Circuits, Systems, and Signal Processing (2022) 41:2061–2087
https://doi.org/10.1007/s00034-021-01875-7

Incorporation of Happiness in Neutral Speech by Modifying Time-Domain Parameters of Emotive-Keywords

Anushiya Rachel Gladston[1] · S. Sreenidhi[2] · P. Vijayalakshmi[3] · T. Nagarajan[4]

Received: 30 August 2020 / Revised: 5 October 2021 / Accepted: 6 October 2021 /
Published online: 10 November 2021

Abstract

Human-computer interactions can be enhanced by making machines recognize the emotional state of a user and respond accordingly. This necessitates text-to-speech systems that can produce natural emotional speech. While several existing methods are data driven, the current work attempts to incorporate happiness into neutral speech using signal processing algorithms. Analysis shows that it is mainly the speech-rate, pitch-period, and energy that exhibit variations due to emotion. Further, emotion is predominantly expressed in certain emotive words in the sentence. In this regard, several variations are introduced into the parameters mentioned before, and it is observed that fitting a hat-shaped pitch-contour onto the emotive keywords in a sentence and increasing their energy, suffices to incorporate happiness into neutral speech. An HMM-based approach is used to spot the keywords. Linear prediction-based synthesis and time-domain pitch-synchronous overlap-and-add method are then used to modify the keywords and synthesize emotional speech. The latter produces happy speech of better quality with a mean opinion score (MOS) of 2.51, out of a maximum of 3. Further, to verify if modifying the keywords would suffice, happy speech is also

✉ Anushiya Rachel Gladston
anushiyarachel@karunya.edu ; anushiyarachel89@gmail.com

S. Sreenidhi
sreenidhissn@gmail.com

P. Vijayalakshmi
vijayalakshmip@ssn.edu.in

T. Nagarajan
nagarajant@snuchennai.edu.in

1 Karunya Institute of Technology and Sciences, Coimbatore, India

2 Edgeverve Systems Ltd., Chennai, India

3 SSN College of Engineering, Chennai, India

4 Shiv Nadar University, Chennai, India

synthesized by modifying all the words of a neutral utterance to match the corresponding natural happy speech. An MOS of 2.34 is obtained for speech synthesized by this method, revealing that modifying the keywords would suffice to incorporate happiness into neutral speech. Finally, the use of the proposed method as a post-processing module in a text-to-speech synthesis system, to generate happy speech instead of neutral speech, is also demonstrated.

Keywords Emotion incorporation · Keyword-spotting · Polynomial curve fitting · PSOLA · Emotional speech synthesis

1 Introduction

A text-to-speech (TTS) system is one that is capable of producing intelligible and natural speech corresponding to any given text. Conventional synthesizers generate speech that is devoid of emotion. To improve human-computer interactions it would be more desirable if the systems were capable of producing speech with appropriate emotions. This is essential in applications where speech is the primary mode of communication with the machine. For example, in automatic call centers and e-learning, a system that is capable of recognizing the emotion of the user and responding appropriately, would be preferable. Further, TTS systems can be used to assist people who are visually challenged, and it is important to deliver information in the right emotion. TTS systems that generate emotional speech can also be used to assist people who have difficulties speaking and expressing their emotions.

Emotional speech can be synthesized in one of three ways: (i) training TTS systems with emotional data, (ii) making modifications to the synthesis algorithm in conventional TTS systems, or (iii) modifying the speech signal, post-synthesis. Some of the popularly used techniques for speech synthesis are unit selection synthesis (USS), HMM-based speech synthesis and DNN-based speech synthesis. Unit selection synthesis involves the concatenation of pre-recorded speech units to synthesize speech. The first method, which requires a TTS system to be trained with emotional speech data, can be applied to a USS system, as in [5], where based on the required emotion, speech units are chosen from the appropriate emotion-specific inventory.

The second method of emotional speech synthesis, which involves modifying the synthesis algorithm, can be incorporated in formant-based, HMM-based and DNN-based speech synthesis systems. The rule-based formant synthesis produces speech based on a set of rules. In order to generate emotional speech, these rules are fine-tuned based on the emotion-specific variations in the fundamental frequency, formant frequencies, speech rate and articulation precision [12,18]. To synthesize emotional speech in HMM-based speech synthesis, style-dependent or style-mixed modeling approach can be adopted. The former involves training models for each emotion and performing tree-based clustering separately for each emotion, while the latter involves adding the emotion along with the contextual information and performing tree-based clustering on all the emotions at the same time [27]. In [28], another HMM-based approach for emotional speech synthesis in a target speaker's (T's) voice, in the absence of emotional speech from that speaker, is described. Here HMMs are initially trained

for neutral and emotional speech data of different speakers of an emotional database. From this pool of models, an emotional model (E_X) and a neutral model (N_Z), corresponding to speakers X and Z respectively, that are closer to a target speaker's neutral model (N_T) are chosen. N_T and E_X are interpolated to form a new model, E_Y. The target speaker's emotional model, E_T is then derived from E_Y and E_Z (emotional model corresponding to the speaker Z), and used to synthesize emotional speech. In order to synthesize emotional speech using a DNN-based system, [25] proposes the use of global style tokens, which are a bank of embeddings trained within Tacotron, a DNN-based end-to-end TTS system.

In order to incorporate emotion in speech, post-synthesis, model-based or signal processing algorithms can be used. The model-based approaches involve training Gaussian mixture models (GMM), linear modification models (LMM), classification and regression trees (CART), etc., with neutral and emotional speech data. A GMM rule-based approach to synthesizing emotional speech is described in [7], where the spectral features, namely the line spectral frequencies extracted from neutral and emotional speech are used to train a GMM and derive a conversion function. A rule-based method is then used to modify the prosody. In [21], the pitch contour, duration, and stress information are extracted from emotional speech, at the syllable level. These are used to train an LMM, GMM, and CART for incorporating emotion in neutral speech. In this, LMM directly modifies the prosodic features by a certain factor. GMM, and CART are trained to establish a mapping between neutral and emotional speech, and both outperform LMM. While GMM enables a smooth and continuous conversion to emotional speech, it does not incorporate linguistic information, however CART takes into account the contextual features as well. Therefore GMM is better suited for a small amount of training data, while CART performs well with a large context-balanced data set as well. A hierarchical prosody conversion technique is described in [26], where prosodic features are extracted at the sentence, prosodic word, and sub-syllable levels. GMMs are trained to convert these features at the sentence and word levels. At the sub-syllable-level, GMM is used to convert the duration, while CART is used to transform the pitch. Further, a GMM-based spectrum conversion is also performed. Here, STRAIGHT algorithm is used to extract the excitation and spectral features and also for synthesis. More recently deep learning methods have also been used to introduce emotion. In [1], a variational autoencoder is used to model global characteristics of speech, such as speaking styles, thereby aiding in making the speech synthesized by an autoregressive synthesis system more expressive.

The afore mentioned techniques require a large amount of emotional speech data. However, collecting emotional data is not a trivial task and several factors come into play, such as the text collection, degree of emotion [4], recording conditions, and choice of speakers [12]. Therefore, when there is a scarcity in emotional data, signal processing algorithms can be applied to neutral speech to incorporate emotions. As evident from the discussion so far, variations in speech due to emotions occur primarily in the pitch frequency, formant frequencies, energy, and speech rate [12,21,29]. Therefore, these parameters can be extracted from a given speech signal and modified appropriately to incorporate emotion.

An important requirement when modifying prosodic features of a speech signal is an estimate of pitch marks or instants of excitation. Several algorithms have been proposed

in literature, for the estimation of instants of excitation, namely, group delay-based algorithm, Dynamic Programming Projected Phase-Slope Algorithm (DYPSA), Zero Frequency Filtering (ZFF), etc. In [19], a group delay-based algorithm is discussed, where the unwrapped phase slope function of the short-time Fourier transform of the LP residue is initially calculated. The instants where the phase slope function makes positive zero crossings are identified as the instants of significant excitation. DYPSA [11] consists of three primary steps. First, candidate instants are located from the zero crossings of the energy-weighted formulation of the phase slope function of the LP residue. Then, the missing instants are identified by phase slope projection and the false instants are eliminated by dynamic programming. ZFF described in [13], low pass filters the speech signal by passing it twice through a 0 Hz resonator and removes the trend in the filtered signal. The positive zero crossings in the signal so obtained, indicate the location of the instants of excitation. More recently, a phase-difference-based approach has been proposed in [3]. This algorithm considers a symmetrized speech signal to be the Fourier Transform of an arbitrary even signal, so that the negative valleys at the instants of excitation would correspond to zeros that lie outside the unit circle, on the z-plane. The phase-difference (PD) is observed to be equal to 2π at the angular locations of zeros lying outside the unit circle. Therefore, this algorithm identifies the locations in the PD spectrum of the even signal that have a value of 2π to identify instants of excitation. The analysis in [3] reveals that this algorithm outperforms other state-of-the-art algorithms.

Once the instants of excitation are estimated, the techniques described in [17] and [15] to modify the pitch contour and duration of speech can be used to incorporate emotions. In the modification of duration of a speech signal, major changes occur in the vowel regions. In [17] it is suggested that the duration varies non-uniformly in different vowels. Therefore, instants of significant excitation are derived and the vowel regions are identified. The authors of [17], then classify the vowels using HMMs and depending on the identity of the vowel, the epoch intervals are modified by different factors. The modified instants derived, are later used to synthesize speech. Similarly, in [24], the energy, pitch period and the duration of vowels in a given neutral speech signal are modified non-uniformly, based on the context in which these vowels occur, to incorporate emotion. In [15], the pitch contour is modified by resampling the LP residue. The samples around the instants of excitation are left unmodified to preserve naturalness. The modified residue is then used to synthesize emotional speech. In addition to modifying the duration and pitch contour, the system features can also be altered, as in [8], by replacing the LPCs of neutral speech by the best matching LPCs of emotional speech, using dynamic time warping (DTW). In [30], duration, pitch contour, and formant frequencies are modified. PSOLA is then used to synthesize emotional speech with the modified parameters.

Excessive processing of the signal could lead to degradation in the quality of emotional speech. In [12], it is noted that emotion is observed in certain content words, in the sense that the stress on the word, pitch frequency and articulation precision varied in accordance with the emotion. In this regard, [10] describes a unit selection synthesis system, where only certain words that are identified to be important are synthesized from an emotional database, while the others are synthesized from a neutral database. The Dictionary of Affect, which scores each word based on the

valence (negative/positive rating) and arousal (mild/intense rating), is used to identify the words which are to be synthesized with the particular emotion. Therefore, instead of carrying out signal processing algorithms on the entire speech signal, only certain keywords may be modified.

The current work focuses on incorporating happiness into neutral speech using signal processing algorithms to modify only the emotive-keywords. Initially, the variations in time-domain parameters, namely, the pitch, duration, and amplitude are analyzed. The PD algorithm is used to derive the instants of excitation, and based on the observations, the appropriate changes are incorporated in the emotive keywords, which are identified using an HMM-based speech recognition system. LP-based or TD-PSOLA-based synthesis is then carried out. The emotional speech synthesized by this technique is evaluated by the mean opinion score and a GMM-based emotion recognition system. Further, the time-domain parameters of all the words in a neutral utterance are modified in accordance with the corresponding emotional speech. Emotional speech so synthesized is compared with the emotional speech synthesized by modifying only the keywords, to verify if there is no significant difference in the extent to which emotion is perceived.

In this regard, the current work primarily differs from the existing signal processing methods described earlier in two ways: (i) While most existing methods modify prosodic features in all words of a neutral utterance, the proposed work aims to incorporate emotion by modifying only the emotive-keywords in the utterance. (ii) While several existing methods operate at the vowel/sub-word level, the current work performs analyses predominantly at the word level and incorporates modification at the word level as well.

The paper is organised as follows: Sect. 2 describes the speech corpora used to analyze emotional speech, Sect. 3 discusses the parameters that affect emotion in speech, Sect. 4 describes how the time-domain parameters are modified in the current work, Sect. 5 describes HMM-based keyword spotting, Sect. 6 elaborates the incorporation of happiness in neutral speech, Sect. 7 discusses the quality of the synthesized emotional speech, Sect. 8 describes the use of the proposed method in a TTS system, and Sect. 9 concludes the paper.

2 Speech Corpora

In the current work, analysis on emotional speech is carried out on the following corpora:

2.1 Berlin Database

The Berlin database [6] is an acted emotional database. It consists of German speech data in 6 emotions, namely, happiness, sadness, anger, fear, disgust, and surprise, along with neutral speech. 10 sentences per emotion are collected from 10 speakers (5 male and 5 female), 9 of whom are trained actors. The recordings are performed in

an anechoic chamber, at a sampling rate of 48 kHz. The data is then down-sampled to 16 kHz.

2.2 Surrey Audio-Visual Expressed Emotion (SAVEE) Database

The SAVEE database [9] is an audio-visual emotional database. It consists of English data recorded from four amateur, male, native English speakers, at a sampling rate of 44.1 kHz. For the current work, the audio is extracted and down-sampled to 16 kHz. The database covers the same emotions as the Berlin database. There are 15 sentences per emotion, three of which are common to all emotions, two are emotion specific and the rest are generic sentences, different for each emotion. The common and emotion specific sentences are also recorded without emotion, along with 15 other sentences. All sentences are chosen from the TIMIT database and are phonetically-balanced.

2.3 Own Database

In addition to the above mentioned corpora, to further analyze happy speech, a speech corpus is developed by the authors. 20 English sentences and 10 Tamil sentences are recorded by a native-Tamil female speaker, with and without happy emotion. Another 178 English and 40 Tamil sentences are collected from the same speaker in a neutral tone. These sentences are framed using 63 keywords and their derivatives, such that there is at least one emotive-keyword per sentence. All sentences used in the database are emotion-specific. The recording is carried out in a laboratory environment, at a sampling rate of 16 kHz. The speech data is manually segmented at the word-level. Further, the phoneme-level segmentation is derived by initially performing forced-Viterbi alignment using monophone models trained on the TIMIT database. This procedure involves the following steps.

- 39 dimensional Mel frequency cepstral coefficients are extracted from the training data of the TIMIT corpus.
- Context-independent phoneme models, with 5 states and number of mixture components based on the occurrence of each phoneme, are then trained.
- The forced-Viterbi alignment procedure is carried out on the neutral and happy speech of our own database.

The phoneme boundaries obtained are then manually corrected.

3 Analysis of Happy Speech

Literature suggests that emotion in speech primarily affects the short-time energy or intensity, pitch frequency, formant frequencies, and speech rate. Therefore, in the current work, an analysis is carried out on these parameters, with the corpora described in Sect. 2.

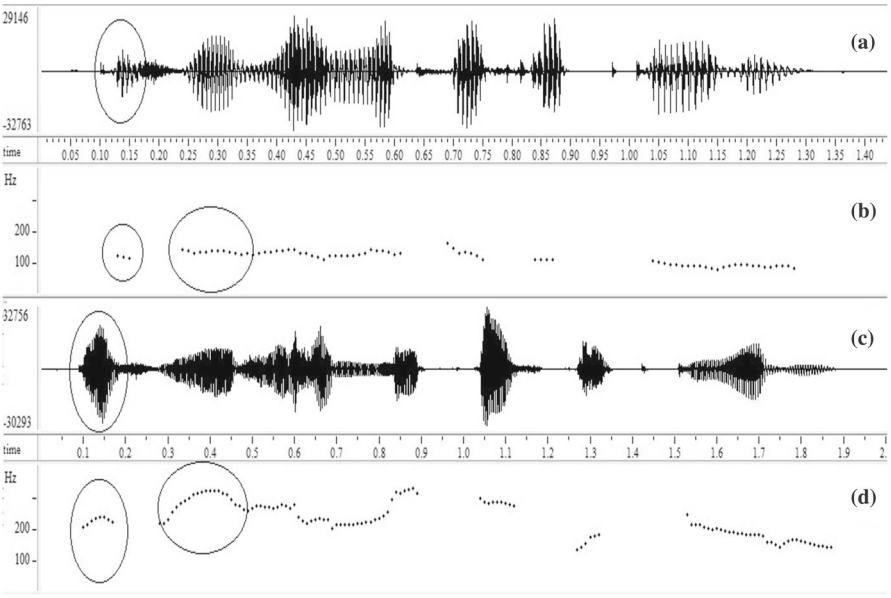

Fig. 1 Comparison of neutral and happy speech of Berlin database **a** Neutral speech **b** Pitch contour of neutral speech **c** Happy speech **d** Pitch contour of happy speech

3.1 Pitch Contour

The pitch contours of neutral speech and the corresponding happy speech are derived using the Entropics Signal Processing Software (ESPS) algorithm [20] and compared. The analysis is initially performed on the speech recorded from a male speaker of the Berlin database. It is observed that the neutral speech being monotonous, has an almost flat pitch contour, while the contour of happy speech exhibits greater amount of variations. Further, the mean pitch frequency of happy speech is approximately 1.5 times higher than that of neutral speech. These observations are portrayed in Fig. 1, which shows the neutral and happy versions of the sentence "Das will sie am Mittwoch abgeben" (meaning, "She will hand it in on Wednesday"), spoken by a male speaker, and their pitch contours.

The analysis is then extended to the SAVEE database (5 sentences per speaker) and the 20 sentences from our own corpus and a similar inference is made. It is also observed that the variations in pitch contour are predominantly observed in the stressed words and these words possess a hat-shaped contour. Figure 2a, c show the neutral and happy versions of the sentence, "Those musicians harmonize marvelously", uttered by a male speaker in the SAVEE database, and Fig. 2b, d show the corresponding pitch contours. The circled regions highlight the stressed word, namely, "marvelously", that possess a hat-shaped contour.

Analysis reveals that the stressed words are generally those that express happiness (such as "great", "happy", "awesome", etc.) and are therefore referred to as emotive-keywords, henceforth.

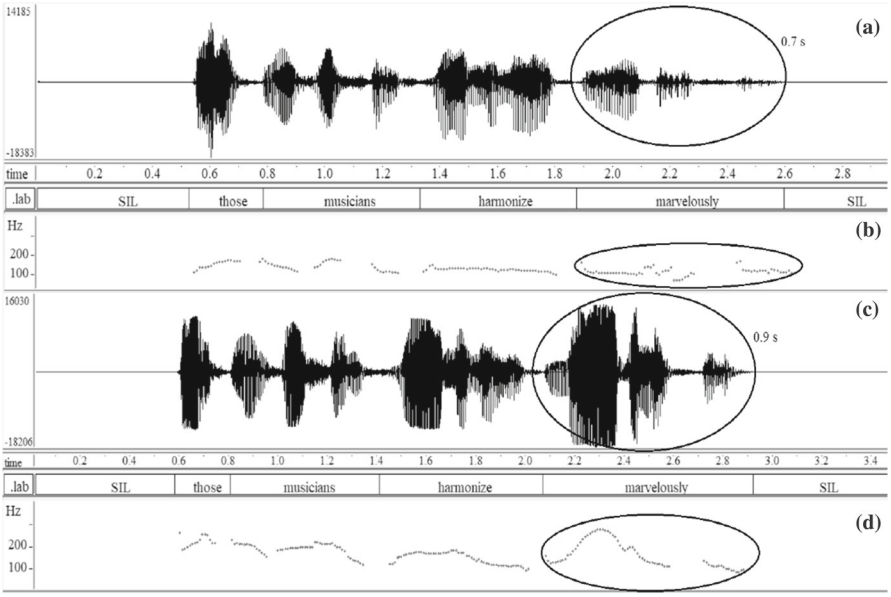

Fig. 2 Comparison of neutral and happy speech of SAVEE database **a** Neutral speech **b** Pitch contour of neutral speech **c** Happy speech **d** Pitch contour of happy speech

3.2 Speech Rate

The speech rates of neutral and happy speech are compared in terms of the duration of each word in an utterance and that of the utterance as a whole. It is observed that the overall duration of happy speech is less than that of the corresponding neutral speech, that is, the speech rate of happy speech is higher. However, the duration of certain emotive-keywords are greater, owing to the emphasis placed on them. The extent to which the duration of happy speech varies from the corresponding neutral speech, differs with the context and the speaker. These inferences can be observed in Fig. 2a, c, where the variation in duration of each word is different, and the duration of the keyword "excitement" is 1.2 times longer in the happy speech than in the neutral speech.

Since the change in speech rate between neutral speech and the corresponding happy speech is inconsistent at the sentence level and the word level, an analysis is carried out on the vowels within the keywords. The 20 English sentences, with parallel happy and neutral utterances, from our own database and the emotion-specific sentences from the SAVEE database are considered for this analysis, resulting a total of 113 vowels. The analysis reveals that approximately 60% of the vowels in happy speech and neutral speech have a similar duration, while around 5% of the vowels in happy speech have a lower duration than that of the vowels in neutral speech. Around 28% of the vowels in happy speech have a duration that is between 1.2 and 2 times higher and only about 6% have a duration that is more than twice that of the vowels in neutral speech. This is depicted in Fig. 3, where the x-axis denotes the number of vowels and the y-axis

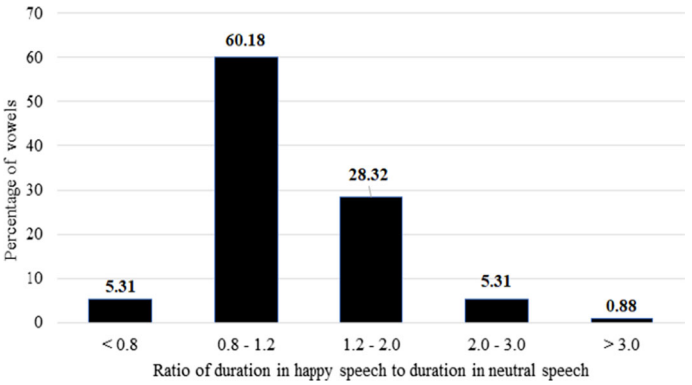

Fig. 3 Duration analysis on the vowels within the emotive-keywords

denotes the ratio of the duration of a vowel in happy speech to that of the same vowel in neutral speech.

3.3 Short-Time Energy

The short-time energy of a signal is directly related to its amplitude or intensity. It is computed frame-by-frame as shown in the following equation,

$$E_n = \sum_{m=-\infty}^{\infty} [x\,(m)\,w\,(n-m)]^2 \tag{1}$$

where w is a Hamming window of size 25 ms, in the current work. Comparison of the energy/amplitude of happy and neutral speech reveals that happy speech possesses a higher amplitude owing to the excitement in the speaker's voice. Further, the intensities of the emotive-keywords in happy speech are greater that those of the other words in the utterance. This can be observed in Fig. 2a, c.

3.4 Formant Frequencies

In order to analyze the effect of happiness on the formant frequencies, the linear prediction (LP) spectra of order 20, of neutral and happy speech are compared. It is observed that the first formant remains almost unaffected, while the second and third formants of happy speech vary from those of the corresponding neutral speech. This is observed in Fig. 4, where the LP spectra of the vowel /ao/ of neutral and happy speech are shown. However, in analyzing the second and third formants of all the vowels in the database, the variations observed are not consistent, and so the current work focuses on modifying the time-domain parameters, namely, the pitch period, duration, and energy, to incorporate happiness into neutral speech.

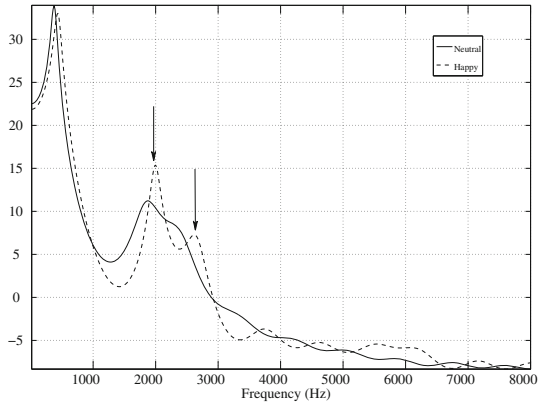

Fig. 4 Comparison of LP spectra of the vowel /ao/ in neutral and happy speech

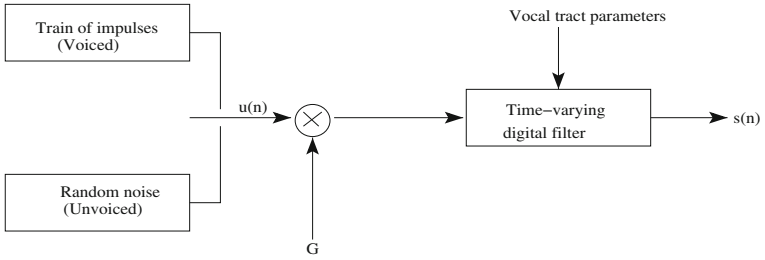

Fig. 5 Source-filter model for speech production (redrawn from [14])

4 Modification of Time Domain Parameters

Analysis in Sect. 3 reveals that emotion in speech mainly affects the pitch contour, speech rate, and the energy. To modify these parameters, speech is first decomposed into its source and system components. The desired changes are incorporated into the source and system, and speech possessing the desired characteristics is synthesized, either using the source-filter model or TD-PSOLA.

4.1 Decomposition of Speech

Liner predictive analysis [14] is used to decompose speech into its source and system components. It works on the principle that a sample of speech signal, can be represented as a linear combination of p past samples. Consider the source-filter model of speech production in Fig. 5, based on which, a speech signal $s(n)$ is generally represented by the following equation,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{2}$$

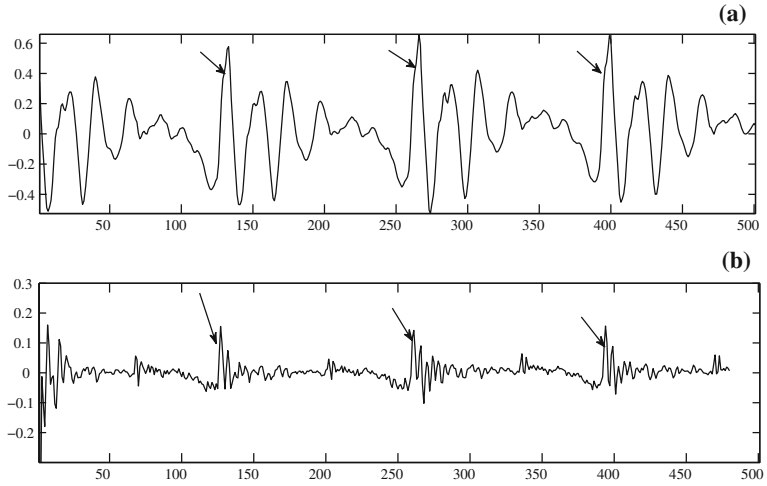

Fig. 6 **a** Segment of the vowel /a/ **b** LP residue

where a_k and G are the coefficients and gain parameter of the time-varying digital filter (system) and $u(n)$ is the excitation (source).

A linear predictor of $s(n)$ of order p, with prediction coefficients α_k, is given by

$$\tilde{s(n)} = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3}$$

The transfer function of this predictor is the following

$$P(z) = \frac{S'(z)}{S(z)} = \sum_{k=1}^{p} \alpha_k z^{-k} \tag{4}$$

The basic requirement of linear prediction is to accurately estimate the coefficients α_k, which are the parameters of the filter transfer function, that is, the system features of a speech signal. These are estimated by minimizing the prediction error, $e(n)$ defined as

$$e(n) = s(n) - \tilde{s(n)} \tag{5}$$

If the coefficients α_k are correctly estimated, that is, if $\alpha_k = a_k$, then, from Eqs. 2 and 3, the prediction error, $e(n)$ is defined by the following equation

$$e(n) = Gu(n) \tag{6}$$

Therefore, the prediction error or the LP residue gives the source or excitation. The residue contains prominent peaks at the instants of excitation, as shown in Fig. 6 for a segment of the vowel /a/. These instants are derived using the PD algorithm, described in Sect. 1.

In the current work, a linear predictor of order 20 is used. Once the source and system parameters, namely, the instants of excitation and the LP coefficients are estimated, they are modified as described in the following section.

4.2 Modification of Source and System Parameters

4.2.1 Pitch Modification

In order to modify the pitch frequency, modifying the source will suffice. Initially, the instants of excitation, $u(n)$ are to be derived using one of the algorithms described in Sect. 1. The current work uses the PD algorithm [11]. Once the instants are derived, the following modifications can be imposed on the fundamental frequency.

- Uniformly increasing or decreasing the frequency, while maintaining the contour
- Fitting a rising contour
- Fitting a falling contour
- Fitting a hat-shaped contour
- Fitting a bucket-shaped contour

To increase or decrease the pitch frequency without changing the contour, the instants of excitation are interpolated or decimated by the desired factor. Figure 7 shows the instants of excitation derived from a segment of the vowel /a/, uttered by a male speaker (pitch period of 5.6 ms), the instants interpolated by a factor of 1.5 (resulting in a pitch period of 3.7 ms), and the instants decimated by a factor of 1.5 (resulting in a pitch period of 8.4 ms).

In order to modify the pitch contour of a speech signal, polynomial curve fitting is used. Based on the required shape, the order of the polynomial is first chosen. For a rising or falling contour, a polynomial of order 1 is used, whereas for a hat-shaped or bucket-shaped contour, a polynomial of order 2 or higher (up to 5) is used. These polynomials are designed based on the desired minimum and maximum pitch frequencies, and the required pitch contour and the corresponding instants of excitation, $u'(n)$ are derived.

4.2.2 Duration Modification

To modify the speech rate or duration, the instants of excitation are first extracted using the PD algorithm, as in the case of pitch modification. The instants are replicated or deleted to increase or decrease the duration respectively. In order to use the source-filter model for synthesis, the system features, namely the LP coefficients are also replicated or deleted by the same factor as the instants of excitation. To use TD-PSOLA, modifying the instants would suffice.

4.3 Synthesis

Now that the instants of excitation and the LP coefficients are modified, speech with the desired characteristics is synthesized either using the source-filter model or TD-PSOLA as described below.

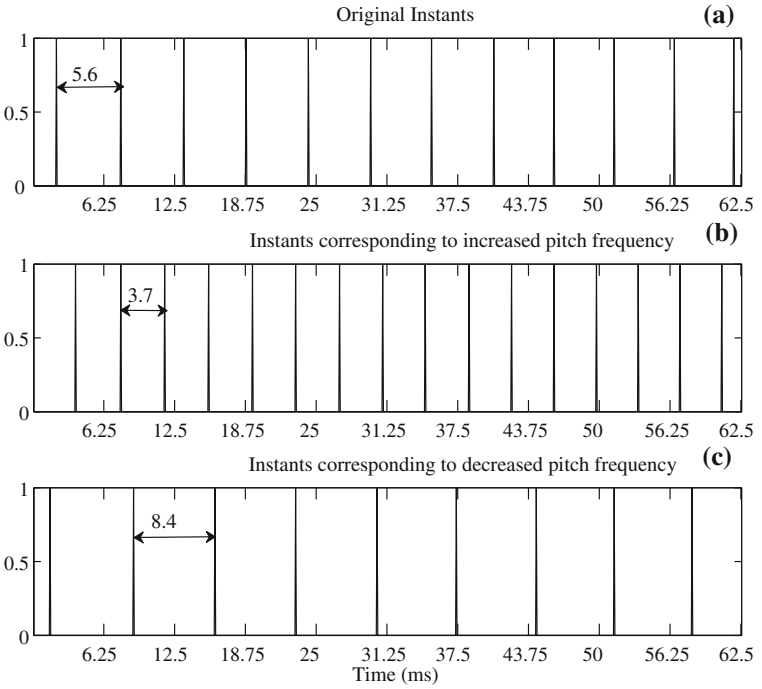

Fig. 7 Instants of excitation from a segment of the vowel /a/ spoken by a male speaker **a** Original speech **b** Speech with pitch period decreased by a factor of 1.5 **c** Speech with pitch period increased by a factor of 1.5

4.3.1 Source-Filter Model

The source-filter model of speech is shown in Fig. 5, where a time-varying digital filter is excited by a train of impulse (voiced sounds) or random noise (unvoiced sounds), to generate a speech signal. In this regard, a train of impulses is generated with the modified instants of excitation, $u'(n)$ and is used to excite a filter with the modified coefficients, α'_k, to synthesize speech, $s'(n)$, with the preferred pitch contour and duration.

4.3.2 TD-PSOLA

Time domain pitch synchronous overlap and add method is commonly used to make prosodic modifications directly on the speech signal, thereby retaining a high level of naturalness. It works pitch synchronously and therefore requires an estimate of the pitch marks or instants of excitation. The speech signal to be modified is segmented using a Hamming window of size equal to the pitch period, such that each segment is centred at the instant of excitation, as shown in Fig. 8.

The segments of speech are represented as,

$$s_i(n) = s(n)w(n - iP) = s(n - iP)w(n - iP) \tag{7}$$

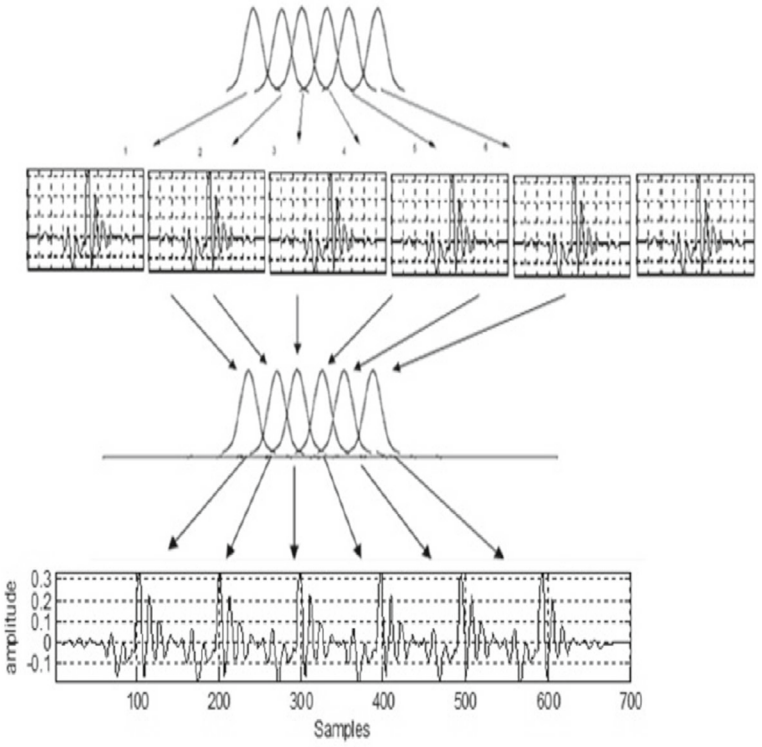

Fig. 8 Pitch/duration modification using TD-PSOLA

where P is the pitch period and w is a Hamming window.

To modify the pitch contour or duration, based on the new instants of excitation, the segments of speech, $s_i(n)$ containing instants of excitation, closer to the new instants, are selected. These segments are overlapped and added as in Eq. 8, to obtain the speech signal, $s'(n)$ with the desired pitch contour and duration.

$$s'(n) = \sum_{i=1}^{N} s_i(n) \tag{8}$$

Figure 9 illustrates the use of PSOLA to reduce the pitch period of a signal by a factor of 2.

To illustrate pitch contour and duration modification using TD-PSOLA, a segment of the vowel /a/, modified to bear the four different pitch contours (with maximum and minimum pitch periods of 10.4 ms and 6.4 ms, respectively), is shown in Fig. 10, and the sentence "Will you tell me why", spoken by a male speaker of the SAVEE database and the duration modified (by a factor of two) version of the speech signal are shown in Fig. 11.

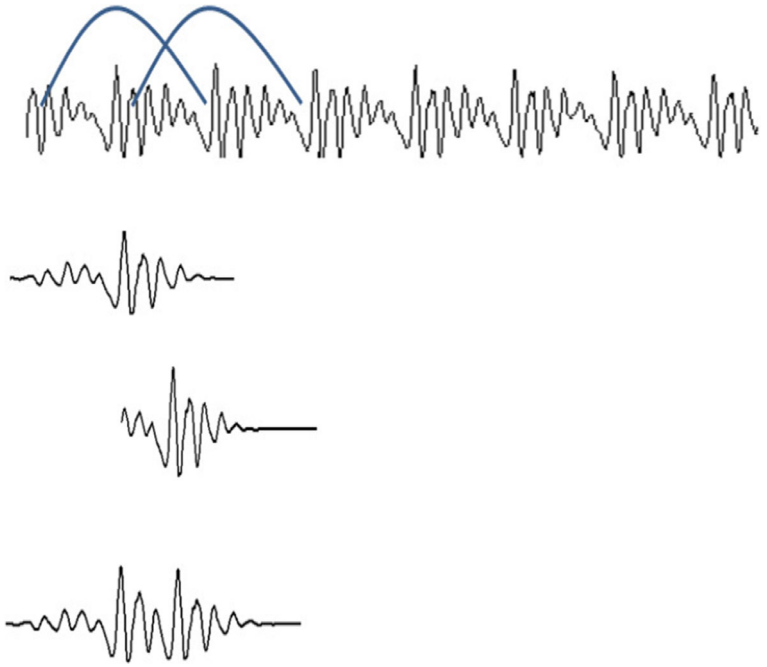

Fig. 9 Reducing the pitch period by a factor of 2 using TD-PSOLA

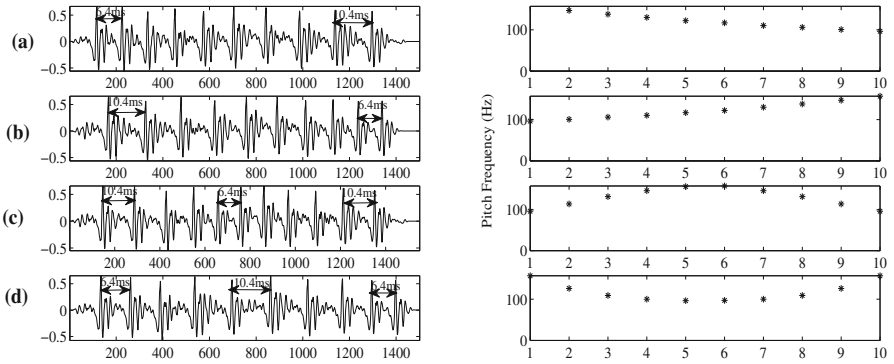

Fig. 10 Segment of the vowel /a/ bearing different pitch contours **a** Falling **b** Rising **c** Hat-shaped **d** Bucket-shaped

4.4 Energy Modification

The short-time energy represents the amplitude variations in the speech signal. Since energy is proportional to the magnitude of the speech signal, to increase or decrease the energy of the speech signal, the amplitude is increased or decreased respectively. Figure 12 shows the sentence, "I feel great" spoken by a female speaker, before and after energy modification. Figure 12b, d show the short-term energy of the two utterances.

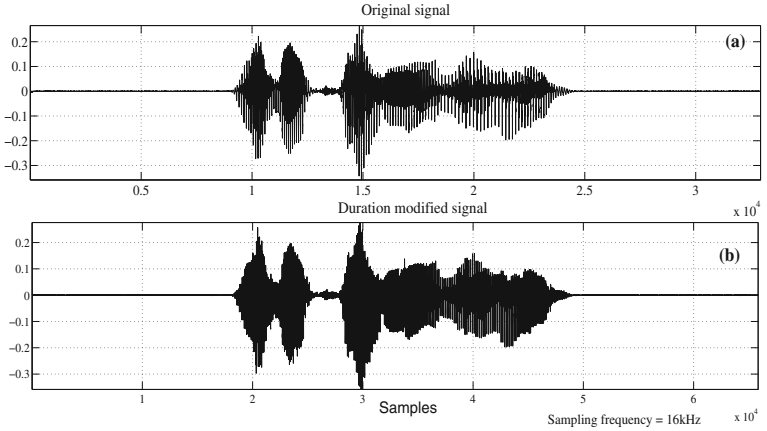

Fig. 11 Duration modification **a** Original speech signal **b** Speech signal with duration increased by a factor of 2

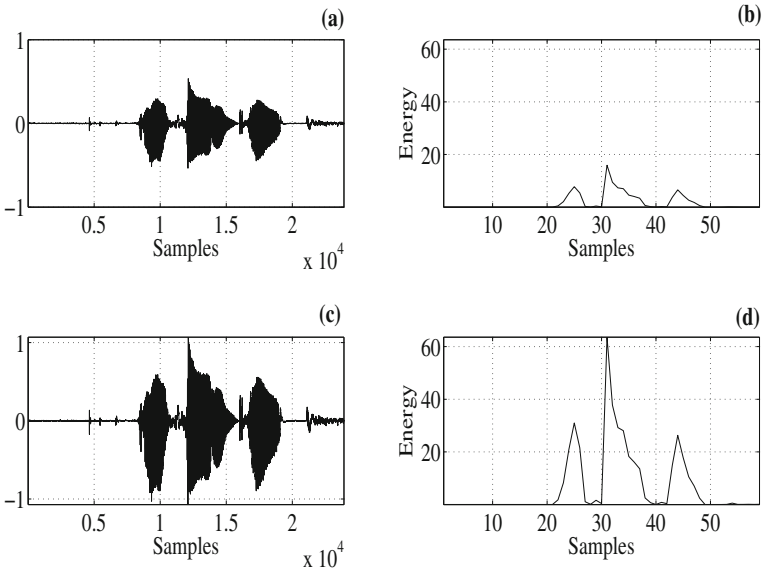

Fig. 12 Energy modification **a** Original neutral speech signal **b** Energy of the neutral speech signal **c** Modified speech signal **d** Energy of the modified speech signal

5 HMM-Based Keyword Spotting

As discussed in Sect. 3, it is clear that the happiness in speech is primarily reflected in the pitch contour, speech rate, and energy, specifically in the emotive-keywords. This necessitates the detection of keywords from a given speech utterance.

Initially keyword spotting is performed using manually derived word-level segmentation. Each word in an utterance is compared with a dictionary of emotive-keywords.

The dictionary consists of 178 emotive-keywords, out of which 63 words are root words and the rest are derivatives of these words. The keywords words of a given neutral utterance are used for further processing.

In order to eliminate the need for manual segmentation, HMM-based keyword spotting is used, where a speech recognition system is trained with six minutes of data, consisting of the 178 English sentences from our own database. 39-dimensional Mel frequency cepstral coefficients (MFCC) are extracted from the training data and context-independent phoneme models are trained with three states and number of mixture components per state based on the occurrence of each phoneme in the database. These models are used along with a word-level dictionary and network to recognize the given text and obtain the segmentation information. The keywords are then identified through comparison with the dictionary of emotive-keywords.

Although the current work employs a dictionary with 178 words to identify the emotive-keywords, it is observed that these words are predominantly nouns, adjectives, adverbs, and interjections. Therefore, this dictionary could be replaced by a part-of-speech tagger to ensure the identification of a much larger number of keywords. Once the emotive-keywords are identified, the time-domain parameters of these words are modified as described in Sect. 4.

6 Incorporation of Happiness in Neutral Speech

Happiness can be incorporated into neutral speech in one of two ways, namely,

- Replicating the variations in the parameters of natural happy speech in all the words of the corresponding neutral speech
- Modifying the parameters of the emotive-keywords, based on the analysis in Sect. 3

These two methods are elaborated below and their performance is compared in Sect. 7, to verify if modifying the keywords in a given neutral utterance would suffice to incorporate happiness.

6.1 Replicating Natural Happy Speech

Owing to the availability of parallel data for neutral and happy speech, the following technique is adapted to incorporate happiness in neutral speech of the SAVEE database. Initially, word-level boundaries are derived for the happy and neutral utterances. Word-by-word manipulation of the time-domain parameters is then carried out. Instants of excitation are derived using the PD algorithm. The ratio between the duration of happy speech and neutral speech is calculated. Based on this ratio, the duration of neutral speech is first modified as described in Sect. 4.2.2.

Now that the duration of neutral speech matches that of the corresponding happy speech, pitch contour modifications are performed using polynomial curve fitting. Polynomials of orders one to ten are fitted on to the pitch contour of happy speech. The polynomial that best captures the shape of the happy pitch contour is used to fit the new contour on to the neutral speech. In order to determine the best order, residues (error between the actual contour and the one obtained by fitting a polynomial) corresponding

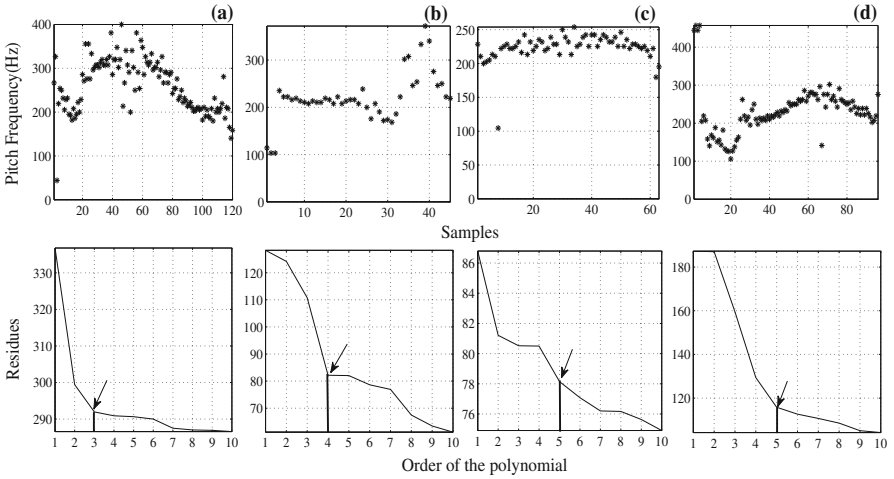

Fig. 13 Residues obtained when pitch contour of each word is modeled by polynomials of orders 1–10 **a** Hurray **b** Caught **c** Winning **d** Thank

to each polynomial are considered. The order beyond which the slope of the residues decreases is chosen. This is illustrated in Fig. 13. Consider Fig. 13a, which shows the residues obtained for different polynomials fitted on to the word "hurray". It is observed that the slope of the residues decreases beyond an order of three. Therefore, a polynomial of order three is used to modify the pitch contour of this word. The order of the polynomial to be used varies with the pitch contour of each word. This is shown in Fig. 13 for the words "caught" (order 4), "winning" (order 5), and "thank"(order 5).

Once the new pitch contour and hence new instants of excitation are derived, they are used to modify the neutral speech to bear the pitch contour of the corresponding happy speech, as described in Sect. 4.2.1. The intensity of the neutral speech is also modified to match that of the happy speech. Figure 14 shows the sentence "I feel great" spoken by a female speaker in neutral and happy tones, and also the happy speech synthesized in the method described above.

The three common and two emotion-specific sentences of the SAVEE database and 20 sentences from our own database are converted from neutral to happy, in the method discussed above. The synthesized happy speech sounds quite similar to the natural happy speech. However, such a technique would require parallel emotional and neutral data. Therefore, in line with the inferences derived in Sect. 3, the following section discusses the modification of only the emotive-keywords to incorporate happiness into neutral speech.

6.2 Introducing Variations in Emotive Words

The analysis in Sect. 3 revealed that emotive keywords bear a hat-shaped pitch contour, lower speech rate, and raised energy. To verify if these modifications would suffice to incorporate happiness in neutral speech, all possible variations of the time domain

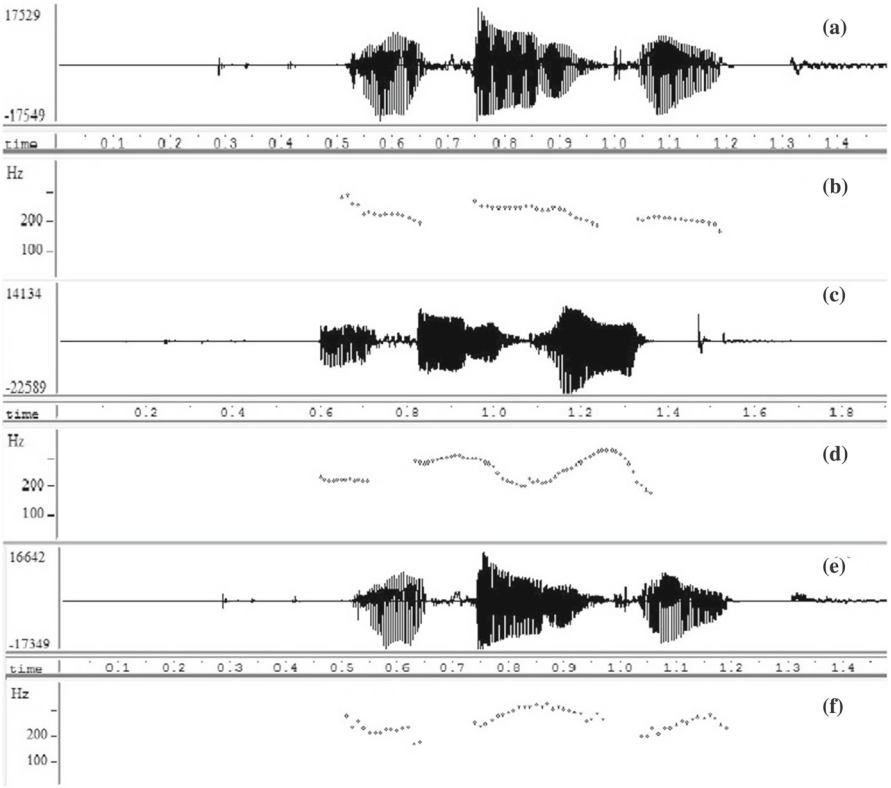

Fig. 14 Replicating natural happy speech **a** Neutral speech **b** Pitch contour of neutral speech **c** Happy speech **d** Pitch contour of happy speech **e** Synthesized happy speech **f** Pitch contour of synthesized happy speech

parameters are incorporated into the keywords and the combination that best portrays happiness is identified.

Figure 15 shows the variations imposed on the time-domain parameters, in the current work. 29 different combinations of these variations are incorporated in neutral speech. Some of these combinations are as follows, two of which are portrayed in Fig. 16, for the sentence, "I feel good now", from our database.

- Rising pitch contour, with minimum and maximum pitch periods equal to ±50% of the average pitch period, and twice the original intensity and duration
- Hat-shaped pitch contour, with minimum and maximum pitch periods equal to ±60% of the average, and twice the original duration
- Hat-shaped pitch contour, with minimum and maximum pitch periods equal to ±30% of the average, and 1.5 times the original intensity
- Half of the original duration and twice the original intensity
- Twice the original duration and intensity

It is observed that happiness is best portrayed by doubling the intensity of the keyword and fitting a hat-shaped pitch contour with minimum and maximum pitch

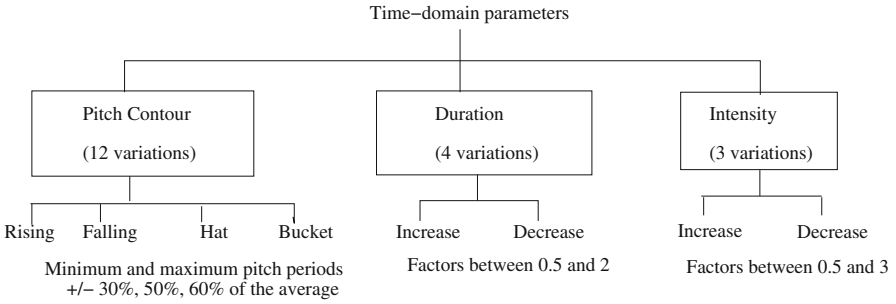

Fig. 15 Variations in time-domain parameters

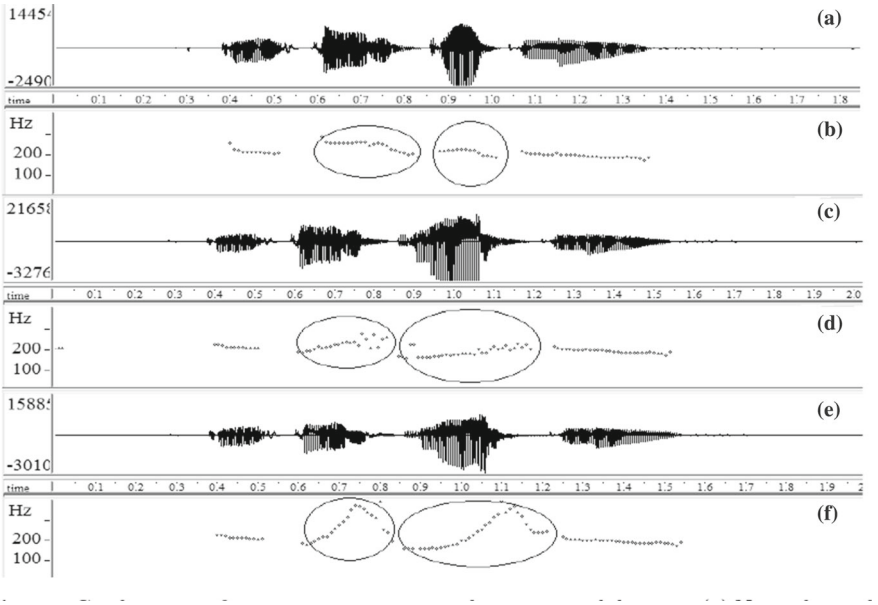

Fig. 16 Combination of variations in energy, pitch contour and duration **a** Neutral speech signal **b** Pitch contour of neutral speech signal **c** Modified speech signal with rising contour for keyword and twice the duration and intensity **d** Pitch contour of modified speech **e** Modified speech signal with hat-shaped contour for keyword and twice the duration **f** Pitch contour of modified speech

periods being ±60% of the average pitch period, using a polynomial of order 2. Modifying the speech rate of the keyword did not have a prominent impact on the perception of happiness.

From the above analysis, happiness is incorporated in the neutral speech by spotting the keywords in the speech utterance using HMM-based keyword spotting, discussed in Sect. 5, and doubling the intensity and fitting a hat-shaped contour as described in Sect. 4.

It is observed that increasing the intensity of the keyword alone, the abrupt change in energy causes the synthesized speech to sound slightly annoying. Therefore, to improve the quality of the synthesized happy speech, the energy is also varied gradually

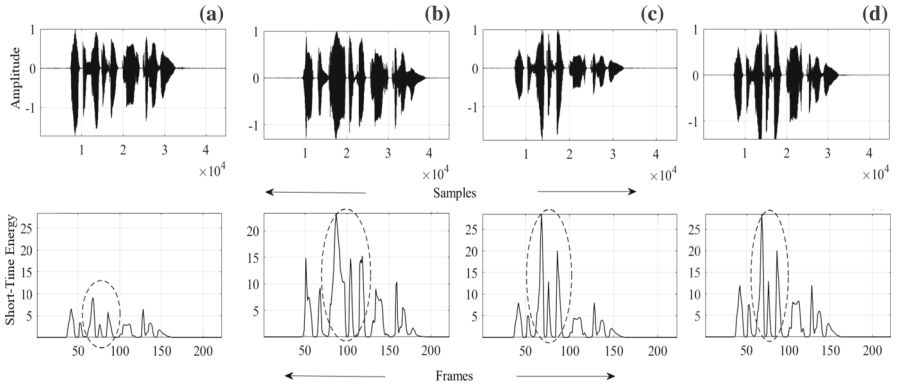

Fig. 17 Gradual increase in energy throughout the sentence **a** Neutral speech **b** Natural happy speech **c** Synthesized happy speech with energy of keywords abruptly modified **d** Synthesized happy speech with energy of keywords gradually modified

Table 1 MOS obtained for LP and TD-PSOLA-based synthesis speech

Synthesis technique	MOS (out of 3)
LP-based synthesis (Modifying only keywords)	2.02
TD-PSOLA-based synthesis (Modifying only keywords)	2.51
TD-PSOLA-based synthesis (Modifying all the words)	2.34

in the adjacent words. This is portrayed in Fig. 17. The sentence, "What a beautiful place to live in!", is considered. The encircled region in the figure, corresponds to the emotive-keyword, "beautiful". Figure 17 shows the energy plot when amplitude of the keywords is abruptly increased by a factor of 2, and when amplitude of the adjacent keywords are also increased gradually.

7 Performance Analysis

The quality of the synthesized speech is assessed subjectively using the mean opinion score and degradation mean opinion score and objectively using a GMM-based emotion identification system. The language-dependence/independence of the proposed method is also analyzed.

7.1 Mean Opinion Score

100 sentences (50 English and 50 Tamil) of our own database, synthesized by the LP-based approach and TD-PSOLA were distributed among 50 naïve listeners. The listeners were asked to rate the extent to which happiness is perceived on a three-point grading scale. A score of 1 implies that happiness is imperceivable and a score of 3 implies that happiness is clearly perceived in the synthesized speech. The mean

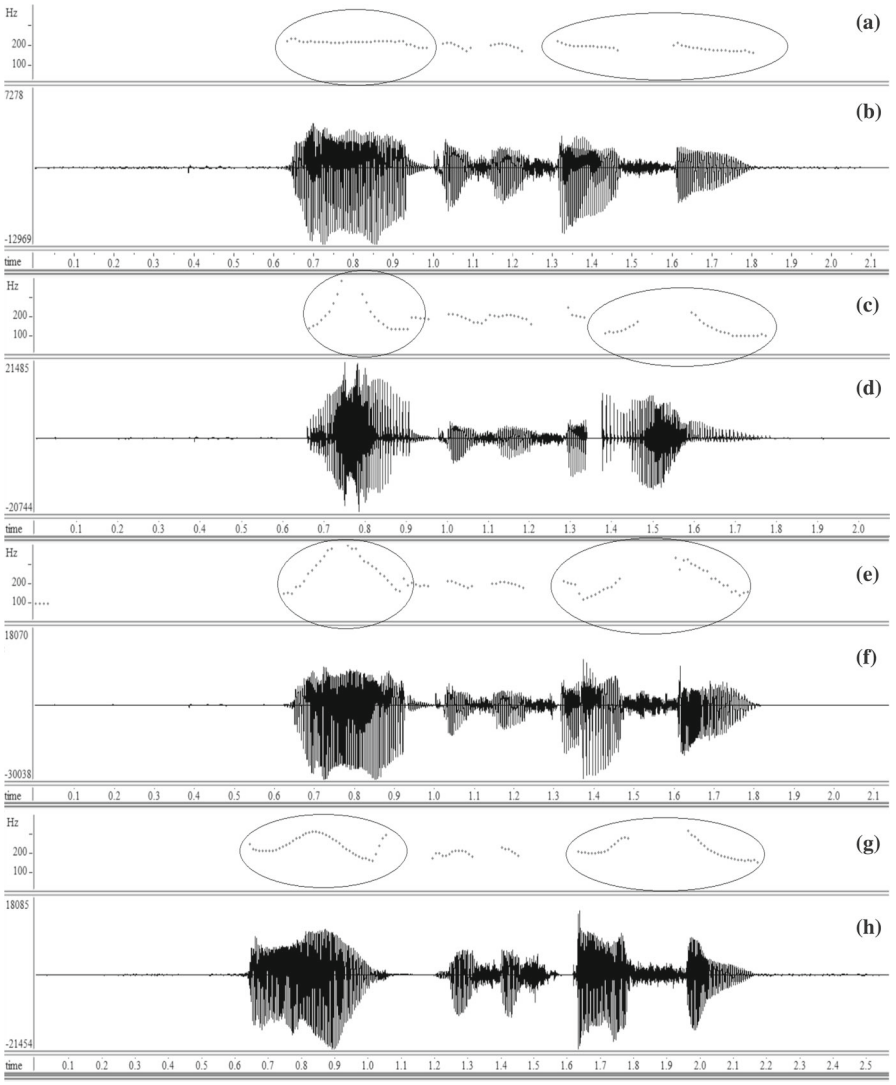

Fig. 18 Comparison of synthesized happy speech **a** Pitch contour of neutral speech **b** Neutral speech signal **c** Pitch contour of happy speech synthesized by the LP-based method **d** Happy speech synthesized by the LP-based method **e** Pitch contour of happy speech synthesized by using TD-PSOLA **f** Happy speech synthesized by using TD-PSOLA **g** Pitch contour of natural happy speech **h** Natural happy speech signal

opinion score (MOS) is obtained by taking the average of the scores assigned for each technique, by the listeners. The MOS obtained for each technique is shown in Table 1.

From the MOS, it is clear that in both techniques, happiness is perceived. The higher score obtained for TD-PSOLA can be attributed to the synthesized speech being more natural. This is because, modifications are directly imposed on the speech waveform in TD-PSOLA, whereas, in LP-based synthesis, speech is decomposed into

Table 2 DMOS obtained for LP and TD-PSOLA-based synthesis speech

Synthesis technique	DMOS (out of 5)
LP-based synthesis (Modifying only keywords)	2.88
TD-PSOLA-based synthesis (Modifying only keywords)	3.96
TD-PSOLA-based synthesis (Modifying all the words)	3.66

the source and system components, and then modified. Further, the use of a train of impulses to model the source, also decreases the naturalness. Figure 18 shows happy speech synthesized by both techniques, for the sentence "Wow, this is awesome", spoken by the female speaker from our database, and it is clearly observed that LP-based synthesis produces speech that is degraded when compared to happy speech synthesized by TD-PSOLA.

In order to quantify the amount of degradation introduced by both techniques, a degradation mean opinion score (DMOS) has also been computed. The listening test to obtain the DMOS included the same 50 listeners and the 100 sentences considered to compute the MOS. The listeners were asked to listen to synthesized happy speech and the corresponding natural neutral speech and indicate the level of degradation perceivable in the former, on a scale on 1–5. A score of 1 indicates that the synthesized happy is speech is completely degraded and unpleasant to listen to and a score of 5 indicates that the synthesized speech sounds as good as the corresponding neutral speech. The DMOS obtained for each technique is tabulated in Table 2, which clearly reiterates the previous inferences.

Happy speech synthesized by modifying all the words in a sentence and by modifying only the keywords, are also analyzed. In this regard, 30 sentences[1] from our own database, synthesized by modifying all the words in the given neutral speech (using TD-PSOLA), are also scored by the same 50 listeners as before. An MOS of 2.34 is obtained, revealing that modifying only the keywords in a neutral utterance would suffice to incorporate happiness. While modifying all the words in the neutral utterance is expected to produce a better result, the lower score could be attributed to the larger amount of degradation in the happy speech synthesized this way. This is evident from Table 2, where the DMOS obtained when all the words are modified is 3.66, while that obtained when only keywords are modified is 3.96.

As mentioned previously, 50 Tamil utterances are also considered in the listening tests. This is primarily done to test if the analysis results that were derived from the English utterances and the proposed method of modifying only the emotive-keywords to incorporate happiness could be applied to other languages as well. The MOS obtained when only the 50 Tamil utterances are considered is 2.53, which is similar to the score obtained when the English utterances are considered as well. This reveals that the inferences from the analyses and the proposed method could possibly be extended to other languages as well.

[1] When modifying all the words in neutral speech to synthesize happy speech, the modifications are incorporated with reference to the corresponding happy speech. Therefore, since only 30 neutral utterances from our own database have a corresponding happy utterance, these have been considered in computing the MOS and DMOS for this method.

7.2 GMM-Based Emotion Recognition

To objectively, analyze the quality of the speech synthesized by modifying the keywords, using TD-PSOLA, a GMM-based emotion recognition system is developed. GMMs with 3 mixture components are trained for neutral and happy speech. Pitch contours of the keywords of happy and neutral speech are modelled using fourth order polynomials. Further, energy is extracted frame-by-frame from the keywords. The polynomial coefficients and the average short-time energy derived, are used to train the GMMs. 100 synthesized happy and neutral utterances are tested using this system and a 93% recognition accuracy is obtained.

8 Incorporation of Happiness in Speech Synthesized by a TTS System

The proposed method of incorporating happiness in neutral speech can be easily employed in any TTS system as a post-processing module, since the method operates independently of the TTS algorithm. In order to test the feasibility of doing this, an HMM-based Indian English text-to-speech synthesis system is considered. The TTS system is trained on 5 hours of data recorded from a professional, native-Tamil female speaker in a studio environment, at a sampling rate of 48 kHz. This data is a subset of the speech corpus collected as a part of a TTS project funded by the Ministry of Electronics and Information Technology, Government of India and is available at https://www.iitm.ac.in/donlab/tts/index.php. Details on the development of an HMM-based TTS system are elaborated in [2] and the current work adopts a similar procedure. Once the HMM-based speech synthesis system (HTS) is developed, the proposed method is added to the system as a post-processing module. The steps involved in obtaining happy speech from this modified TTS system are as follows:

- Given a text, the corresponding neutral speech is first synthesized and its time-aligned phonetic transcription is obtained.
- From the time-aligned phonetic transcription, the time-aligned word-level transcription is obtained.
- The emotive-keywords are then identified using the Stanford part-of-speech tagger [22,23] (adjectives, adverbs, interjections and nouns are considered to be the emotive-keywords as discussed in Sect. 5).
- TD-PSOLA is then used to fit a hat-shaped contour on to the keyword, the energy of the keyword is doubled and that of the adjacent words are increased appropriately, thus resulting in the desired happy speech.

In order to test if a part-of-speech tagger suitably replaces the dictionary of keywords used in the previous analyses, 50 utterances that are synthesized by carrying out the above-mentioned steps are evaluated by 50 listeners. An MOS of 2.52 out of 3 and a DMOS of 3.8 out of 5 are obtained, which are close to those obtained in Sect. 7. This reveals that the proposed method could be easily incorporated with a TTS system and also that the dictionary of keywords could well be replaced by a part-of-speech tagger.

Although the current work incorporates the proposed method in an HMM-based TTS system, a similar procedure could be used with the recent DNN-based TTS systems as well.

9 Conclusion

Emotion is the state of mind of a person and is reflected in speech. Emotion in speech primarily affects the pitch period, formant frequencies, speech rate, and energy. In this regard, the afore-mentioned parameters are analyzed for happy speech. It is seen that variations in these parameters are predominantly observed in the emotive-keywords. The keywords possess a hat-shaped pitch contour, increased energy, and reduces speech rate, owing to the stress placed on them. To verify these inferences, different combinations of variations on the time-domain parameters are incorporated in the keywords of neutral speech and the quality of synthesized speech is analyzed. It is inferred that happiness is best perceived when the keyword is fitted with a hat-shaped contour, using a polynomial of order 2, and minimum and maximum pitch period equal to $\pm 60\%$ of the average, and its intensity is doubled. Happy speech is synthesized using a source-filter model (filter coefficients derived using linear prediction) and TD-PSOLA. The latter outperforms the former in terms of naturalness, as revealed by the MOS of 2.51 and the DMOS of 3.96. Further, instead of modifying only the keywords, all the words in a neutral sentence are modified to replicate the characteristics of the corresponding happy speech. An MOS of 2.34 obtained for the speech synthesized by this technique reveals that modifying the keywords would suffice to incorporate happiness. This is assessed objectively by a GMM-based emotion recognition system as well. Further, the feasibility of incorporating the proposed method as a post-processing module in a TTS system to generate happy speech and replacing the dictionary of emotive-keywords by a part-of-speech tagger is also demonstrated.

Availability of Data and Material The datasets used for analysis are open source and the links to obtain the same are as follows:
- Berlin database: https://www.kaggle.com/piyushagni5/berlin-database-of-emotional-speech-emodb
- SAVEE database: http://kahlan.eps.surrey.ac.uk/savee/Download.html
- TTS database: https://www.iitm.ac.in/donlab/tts/index.php

Declarations

Conflict of interest The authors declare that they have no competing interest.

References

1. K. Akuzawa, Y. Iwasawa, Y. Matsuo, Expressive speech synthesis via modeling expressions with variational autoencoder. INTERSPEECH, 3067–3071 (2018)
2. G. Anushiya Rachel, V. Sherlin Solomi, K. Naveenkumar, P. Vijayalakshmi, T. Nagarajan, A small-footprint context-independent HMM-based synthesizer for tamil. Int. J. Speech Technol. **18**, 405–418 (2015)

3. G. Anushiya Rachel, P. Vijayalakshmi, T. Nagarajan, Estimation of glottal closure instants from degraded speech using a phase-difference-based algorithm. Comput. Speech Lang. **46**, 136–153 (2017)
4. A.W. Black, Unit selection and emotional speech. Eurospeech, 1–4 (2003)
5. M. Bulut, S.S. Narayanan, A.K. Syrdal, Expressive speech synthesis using a concatenative synthesizer. in 7th International Conference on Spoken Language Processing, pp. 1265–1268 (2002)
6. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech. INTERSPEECH, 1–5 (2005)
7. L. Cen, P. Chan, M. Dong, H. Li, Generating emotional speech from neutral speech. in 7th International Symposium on Chinese Spoken Language Processing, (2010), pp. 383–386
8. D. Govind, S.R.M. Prasanna, Expressive speech synthesis using prosodic modification and dynamic time warping. Natl. Conf. Commun., 290–293 (2009)
9. S. Haq, P.J.B. Jackson, J.D. Edge, Audio-visual feature selection and reduction for emotion classification. in International Conference on Auditory-Visual Speech Processing 185–190 (2008)
10. G.O. Hofer, K. Richmond, R.A. Clark, Informed blending of databases for emotional speech synthesis. INTERSPEECH 501–504 (2005)
11. P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans Audio Speech Lang Process **15**, 34–43 (2007)
12. I.R. Murray, J.L. Arnott, Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Comput Speech Lang **22**, 107–129 (2008)
13. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process **16**, 1602–1613 (2008)
14. L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals* (Prentice Hall, 1978)
15. K.S. Rao, Unconstrained pitch contour modification using instants of significant excitation. Circuits Syst Signal Process **31**, 2133–2152 (2012)
16. K.S. Rao, S.R.M. Prasanna, B. Yegnanarayana, Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. IEEE Signal Process. Lett. **14**, 762–765 (2007)
17. K.S. Rao, A.K. Vuppala, Non-uniform time scale modification using instants of significant excitation and vowel onset points. Speech Commun. **55**, 745–756 (2013)
18. M. Schroder, *Expressive Speech Synthesis: Past Present, and Possible Futures. Affective Information Processing* (Springer, Berlin, 2009)
19. R. Smits, B. Yegnanarayana, Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. **3**, 325–333 (1995)
20. D. Talkin, *A Robust Algorithm for Pitch Tracking. RAPT* (Elsevier, Amsterdam, 1995)
21. J. Tao, Y. Kang, A. Li, Prosody conversion from neutral speech to emotional speech. IEEE Trans. Audio Speech Lang. Process. **14**, 1145–1154 (2006)
22. K. Toutanova, D. Klein, C. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network. HLT-NAACL **2003**, 252–259 (2003)
23. K. Toutanova, C. D. Manning, Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), 63–70 (2000)
24. H.K. Vydana, S.R. Kadiri, A.K. Vuppala, Vowel-based non-uniform prosody modification for emotion conversion. Circuits Syst. Signal Process. **35**, 1643–1663 (2016)
25. Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, R. A. Saurous, Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. https://arxiv.org/abs/1803.09017 (2018)
26. C.-H. Wu, C.-C. Hsia, C.-H. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. IEEE Trans. Audio Speech Lang. Process. **18**, 1394–1405 (2010)
27. J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, Acoustic modelling of speaking styles and emotion expressions in HMM-based speech synthesis. IEICE Trans. Inf. Syst. **E88–D**, 502–509 (2005)
28. C.-Y. Yang, C.-P. Chen, A hidden Markov model-based approach for emotional speech synthesis. in 7th ISCA Workshop on Speech Synthesis, pp. 126–129 (2010)
29. S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, An Acoustic Study of Emotions Expressed in Speech. INTERSPEECH 2193–2196 (2004)
30. H. Zhang, Y. Yang, Fundamental frequency adjustment and formant transition-based emotional speech synthesis. in 9th International Conference on Fuzzy Systems and Knowledge Discovery, (2012), pp. 1797–1801