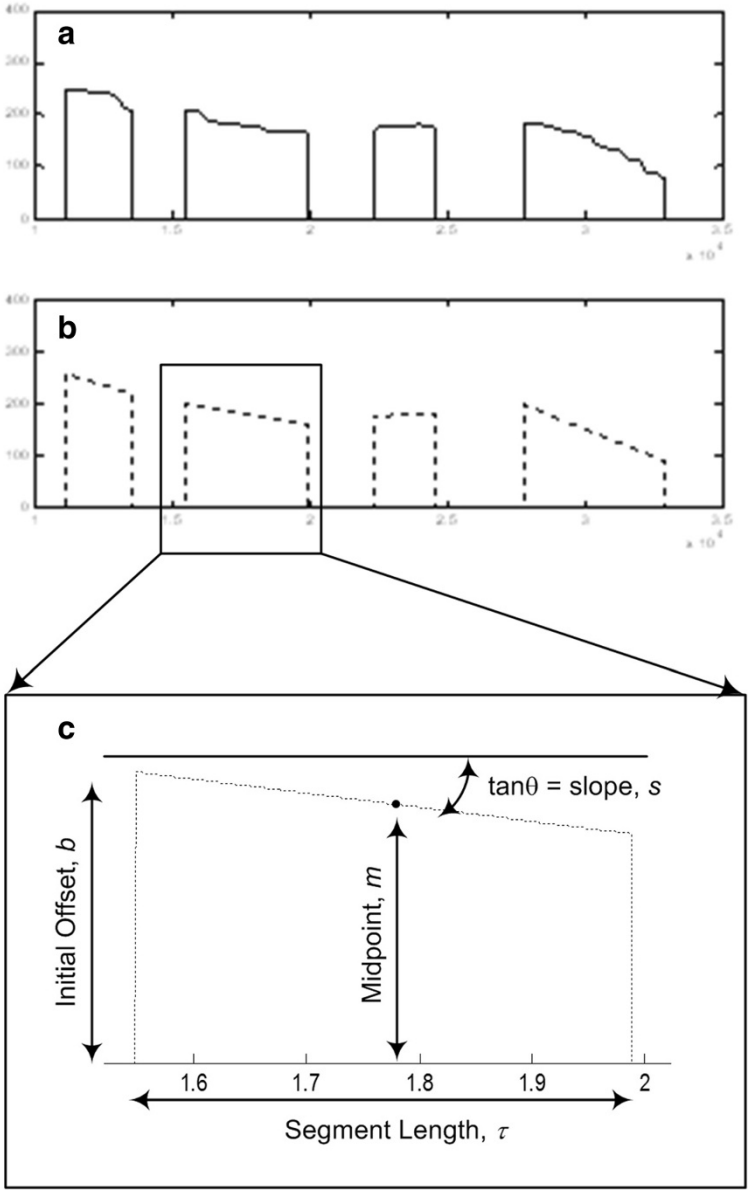

Figure 3 Approximating parameter contours. (a) Estimated F_0 contour. **(b)** Linear approximation of F_0 contour. **(c)** Linear model parameters - τ, s and b (or m).

use of the source-filter model of speech production, and consequently the contours considered are those of the parameters of this model. Specifically, the vocal tract is parameterised by the contours of the first three formant frequencies (F_1, F_2 and F_3 contours) and the contours of the magnitudes of the all-pole vocal tract spectrum at these frequencies (A_1, A_2 and A_3 contours), the glottal flow derivative (considering glottal flow and lip radiation together) by the glottal formant contours (F_g and A_g contours) and the source excitation rate using F_0 contours. Hence, an utterance can be compactly represented as a sequence of vectors, $\mathbf{V}$:

$$\mathbf{V} = \left\{ \boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, ..., \boldsymbol{v}^{(N)} \right\} \tag{3}$$

where N is the number of voiced segments in the utterance, and $\boldsymbol{v}^{(i)}$ is a vector corresponding to the ith voiced segment,

$$\boldsymbol{v}^{(i)} = \begin{bmatrix} s_{F_1}^{(i)} \\ b_{F_1}^{(i)} \\ s_{F_2}^{(i)} \\ b_{F_2}^{(i)} \\ M \\ \tau^{(i)} \end{bmatrix}, \tag{4}$$

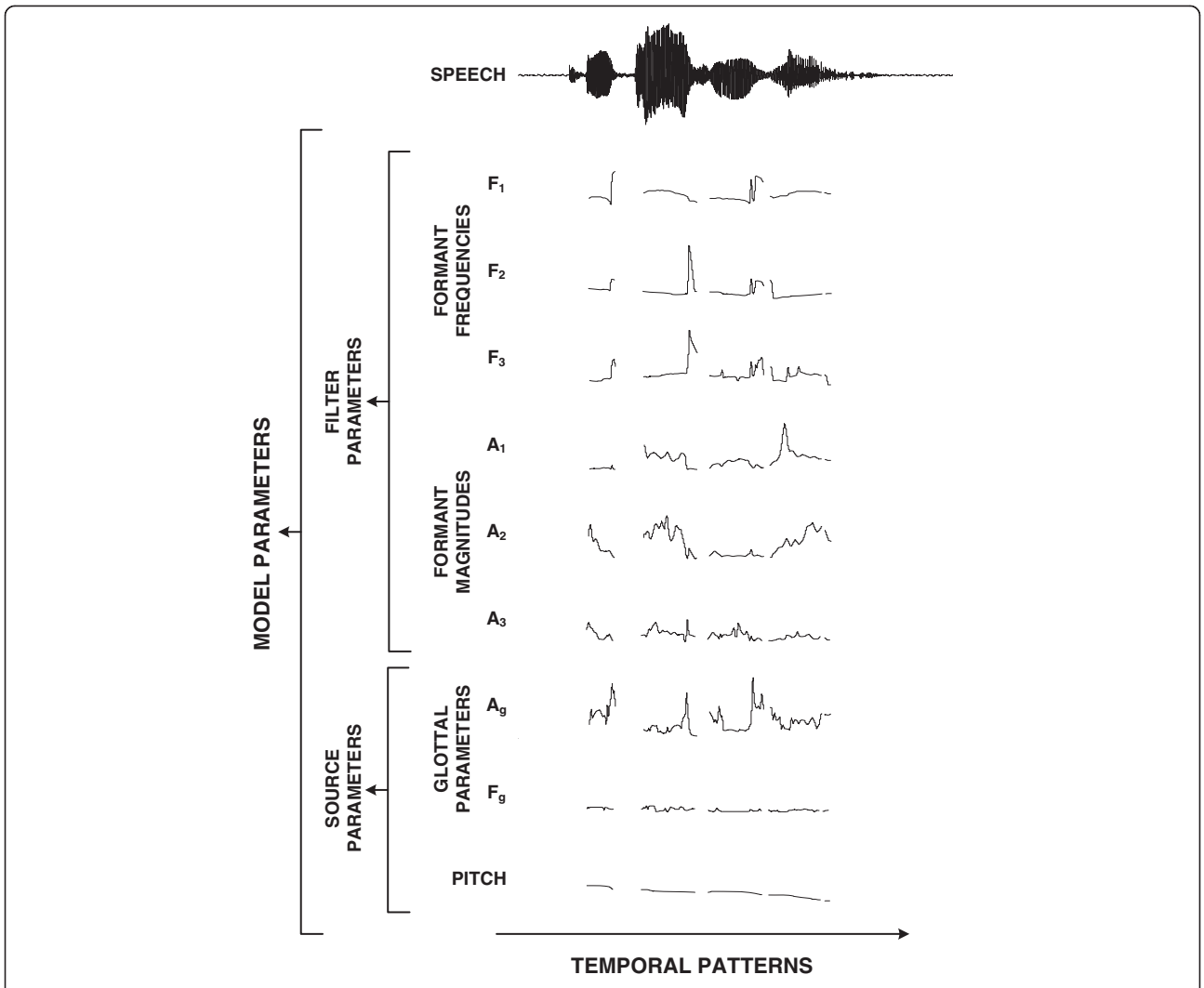

Figure 4 Parameter contour examples for the phrase 'two thousand one'. Elements of the source-filter model-based speech characterisation including glottal parameters.

where $\tau^{(i)}$ is the length of the ith voiced segment; $s_P^{(i)}$ and $b_P^{(i)}$ are the slope and initial offset (as previously mentioned, the midpoint, $m_P^{(i)}$, may be used in place of $b_P^{(i)}$in $\boldsymbol{\nu}^{(i)}$) that describe the contour of parameter P in the ith voiced segment; and P is one or more of the source-filter model parameters, i.e.,

$$P \in \{F_1, F_2, F_3, A_1, A_2, A_3, F_g, A_g, F_0\} \tag{5}$$

Note that the lengths of the contours are identical for all model parameters within each voiced segment. Hence, only one element, $\tau^{(i)}$, in each vector is required to represent the contour length. Thus, $\boldsymbol{\nu}^{(i)}$ is a $2K + 1$ dimensional vector, where K is the number of parameters chosen (from P) to represent each voiced segment, when slope and initial offset (or midpoint) are both used to represent a contour. If the slope values of some or all parameters are dropped, the vector has a lower dimension. Figure 4 depicts all the parameter contours considered in this work.

4. Emotional speech corpora

The experiments reported in this paper were performed using one of two databases, namely, the Emotional Prosody Speech and Transcripts corpus (herein referred to as the LDC corpus) [42] and the German FAU Aibo corpus [43]. The two databases have significant differences, with the LDC corpus containing 'acted' emotional speech collected from seven professional actors with the speech comprising preselected, semantically neutral phrases. The Aibo corpus on the other hand contains 'elicited' emotional speech from 51 children with no restrictions on what is being said. Moreover, the emotional categories present in both corpora are different, with the

LDC corpus emphasising high-intensity 'prototypical' emotions and the Aibo corpus focussing on low-intensity 'non-prototypical' emotions. In addition, the LDC corpus contains English speech while the Aibo corpus contains German speech. A primary motivation in the choice of corpora was the significant differences between the two corpora which suggest that trends common to the experimental results obtained from both corpora are much more likely to be independent of the test data and hence more likely to be applicable to any other emotional data as well.

4.1. LDC emotional speech and transcripts corpus

The Emotional Prosody Speech and Transcripts corpus contain audio recordings, recorded at a sampling rate of 22,050 Hz, and the corresponding transcripts (word level transcripts that lack time stamps). The recordings were made by professional actors reading a series of semantically neutral utterances consisting of dates and numbers, spanning 14 distinct emotion categories, selected based on a German study [44], and a 'neutral' category that does not involve any emotional state. Of these fifteen categories, five were chosen to conduct the five-class emotion classification experiments reported in this paper. These five emotional classes are neutral, anger, happiness, sadness and boredom.

Four female and three male actors participated in the creation of this database and were provided with descriptions of each emotional context, including situational examples adapted from those used in the German study. Flashcards were used to display series of four syllable dates and numbers to be uttered in the appropriate emotional category. During the recording, the actors repeated each phrase as many times as necessary until they were satisfied that the emotion was expressed and then moved onto the next phrase. Only the last instance of each phrase was included in all the experiments reported herein. This provided about 8 to 12 utterances per speaker for each of the five emotional categories. While the phrases recorded for all emotions were not identical, they were very similar to each other and contained numerous words that were common (e.g. 'two thousand and one' and 'two thousand and twelve' or 'December second' and 'December twenty-first').

4.2. FAU Aibo emotion corpus

The German FAU Aibo Emotion Corpus [43] consists of spontaneous emotionally coloured children's speech with recordings of 51 German children (30 females and 21 males) aged between 10 and 13 from two different schools. The speech signals were recorded at a sampling rate of 48 kHz with 16-bit quantisation and then down-sampled to a rate of 16 kHz. The children were given different tasks where they had to direct Sony's dog-like robot Aibo to certain objects and along given 'parcours'. They were told that they should talk to Aibo like a real dog and in particular reprimand it or praise it as appropriate. However, Aibo was remote-controlled from behind the scenes and at certain positions it was made to disobey or act in some manner not instructed by the child 'controlling' it in order to elicit an emotionally coloured response from the child. It is important to note that the real purpose of the data collection experiment, the elicitation of the emotions, was not known to the children, and none of them realised that Aibo was remote-controlled.

The recorded speech data was then annotated independently at the word level by five listeners using 11 emotional categories. The chosen units of analyses were, however, not single words but short phrases referred to as 'chunks' which were semantically and syntactically meaningful. The emotional categories assigned to each word in the chunk by all five listeners were combined heuristically to generate a single emotion label for the chunk. All classification experiments reported in this paper based on this database were setup (with regards to the training and test data sets and the performance measure) along the lines of the five-class problem set out in the INTERSPEECH 2009 Emotion Challenge [45] that made use of this database.

5. Validating linear approximations - listening test

Listening tests were conducted to determine whether linear approximations to pitch contour segments and glottal parameter contour segments retained sufficient information about the emotions being expressed. Specifically, in addition to previous results for pitch contours (F_0) which suggested that a significant amount of information was retained [21], a further listening test was conducted to determine whether linear approximations to glottal model parameter contours (F_g, A_g and F_c) were able to sufficient to retain emotion-specific information. Even though F_c contours were not used as features in the automatic classification systems used in the work reported in this paper, they cannot be ignored for speech re-synthesis since the absence of the −12 dB/oct slope would be equivalent to high-pass filtering the speech signal. With regards to linear approximations of formant contours (F_{1-3} and A_{1-3}) while they capture broad trends that may be useful in a classification scenario, if they were used in speech synthesis, the approximation errors will give rise to significant distortion and the synthesised speech would not be useful for listening tests. More complex models of formant contours will result in less error and consequently less distortion at the cost of a larger number of parameters. The use of more complex models may be investigated in the future.

5.1. Speech re-synthesis

In the work reported herein, the purpose of the speech re-synthesis procedure is to determine temporal contours of the speech production model parameters and then to re-synthesise speech from these parameter contours or their linear approximations. A synthesis method based on a non-stationary AM-FM type representation of speech, which is very close to the sinusoidal representation [46], was used. This method was chosen since all the estimated parameter contours, particularly the pitch contour, can be directly incorporated without any further processing:

$$s(t) = \sum_{k=1}^{N} V(kf(t), t) \cdot G(kf(t), t) \cdot \sin\left(\int_0^t kf(\tau) d\tau\right) \quad (6)$$

where $f(t)$ is the F_0 contour, N is the number of harmonics, $V(f, t)$ and $G(f, t)$ are estimates of the contributions of the vocal tract and the glottal source, respectively, towards the speech spectral magnitude (i.e. the vocal tract and glottal spectra) as a function of frequency and time.

The pitch contours were estimated using the RAPT algorithm [47], and the vocal tract and glottal spectra were estimated using the pitch synchronous IAIF algorithm [33]. The three glottal parameter contours (F_g, A_g and F_c) were estimated from the glottal spectrum as outlined in Section 2.2, and their linear approximations were obtained as described in Section 3. Based on these linear approximations, the glottal contribution to the amplitudes of the pitch harmonics, $G(f, t)$, was computed as per the stylised glottal flow derivative spectrum (Figure 1) used to estimate the glottal parameters. Thus, the parameters of the re-synthesised speech samples were identical to those of the original samples except for the glottal parameters, allowing for a subjective evaluation of the linear approximation to these parameter contours. While this speech synthesis procedure is by no means state-of-the-art, it has sufficiently high quality, as indicated by comparisons between synthesised speech (synthesised without using any linear approximations) and the original speech, which suggested that the listeners could not distinguish between the two versions [21].

It should be noted that the representation of speech as a sum of harmonic sinusoids used in this re-synthesis method holds only for voiced speech and was only applied to segments where pitch estimates were available. The unvoiced segments of the original speech samples were retained during re-synthesis. This was deemed acceptable, since the automatic emotion recognition system utilised only voiced segments, as is common with most other emotion recognition systems.

5.2. Subjective evaluation

A listening test was conducted to determine whether linear approximations to glottal parameter contours are able to retain emotion-specific information. Fourteen untrained listeners were involved in this test. For each listener, 30 speech utterances were drawn at random from the LDC corpus such that there were 6 utterances from each of the five emotions (neutral, anger, sadness, happiness and boredom). The re-synthesis method outlined in Section 5.1 was used to generate an alternative version of each of the 30 utterances using linear approximations to the glottal parameter contours. All 60 utterances were then presented to the listener in random order and for each he/she was asked to pick one out of the five emotional categories that they could associate with the utterance. Their decisions were then analysed to determine for how many of the 30 possible pairs the listener selected the same emotion for both utterances. The decisions of all 14 listeners were then combined to compute the percentage of speech samples for which listeners associated the same emotion with both the original and re-synthesised version. It should be noted that the measure of primary interest was how often listeners picked the same emotions for both versions of an utterance (one synthesised with linear approximations to parameter contours and one synthesised with actual contours) and not the actual accuracy of subjective classification since the aim of the listening tests was to validate the use of linear approximations.

The number of pairs where the listener assigned the same emotion to both versions, out of a maximum possible of 30, is given (as percentages) in Figure 5. Taken together across all 14 speakers, the same emotion was associated with both versions (re-synthesized using (6) with actual glottal parameter contours F_g, A_g and F_c and re-synthesized with linear approximations to glottal parameter contours) in 65.5% of the cases, and no individual result was poorer than 56.7%. These results suggest that the linear approximations to the glottal parameter contours are able to preserve emotion-specific information to a reasonable extent. This is in agreement with the results of a previous listening test that suggested linear approximations to pitch parameter contours preserve a significant amount of emotion-specific information [21].

In order to verify that the results reported in Figure 5 are not due to the listeners picking the same emotional label for all utterances, the distribution of how many utterances were assigned to each label was determined, and the results indicated that the distribution is not skewed severely in favour of any particular emotion. The fractions of utterances assigned to each of the five labels were 36.2% (neutral), 20% (anger), 15.7% (sadness), 10.2% (happiness) and 17.9% (boredom). It should also be noted that the

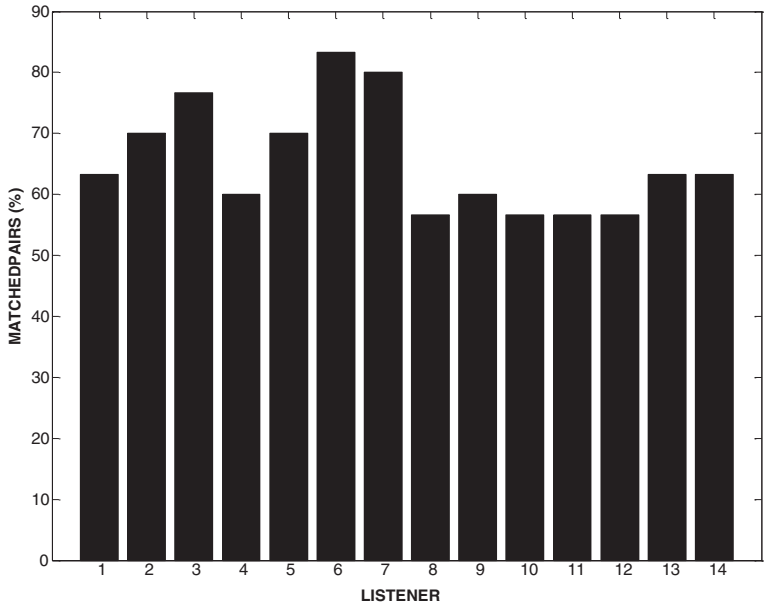

Figure 5 Percentage of pairs for which both versions were assigned the same emotion by the listener. Each listener was given 30 pairs, one version using actual contours of and one using linear approximations to F_g, A_g and F_c.

listeners were not aware that there were an equal number of samples from each of the five emotional categories.

6. Automatic classification system

6.1. The front-end

In addition to subjective evaluations, the usefulness of linear approximations to speech parameter contours to automatic classification systems was evaluated. The front-end of these systems represents the linear approximation to each parameter (P) contour in a voiced segment (i) by its slope ($s_P^{(i)}$) and initial value ($b_P^{(i)}$), as explained in Section 3. Apart from these, the length of each segment is given by a single value ($\tau^{(i)}$) and thus each voiced segment is represented by a $2K + 1$ dimensional vector, $v^{(1)}$, (where K is the number of parameters considered). Each utterance is then represented as a sequence of N such vectors, $\mathbf{V} = \left\{\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(N)}\right\}$, (where N is the number of voiced segments in the utterance) by the front-end.

6.2. The back-end

The choice of back-end is dictated by the requirement that it should be capable of modelling utterances represented by a sequence of vectors to capture longer-term temporal information, and hence hidden Markov models (HMMs) were chosen. While this in itself is not novel, in most cases when HMMs are used in an AER system, they are utilised to directly model the temporal evolution of selected features (e.g. model pitch contours and contours of cepstral coefficients), i.e. they model sequences of feature vectors, each one extracted from a frame of speech, spanning an utterance. There is less agreement, however, on the optimal number of states for these HMMs, with some studies suggesting four states [48,49] and others suggesting thirty-two or more [13,14]. However, in the system described in this paper, the parameters of the linear approximations to these contours describe the temporal evolution within each voiced segment, and the states of the HMM only describe the change *across different segments*. In a sense, the HMMs in the described system only model a second higher level of temporal variation instead of all the dynamic information. The number of states in each HMM determines how much of the variations in the contours between voiced segments in an utterance are modelled, which, in turn, depend on how many voiced segments are present in each utterance. A three-state HMM has sufficiently many states to model the variation in the initial, central and terminal segments of the utterance without overfitting and losing the ability to generalise. Preliminary experiments on the LDC corpus supported this choice. Each state utilised a 4-mixture Gaussian mixture model. For experiments performed on the FAU Aibo corpus, the number of states and mixtures were picked based on a preliminary search and the results are discussed in Section 7.2. All HMMs utilised in the systems proposed in this paper had linear left-right topology.

Previously, a modified feature warping technique was proposed as a means of speaker normalisation [50] and was applied to all features in experiments reported in this paper unless otherwise stated.

6.3. GMM-based classification system

In order to facilitate comparisons to help determine the significance of temporal information contained in parameter contours as opposed to the statistical distributions of parameter values, another classification system based on Gaussian mixture models (GMMs) was set up. The features used by this system are the frame-based values of the source-filter model parameters (F_0, A_g, F_g, A_{1-3}, F_{1-3}) obtained prior to approximating their contours with straight lines. The back-end of this system is a 128-mixture GMM-based classifier that models the probability distributions of the feature values (without taking into consideration any temporal dependence).

7. Experimental results

7.1. LDC corpus

The automatic classification systems setup for a five-emotion (neutral, anger, sadness, happiness and boredom) classification problem on the LDC corpus was implemented in a speaker-independent configuration. All experiments were repeated seven times in a 'leave-one-out' manner, using data from one of the seven speakers as the test set in turn, and data from the other six speakers as the training set. The accuracies reported are the means of the seven trials.

Classification experiments were performed using the HMM-based system using features based on pitch contour, glottal parameter contours and formant contours individually. In all three cases, two descriptions of the contours were used: one was the two-dimensional representation of each segment of the contour using the slope and initial value, $[s_P^{(i)}, b_P^{(i)}]$, (abbreviated as S-I henceforth) and the other a one-dimensional representation using only the value of the midpoint, $[m_P^{(i)}]$, (abbreviated as MP henceforth), where P can represent any parameter as given by Equation 5. In both descriptions the length of each voiced segment was appended as outlined in Section 6.1. The confusion matrices corresponding to these six experiments are given in Tables 1,2,3,4,5 and 6.

Emotion classification experiments were also performed using a GMM-based system with pitch, glottal parameter and formant parameter values as features in order to compare systems that model temporal information contained in linear contour approximations to those that modelled only the statistical distributions of the parameter values. A summary of the overall accuracies obtained from these experiments is given in Table 7.

These results indicate that the contour modelling approach is better than the static distribution modelling approach for pitch and glottal parameters. Also the pitch, glottal and vocal tract parameters describe independent (based on the source-filter model) and distinct components in the speech production mechanism and are hence independent of each other. Consequently, their contours can be expected to be complementary (unless the back-end models identical information from the different parameter contours in each voiced segment). This suggests that a system, such as the HMM-based one, can be used to model all the contours by simply concatenating the contour descriptors (either slope, initial value or midpoint value or some combination of them) to form the feature vector. It should be noted that while a static modelling approach was better than the contour modelling one for vocal tract parameters, the vocal tract parameters would still contribute towards a combined system if they are complementary to pitch and glottal parameters. Such a system was constructed and its performance was evaluated on the LDC corpus. The classification accuracies obtained are reported in Table 8. This system uses the S-I description for pitch contours and the midpoint value (MP) description for glottal and vocal parameter contours, i.e. the feature vector corresponding to the ith voiced segment is

Table 1 Confusion matrix for the HMM-based system using pitch contours (S-I), $v^{(i)} = \left[s_{F_0}^{(i)}, b_{F_0}^{(i)}\right]$ (overall accuracy 57.1%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*59.6*	0	19.2	6.4	14.9
Anger	0	*78.9*	2.8	16.9	1.4
Sad	13.1	1.6	*49.2*	14.8	21.3
Happy	1.4	32.9	16.4	*43.8*	5.5
Bored	6.5	0	28.6	10.4	*54.6*

Table 2 Confusion matrix for HMM-based system using glottal parameter contours (S-I), $v^{(i)} = \left[s_{F_g}^{(i)}, b_{F_g}^{(i)}, s_{A_g}^{(i)}, b_{A_g}^{(i)}\right]$ (overall accuracy 52.6%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*53.2*	0	14.9	6.4	25.5
Anger	1.4	*80.3*	0	18.3	0
Sad	16.4	3.3	*31.2*	6.6	42.6
Happy	9.6	19.2	6.9	*50.7*	13.7
Bored	13.0	1.3	31.2	9.1	*45.5*

Table 3 Confusion matrix for HMM-based system using formant parameter contours (S-I), $v^{(i)} = \left[s_{F_1}^{(i)}, b_{F_1}^{(i)}, s_{F_2}^{(i)}, b_{F_2}^{(i)}, s_{F_3}^{(i)}, b_{F_3}^{(i)}, s_{A_1}^{(i)}, b_{A_1}^{(i)}, s_{A_2}^{(i)}, b_{A_2}^{(i)}, s_{A_3}^{(i)}, b_{A_3}^{(i)}\right]$ (overall accuracy 41.6%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*55.3*	2.1	14.9	17.0	10.6
Anger	4.2	*46.5*	0	26.8	9.9
Sad	6.6	8.2	*32.8*	24.6	27.9
Happy	1.4	21.9	15.1	*37.0*	24.7
Bored	2.6	9.1	20.8	27.3	*40.3*

Table 4 Confusion matrix for the HMM-based system using pitch contours (MP), $\boldsymbol{v}^{(i)} = \left[m_{F_0}^{(i)}\right]$ (overall accuracy 56.2%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*63.8*	0	14.9	4.3	17.0
Anger	0	*77.5*	2.8	18.3	1.4
Sad	13.1	0	*39.3*	26.2	21.3
Happy	4.1	27.4	16.4	*49.3*	2.7
Bored	10.4	1.3	28.6	7.8	*52.0*

given by $\boldsymbol{v}^{(i)} = \left[\tau^{(i)},\ s_{F_0}^{(i)},\ b_{F_0}^{(i)},\ m_{F_g}^{(i)},\ m_{A_g}^{(i)},\ m_{F_1}^{(i)},\ m_{A_1}^{(i)},\ m_{m}^{F_2(i)},\ m_{A_2}^{(i)},\ m_{F_3}^{(i)},\ m_{A_3}^{(i)}\right]$.

Previously, a listening test was conducted on the LDC corpus and the results were reported in [21]. The overall accuracy obtained by 11 human listeners was 63.6%. A comparison of this accuracy to the different automatic emotion classification systems mentioned above is summarised in Table 9. In addition to determine the effect of longer-term temporal modelling afforded by the hidden Markov models, a one-state HMM system, identical in all other ways to the proposed system, was implemented and its performance is also included in Table 9. Finally, a GMM (128-mixtures)-based system, trained on mel frequency cepstral coefficients (MFCCs) and ΔMFCCs, was also implemented for comparison.

The classification accuracies obtained by the system making use of all the model parameter contours is higher than the accuracies obtained by any of the individual systems, as expected. It should also be noted that the emotion-specific classification accuracies (diagonal elements of the confusion matrix) are all higher than 55%, indicating that the system does not suffer from any inherent bias against one or more of the emotions. Moreover, the performance of the combined system compares very well with the human classification performance. Also, the automatic (GMM-based) system that did not make use of any temporal information had an overall classification accuracy of 59.0%. This system modelled the distribution of all three component (pitch, glottal and vocal tract) parameters.

Table 5 Confusion matrix for HMM-based system using glottal parameter contours (MP), $\boldsymbol{v}^{(i)} = \left[m_{F_g}^{(i)}, m_{A_g}^{(i)}\right]$ (overall accuracy 55.0%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*72.3*	0	14.9	0	12.8
Anger	0	*83.1*	0	15.5	1.4
Sad	6.6	1.6	*36.1*	9.8	45.9
Happy	1.4	23.3	2.7	*54.8*	17.8
Bored	16.9	0	33.8	15.6	*33.8*

Table 6 Confusion matrix for HMM-based system using formant parameter contours (MP), $\boldsymbol{v}^{(i)} = \left[m_{F_1}^{(i)}, m_{F_2}^{(i)}, m_{m}^{F_3(i)}, m_{A_1}^{(i)}, m_{A_2}^{(i)}, m_{A_3}^{(i)}\right]$ (overall accuracy 45.0%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*42.6*	4.3	25.5	14.9	12.8
Anger	0	*66.2*	8.5	16.9	8.5
Sad	6.6	8.2	*37.7*	11.5	36.1
Happy	8.2	19.2	19.2	*37.0*	16.4
Bored	3.9	7.8	29.9	18.2	*40.3*

7.2. FAU Aibo corpus

Similar to the experiments performed on the LDC corpus, those performed on the FAU Aibo corpus were constructed as a five-class emotion classification problem. The emotions involved, however, were different (Anger, Emphatic, Neutral, Positive and Rest) as outlined in Section 4.2. The training and test sets for these classification experiments were taken as given in the guidelines to the INTERSPEECH 2009 Emotion Challenge [45]. Also in accordance with these guidelines, the metric used to quantify performance was the unweighted average recall (UAR) which should take into account relative imbalances in the number of occurrences of the different emotional states. However, the speaker normalisation technique employed in the experiments reported in this paper requires a priori knowledge of the distribution of speech features for each speaker (no knowledge of emotional class is required), and hence this normalisation technique could not have been applicable in the INTERSPEECH 2009 Emotion Challenge. The use of the normalisation technique was deemed acceptable since the aim of the paper is to investigate use of speech parameter contours for emotion recognition and not AER system optimisation and to make the results directly comparable to those obtained in a previous study [21] which is a precursor to the one reported in this paper.

Due to the significant differences between the two databases as outlined in Section 4, the parameters of the back-end, such as the number of states and the number

Table 7 Summary of overall accuracies for systems evaluated on LDC corpus

Parameter	Overall accuracy		
	HMM (S-I) (%)	HMM (MP) (%)	GMM (%)
Pitch (F_0)	*57.1*	56.2	46.6
Glottal (F_g, A_g)	52.6	*55.0*	46.6
Vocal tract (F_{1-3}, A_{1-3})	41.6	45.0	*48.2*

Table 8 Confusion matrix for the HMM-based system using all parameter contours, $\boldsymbol{v}^{(i)} = \left[\tau^{(i)}, s_{F_0}^{(i)}, b_{F_0}^{(i)}, m_{F_g}^{(i)}, m_{A_g}^{(i)}, {}_{m}^{F_1(i)}, m_{A_1}^{(i)}, m_{F_2}^{(i)}, m_{A_2}^{(i)}, m_{F_3}^{(i)}, m_{A_3}^{(i)}\right]$ (overall accuracy 62.6%)

	Neutral (%)	Anger (%)	Sad (%)	Happy (%)	Bored (%)
Neutral	*55.3*	0	23.4	4.3	17.0
Anger	0	*81.7*	0	16.9	1.4
Sad	3.3	0	*57.4*	14.8	24.6
Happy	0	23.3	5.5	*60.3*	11.0
Bored	3.9	0	32.5	7.8	*55.8*

of Gaussian mixtures in each state, used in the experiments performed on the LDC corpus may not be suitable for the Aibo corpus. A preliminary search was initially performed by comparing the accuracies of the systems constructed with a different number of states (ranging from 2 to 5) and a varying number of Gaussian mixtures in each state (ranging from 2 to 10). The features used in these systems were identical to those used in the experiment used to obtain the results in Table 8, i.e. $\boldsymbol{v}^{(i)} = \left[\tau^{(i)}, s_{F_0}^{(i)}, b_{F_0}^{(i)}, m_{F_g}^{(i)}, m_{A_g}^{(i)}, {}_{m}^{F_1(i)}, m_{A_1}^{(i)}, m_{F_2}^{(i)}, m_{A_2}^{(i)}, m_{F_3}^{(i)}, m_{A_3}^{(i)}\right]$. Consistently, the best UARs were obtained by systems using two-state HMMs. The optimal number of mixtures in each state is likely to be more dependent on feature dimensionality, but for the above-mentioned features, the highest UARs were obtained when five or six mixtures were used.

Using this back-end configuration, a series of experiments were carried out to compare the performances of systems that modelled temporal contours of pitch and glottal parameters and vocal tract parameters with the performances of systems that modelled only static distributions. GMM-based systems were used to model distributions without taking into account temporal contours. These experiments also compared the S-I representation to the MP representation for all three parameters and the results are summarised in Table 10. This comparison is identical to the one carried out on the LDC corpus and reported in Table 7.

Table 9 Comparison of overall accuracies obtained on LDC corpus

Classification test	Accuracy (%)
Human listeners (using 11 listeners) [21]	63.6
Automatic - using pitch contours (slope-bias)	57.1
Automatic - using glottal parameter contours (midpoint)	55.0
Automatic - using vocal tract parameter contours (midpoint)	45.0
Automatic - HMM (three-state)-based system using all model parameters	*62.6*
Automatic - HMM (one-state)-based system using all model parameters	59.3
Automatic - GMM-based system using all parameters (frame-based)	59.0
Automatic - GMM-based system using MFCC + ΔMFCC (frame-based)	51.1

Table 10 Summary of overall accuracies for systems evaluated on FAU Aibo corpus

Parameter	Overall accuracy		
	HMM (S-I) (%)	HMM (MP) (%)	GMM (%)
Pitch (F_0)	32.5	*33.3*	28.4
Glottal (F_g, A_g)	33.7	*37.4*	31.4
Vocal tract (F_{1-3}, A_{1-3})	31.6	29.9	*33.1*

It can be seen that the trends observed in the LDC corpus match the ones in these results. Similar to the results in Table 7, systems that make use of temporal information (HMM-based) outperform the static system (GMM) when pitch and glottal parameters are considered, but not when vocal tract parameters are considered.

Following this comparison, another experiment was performed to determine the best feature set. This would, in turn, reveal the best performance possible given the system setup and indicate what information is utilised. The back-end was fixed as a two-state HMM with a 6-mixture GMM modelling each state and various combinations of descriptors (slope, initial and midpoint values) of the different source-filter model parameter contours (Figure 4) used as features and the UAR determined for each setup. The five highest UARs obtained and the corresponding feature sets are reported in Table 11. Table 11 also includes the UAR obtained when the standard parameter contour feature set was used by the system. In comparison, the UAR

Table 11 Top 5 UARs obtained on the Aibo corpus and the corresponding feature set utilised

Feature set $\boldsymbol{v}^{(i)}$	UAR (%)
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{s}_{F_0}^{(i)}, \boldsymbol{b}_{F_0}^{(i)}, \boldsymbol{m}_{F_g}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{s}_{F_1}^{(i)}, \boldsymbol{b}_{F_1}^{(i)}, \boldsymbol{s}_{A_1}^{(i)}, \boldsymbol{b}_{A_1}^{(i)}\right]$	40.3
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{b}_{F_0}^{(i)}, \boldsymbol{m}_{F_0}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{s}_{F_1}^{(i)}, \boldsymbol{b}_{F_1}^{(i)}, \boldsymbol{s}_{A_1}^{(i)}, \boldsymbol{b}_{A_1}^{(i)}, \boldsymbol{\tau}^{(i)}\right]$	39.9
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{s}_{F_0}^{(i)}, \boldsymbol{b}_{F_0}^{(i)}, \boldsymbol{m}_{F_g}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{m}_{F_1}^{(i)}, \boldsymbol{m}_{A_1}^{(i)}, \boldsymbol{\tau}^{(i)}\right]$	39.9
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{s}_{F_0}^{(i)}, \boldsymbol{m}_{F_0}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{m}_{F_1}^{(i)}, \boldsymbol{m}_{A_1}^{(i)}, \boldsymbol{\tau}^{(i)}\right]$	39.8
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{s}_{F_0}^{(i)}, \boldsymbol{m}_{F_g}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{m}_{F_1}^{(i)}, \boldsymbol{m}_{A_1}^{(i)}, \boldsymbol{\tau}^{(i)}\right]$	39.8
$\boldsymbol{v}^{(i)} = \left[\boldsymbol{s}_{F_0}^{(i)}, \boldsymbol{b}_{F_0}^{(i)}, \boldsymbol{m}_{F_g}^{(i)}, \boldsymbol{m}_{A_g}^{(i)}, \boldsymbol{m}_{F_1}^{(i)}, \boldsymbol{m}_{F_2}^{(i)}, \boldsymbol{m}_{F_3}^{(i)}, \boldsymbol{m}_{A_1}^{(i)}, \boldsymbol{m}_{A_2}^{(i)}, \boldsymbol{m}_{A_3}^{(i)}, \boldsymbol{\tau}^{(i)}\right]$	*37.5*

achieved by a GMM-based system jointly modelled the static distributions of pitch, glottal parameters and vocal tract parameters was 35.8% and the baseline UARs of the INTERSPEECH 2009 emotion challenge were 35.9% with dynamic modelling and 38.2% with static modelling (without any speaker normalisation).

The trends in the feature sets corresponding to the five setups reported in Table 11 are potentially more interesting than the UARs themselves. In particular, the use of one-dimensional midpoint values (MP) is better than the use of two-dimensional S-I descriptions for glottal parameter contours (F_g and A_g), and two-dimensional representations are better than one-dimensional ones for pitch contours. This is in agreement with the systems evaluated on the LDC corpus (Tables 1,2,4 and 5).

Also of interest is that these results suggest that information about the first formant is significantly more important than information about the second and third formants. However, classification tests indicated that this does not hold for the LDC corpus where dropping information about the second and third formants caused the classification accuracy to reduce from 62.6% to 57.5%. The reason for this difference in experimental results obtained from the two databases is not clear, but it could be due to the subtlety of the emotional states in the Aibo corpus or the fact that the languages spoken in both databases are different.

8. Conclusion

Commonly automatic emotion recognition systems capture spectral information with frame-based information and temporal information by either computing a range of functionals over all the frame-level features in an utterance or by using a suitable back-end to model temporal variations of these frame-level features. In this paper we explore a combined approach, extracting 'short-term' temporal information in the front-end and modelling 'longer-term' temporal information with the back-end. In particular, this paper extends the idea of modelling F_0 contours using linear approximations in each voiced segment, the focus of earlier work, to other parameters of the source-filter model to parameterise short-term temporal variations prior to segment-by-segment modelling of sequences of these parameter vectors with a back-end. The work also has the advantage of not requiring explicit separation of speech into utterances.

As part of the parameterisation process, this paper has taken a second look at the traditional source-filter model widely used in speech processing tasks and, in particular, the assumption about the vocal excitation that is inherent in common feature extraction procedures. By estimating glottal spectral parameter contours, the system is not constrained by the assumption that the glottal spectrum can be modelled as the response of a system with a fixed transfer function. Earlier work had indicated that linear approximations to pitch contours were acceptable for the purpose of emotion classification, and another listening test conducted as part of the work reported in this paper revealed that similar approximations to glottal parameter contours were acceptable as well.

Furthermore, extending a previously developed AER system, which made use of linear approximations to pitch contours as features, to take into account linear approximations to contours of other parameters of the source-filter model as well, led to a relative increase in classification accuracy of 9.6% on the Emotional Prosody Speech and Transcripts corpus. Comparisons based on experiments using the LDC corpus revealed that an automatic emotion classification system that modelled contours outperformed one that modelled statistical distributions by 6.1% (relative) with regards to classification accuracy, and the classification accuracy of this contour-based system was comparable with human classification accuracy. The paper also includes the classification accuracies obtained when tested on the German FAU Aibo Emotion Corpus on the five-class emotion classification problem originally outlined in the INTERSPEECH 2009 Emotion Challenge. Work is currently underway on collecting an Australian speech corpus with emotional speech, and the use of the proposed contour-based features will be validated on those data in the future. Further, the choice of using linear approximations to model temporal information within voiced speech segments, while the simplest, may not be optimal and is an avenue for future work.

Competing interests
The authors declare that they have no competing interests.

Received: 31 July 2012 Accepted: 24 June 2013
Published: 10 July 2013

References

1. R Barra, JM Montero, J Macias-Guarasa, LF D'Haro, R San-Segundo, R Cordoba, *Prosodic and segmental rubrics in emotion identification, in Proceedings of the 2006 IEEE* (International Conference on Acoustics, Speech and Signal Processing, vol. 1 (IEEE, Piscataway, 2006), 2006), p. 1
2. M Borchert, A Dusterhoft, *Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments, in Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'05)* (Piscataway, IEEE, 2005), pp. 147–151
3. M Lugger, B Yang, *An incremental analysis of different feature groups in speaker independent emotion recognition, in Proceedings of the 16th International Congress of Phonetic Sciences, Saarbruecken, August 2007* (Washington, IEEE, 2007). pp. 2149–2152
4. M Pantic, LJM Rothkrantz, Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE **91**, 1370–1390 (2003)
5. D Ververidis, C Kotropoulos, Emotional speech recognition: resources, features, and methods. Speech Communication **48**, 1162–1181 (2006)
6. L Vidrascu, L Devillers, Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, in

Proceedings of International Workshop on Paralinguistic Speech - 2007. Saarbrücken **3**, 11–16 (August 2007)
7. S Yacoub, S Simske, X Lin, J Burns, Recognition of emotions in interactive voice response systems, in *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003*. Geneva **1–4**, 729–732 (September 2003)
8. D Bitouk, R Verma, A Nenkova, Class-level spectral features for emotion recognition. Speech Communication **52**, 613–625 (2010)
9. M El Ayadi, MS Kamel, F Karray, Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition **44**, 572–587 (2011)
10. C Lee, S Narayanan, R Pieraccini, *Combining acoustic and language information for emotion recognition, in Seventh International Conference on Spoken Language Processing* (September, Denver, 2002), pp. 873–876
11. B Schuller, A Batliner, S Steidl, D Seppi, *Emotion recognition from speech: putting ASR in the loop, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009* (Piscataway, IEEE, 2009), pp. 4585–4588
12. G Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960)
13. B Schuller, G Rigoll, M Lang, *Hidden Markov model-based speech emotion recognition, in Proceedings of International Conference on Acoustics* (Speech, and Signal Processing, (ICASSP'03). vol 2 2nd edn, IEEE, New York, 2003, 2003), pp. 1–4
14. A Nogueiras, A Moreno, A Bonafonte, J Mariño, *Speech emotion recognition using hidden Markov models, in Proceedings of EUROSPEECH-2001* (EUROSPEECH, Scandinavia, 2001), pp. 2679–2682
15. B Schuller, D Seppi, A Batliner, A Maier, S Steidl, *Towards more reality in the recognition of emotional speech, in Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, April 2007*, vol. 4 (Piscataway, IEEE, 2007), pp. 941–944
16. B Vlasenko, B Schuller, A Wendemuth, G Rigoll, *Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing, in Affective Computing and Intelligent Interaction* (Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing, in Affective Computing and Intelligent Interaction, Springer Berlin, Heidelberg, 2007, 2007), pp. 139–147
17. S Planet, I Iriondo, J Socoró, C Monzo, J Adell, GTM-URL contribution to the INTERSPEECH 2009 Emotion Challenge, in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)*. Brighton **6–10**, 316–319 (September 2009)
18. S Wu, TH Falk, W-Y Chan, Automatic speech emotion recognition using modulation spectral features. Speech Communication **53**, 768–785 (2011)
19. P Dumouchel, N Dehak, Y Attabi, R Dehak, N Boufaden, Cepstral and long-term features for emotion recognition, in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)*. Brighton **6–10**, 344–347 (September 2009)
20. T Moriyama, S Ozawa, *Emotion recognition and synthesis system on speech, in IEEE International Conference on Multimedia Computing and Systems, 1999 vol 1*, 1st edn. (IEEE, New York, 1999), pp. 840–844
21. V Sethu, E Ambikairajah, J Epps, Pitch contour parameterisation based on linear stylisation for emotion recognition, in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH-2009)*. Brighton **6–10**, 2011–2014 (September 2009)
22. DG Childers, CK Lee, Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. **90**, 2394–2410 (1991)
23. J Cabral, S Renals, K Richmond, J Yamagishi, *Towards an improved modeling of the glottal source in statistical parametric speech synthesis, in 6th ISCA Workshop on Speech Synthesis* (Bonn, Germany, 2007)
24. C D'Alessandro, B Doval, *Voice quality modification for emotional speech synthesis, in Proceedings of the 8th European Conference on Speech Communication and Technology* (EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, 2003), pp. 1653–1656
25. L He, M Lech, N Allen, On the importance of glottal flow spectral energy for the recognition of emotions in speech, in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. Makuhari, Chiba, Japan **26–30**, 2346–2349 (September 2010)
26. J Tao, Y Kang, *Features importance analysis for emotional speech classification, in Affective Computing and Intelligent Interaction* (ed. by J Tao et al, Springer Berlin, Heidelberg, 2005, 2005), pp. 449–457
27. S Rui, E Moore, *JF Torres, Investigating glottal parameters for differentiating emotional categories with similar prosodics, in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009* (IEEE, New York, 2009), pp. 4509–4512
28. G Fant, J Liljencrants, Q Lin, A four-parameter model of glottal flow. STL-QPSR **4**, 1–13 (1985)
29. AE Rosenberg, Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am. **49**, 583–590 (1971)
30. DH Klatt, LC Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. **87**, 820–857 (1990)
31. R Veldhuis, A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. J. Acoust. Soc. Am. **103**, 566–571 (1998)
32. B Doval, C d'Alessandro, N Henrich. The spectrum of glottal flow models. Acta Acustica united with Acustica **92**, 1026–1046 (2006)
33. P Alku, *Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, in Proceedings of the EUROSPEECH-1991* (ICSA, Mechanicsburg, 1991), pp. 1081–1084
34. J Cabral, S Renals, K Richmond, J Yamagishi, *Glottal spectral separation for parametric speech synthesis, in Proceedings of INTERSPEECH-2008* (ICSA, Brisbane, 2008), pp. 1829–1832
35. M Frohlich, D Michaelis, HW Strube, SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. J. Acoust. Soc. Am. **110**, 479–488 (2001)
36. L Hui-Ling, *JO Smith III, Joint estimation of vocal tract filter and glottal source waveform via convex optimization, in 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (Piscataway, IEEE, 1999), pp. 79–82
37. EL Riegelsberger, AK Krishnamurthy, *Glottal source estimation: methods of applying the LF-model to inverse filtering, in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93, vol 2*, 2nd edn. (Piscataway, IEEE, 1993). pp. 542–545
38. D Vincent, O Rosec, T Chonavel, Estimation of LF glottal source parameters based on an ARX model, in *Proceedings of INTERSPEECH 2005 - EUROSPEECH, 9th European Conference on Speech Communication and Technology*. Lisbon **4–8**, 333–336 (September 2005)
39. PA Naylor, K Anastasis, G Jon, B Mike, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Transactions on Audio, Speech, and Language Processing **15**, 34–43 (2007)
40. D Wang, S Narayanan, Piecewise linear stylization of pitch via wavelet analysis, in *Proceedings of INTERSPEECH 2005 - EUROSPEECH, 9th European Conference on Speech Communication and Technology*. Lisbon **4–8**, 3277–3280 (September 2005)
41. S Ravuri, *DPW Ellis, Stylization of pitch with syllable-based linear segments, in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008* (Piscataway, IEEE, 2008), pp. 3985–3988
42. M Liberman, K Davis, M Grossman, N Martey, *J Bell, Emotional prosody speech and transcripts* (Linguistic Data Consortium (LDC) database, LDC catalog no. LDC2002S28, 2007). http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28. ISBN 1-58563-237-6 (2007)
43. S Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech* (Logos Verlag, Berlin, 2009)
44. R Banse, K Scherer, Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. **70**, 614–636 (1996)
45. B Schuller, S Steidl, A Batliner, *The INTERSPEECH 2009 Emotion Challenge, in INTERSPEECH-2009* (ISCA, Brighton, 2009), pp. 312–315
46. R McAulay, T Quatieri, Speech analysis/synthesis based on a sinusoidal representation. Acoustics, Speech and Signal Processing, IEEE Transactions on **34**, 744–754 (1986)
47. D Talkin, *A robust algorithm for pitch tracking (RAPT), in Speech Coding and Synthesis* (by W Kleijn, K Paliwal, Elsevier, New York, 1995), pp. 495–518
48. TL Nwe, SW Foo, LC De, Silva, Speech emotion recognition using hidden Markov models. Speech Communication **41**, 603–623 (2003)
49. R Huang, C Ma, *Toward a speaker-independent real-time affect detection system, in 18th International Conference on Pattern Recognition ICPR 2006*, vol. 1 (IEEE, New York, 2006), pp. 1204–1207
50. V Sethu, E Ambikairajah, J Epps, *Speaker normalisation for speech-based emotion detection, in 15th International Conference on Digital Signal Processing, 2007* (Piscataway, IEEE, 2007), pp. 611–614

doi:10.1186/1687-4722-2013-19
Cite this article as: Sethu *et al.*: **On the use of speech parameter contours for emotion recognition.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:19.