Int J Speech Technol
DOI 10.1007/s10772-012-9180-2

Expressive speech synthesis: a review

D. Govind · S.R. Mahadeva Prasanna

Received: 31 May 2012 / Accepted: 11 October 2012

Abstract The objective of the present work is to provide a detailed review of expressive speech synthesis (ESS). Among various approaches for ESS, the present paper focuses the development of ESS systems by explicit control. In this approach, the ESS is achieved by modifying the parameters of the neutral speech which is synthesized from the text. The present paper reviews the works addressing various issues related to the development of ESS systems by explicit control. The review provided in this paper include, review of the various approaches for text to speech synthesis, various studies on the analysis and estimation of expressive parameters and various studies on methods to incorporate expressive parameters. Finally the review is concluded by mentioning the scope of future work for ESS by explicit control.

Keywords Expressions · Prosody modification · Text to speech · Emotion conversion · Epochs

1 Introduction

Speech synthesis is the process of converting message written in text to equivalent message in spoken form. Expressive speech synthesis deals with synthesizing speech and adding various expressions related to different emotions and speaking styles to the synthesized speech (Pitrelli et al. 2006; Tao et al. 2006; Erickson 2005; Campell et al. 2006). The dictionary meaning of expression is conveying a thought or an emotion. The expression is defined as the vocal indicator of various emotional states that reflect in the speech waveform (Scherer 1986). The different emotions and speaking styles are also considered as expressions (Pitrelli et al. 2006). Based on this, in the present work, we have considered different emotions as the expressions and hence emotions and expressions are interchangeably used.

The objective of speech synthesis is to synthesize speech waveform from the text. The Schematic block diagram of a speech synthesis system is shown in Fig. 1. The input text is first converted into abstract linguistic representation by the front end text processing stage. This linguistic representation is obtained by performing prosodic annotations on the syntactic, semantic and lexically analyzed text (Klatt 1987). This linguistic representation drives the synthesis routines to get the speech waveform of the input text (Klatt 1980). In the present work, such a system is termed as Neutral Speech Synthesis (NSS) system.

In expressive speech synthesis, along with text, the desired expression also forms an additional input to the text processing stage as shown in Fig. 2. The input text is converted into abstract linguistic representation as in NSS. In addition, the expressive information is also incorporated, either before or after the synthesis of neutral speech. In the former case, the expressive information is coded along with the linguistic information and speech is synthesized from the text using the linguistic and expressive information. In the later case, the speech is synthesized initially without any expression, that is, neutral speech and then later the desired expression is added using a suitable voice transformation technique (Tao et al. 2006; Campell et al. 2006).

D. Govind (✉)
Center for Excellence in Computational Engineering & Networking, Amrita Vishwa Vidyapeetham (University), Coimbatore 641 105, India
e-mail: dgovind@iitg.ernet.in

S.R.M. Prasanna
Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India
e-mail: prasanna@iitg.ernet.in

Published online: 24 October 2012

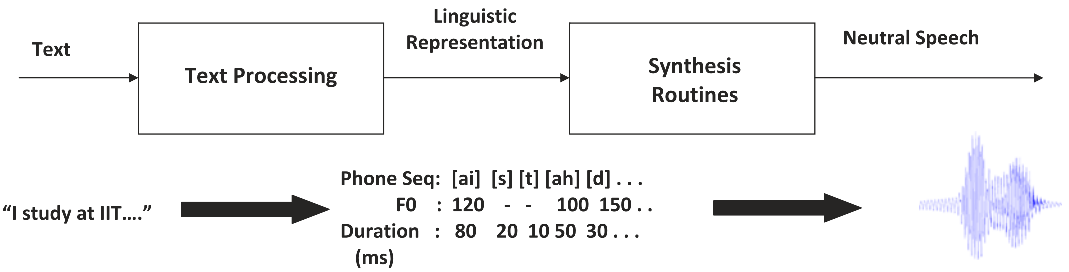
Text
Text Processing
Linguistic Representation
Synthesis Routines
Neutral Speech
"I study at IIT...."
Phone Seq: [ai] [s] [t] [ah] [d] . . .
F0 : 120 - - 100 150 . .
Duration (ms) : 80 20 10 50 30 . . .

Fig. 1 Schematic diagram of neutral speech synthesis

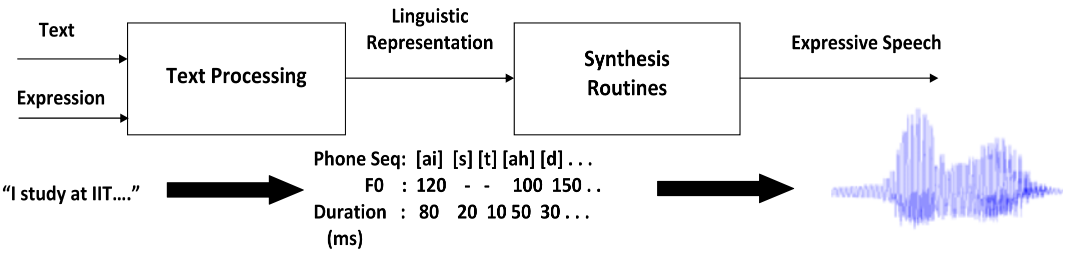

Fig. 2 Schematic diagram of expressive speech synthesis

Speech synthesized in different expressions can be used in story telling applications for children where for effectiveness and drawing attention, different expressions have to be generated in different contexts of the story (Theune et al. 2006). ESS can be used as a part of dialogue system which makes the human computer interaction more natural and effective. Expressive speech analysis can be utilized by the call center managers to identify the emotional state of the operators during conversing with the customers and valuate them based on their emotional maturity. ESS finds application in the financial information system to make announcements in different speaking styles to the users (Pitrelli et al. 2006).

The review of existing approaches for the development of ESS systems are presented in Sect. 2. Among various approaches for the development of ESS systems, the present focus of the development of ESS by explicit control and the issues in that are described in Sect. 3. In ESS by explicit control approach, the perceptual quality of the synthesized expressive speech deeply depended on the quality of the synthesized neutral speech, Sect. 4 reviews different NSS approaches.The various works done on analysis and estimation of expressive parameters are explained in Sect. 5. Section 6 reviews the methods to incorporate the expressive parameters. Section 7 describes the methods to evaluate the synthesized expressive speech. Finally, the scope for the present work is given in Sect. 8

2 Review of existing expressive speech synthesis systems

This section reviews various existing approaches employed for expressive speech synthesis. According to Schroder (2009), the expressive speech synthesis approaches can be broadly classified into the following three categories.

- Expressive speech synthesis by explicit control
- Expressive speech synthesis by playback approach
- Expressive speech synthesis by implicit control

2.1 Expressive speech synthesis by explicit control

Here, the expressive speech is synthesized by modifying the neutrally synthesized speech based on the prosodic rules derived from the expressive speech database of the respective expressions. The expressive speech synthesis systems developed on formant synthesis (Cahn 1989; Murray and Arnott 1993; Burkhardt and Sendilmeier 2000) and diphone concatenation are examples of explicit control (Vroomen et al. 1993; Montero et al. 1999). Various methods developed for neutral to expressive conversion tasks (Tao et al. 2006; Cabral and Oliveira 2006a; Theune et al. 2006) also falls under the category of explicit control.

As the formant speech synthesizers provide flexibility to control the various source and system parameters, earlier developments of expressive speech synthesis systems were on top of the formant speech synthesis systems (Schröder 2001). The affect editor developed by Cahn was the first attempt to synthesize emotional speech using a formant synthesizer (Cahn 1989; Schröder 2001). The control parameters of formant synthesizer are manually tuned for each of the emotions to synthesize the expressive speech. Modification of each of the control parameters for each emotion is done according to various acoustic profiles discussed in the literature (Williams and Stevens 1972; Fairbanks and

Hoaglin 1939). HAMLET, emotion speech synthesis system developed by Murray and Arnott (1995), is a rule based system developed on commercial formant speech synthesis system called DECtalk. In HAMLET the pitch and duration rules and voice quality rules are set in the formant synthesizer and quality of the synthesized emotions are improved heuristically by manual tuning. The development of these rules for emotions is as given in Murray and Arnott (1993). The objective of the perceptual experiments conducted by Felix Burkhardt was to, find out the perceptually relevant acoustic features for each emotion by systematically varying these acoustic parameters during the synthesis of the neutral utterances and find the optimum values of each of the acoustic features for the emotional speech synthesis (Burkhardt and Sendilmeier 2000). According to these perceptual experiments, the pitch parameters like mean pitch and pitch range, suprasegmental parameters like speech rate, and voice quality parameters like phonation and vowel precisions, are found to be significant for effectively synthesizing emotions using formant synthesizers. The studies conducted by Vroomen et al. (1993) on seven emotions (neutral, joy, boredom, anger, sadness, fear, indignation) showed that only intonation and duration are enough to express emotions in the synthesized speech using a diphone synthesizer. Here the emotional speech is synthesized by manipulating the pitch and duration using Pitch Synchronous Overlap Add (PSOLA) of the neutrally synthesized speech. The significance of pitch and duration parameters in emotional speech synthesis is also shown in the studies by Montero et al. (1999) in Spanish using diphone concatenation. This study also showed that the relative contribution of prosody and voice quality depends on the emotions to be synthesized (Schroder 2009).

Apart from the expressive speech synthesis systems developed based on various speech synthesis approaches, there are some works done for neutral to target expressive speech conversion task using the explicit control approach. Tao et al. (2006) achieved expressive speech conversion by prosody (pitch and duration) modification of the neutral expressive speech. This paper compared linear, Gaussian mixture model (GMM) and classification and regression tree (CART) methods for converting neutral speech to target expressive speech for mandarin language. Apart from discrete emotions like angry, happy, sad and fear, the strong, medium and weak versions of each is also considered for synthesis. Direct scaling of sentence F_0 and syllable duration is done in linear modification model and other acoustic features of F_0 contour considered for modification are $F_{0topline}$, $F_{0baseline}$, F_{0avg} and intensity. In GMM based prosody modification, pitch target models are constructed from the tonal representation of the intonation pattern of each syllable for each expression. The pitch target model parameters generated by GMM of the neutral syllable is mapped to that of the target expression to obtain intonation contour. In the case of CART, along with prosody information of target expression, linguistic information obtained from the text is also used to build trees. Listening test indicates that the speech synthesized using GMM (for small data set) and CART (large data set) sounds more expressive compared to linear prosody modification. Cabral and Oliveira (2006a) developed Emo Voice system to incorporate different emotions into the neutral expressive speech in German language. In Emo Voice system the neutral speech is converted to expressive speech by modifying both prosody parameters (pitch, duration and intensity) and excitation source parameters (jitter, shimmer, and glottal wave parameters) by Pitch Synchronous Time Scaling (PSTS) method (Cabral and Oliveira 2006b; Cabral 2006). The rules for the prosody and voice quality modification are derived based on the acoustic profiles presented in Drioli et al. (2003), Whiteside (1998), Zovato et al. (2004). Theune et al. (2006) devised prosodic rules to generate expressions in the story telling style. Story telling expressions are synthesized by modifying the pitch and intensity of various part of the story like suspense, climax etc.

2.2 Expressive speech synthesis by playback approach

In playback approach, the expressive speech is synthesized independently using the respective expressive speech database. Here expressive speech synthesis is achieved either by merely playing back what is available in the database of the target expression or using the models which are trained using the target expression database. The unit selection based and HMM based expressive speech systems trained on the respective expressive database works on play back approach (Yamagishi et al. 2003; Iida et al. 2000; Hofer et al. 2005; Fernandez and Ramabhadran 2007; Pitrelli et al. 2006).

For improved naturalness in the synthesized speech, the emotional speech synthesis systems developed based on unit selection cocatenation were developed. A highly natural synthesized emotional speech is demonstrated by Iida et al. (2000) by storing large databases for each emotion. For synthesizing the target emotion, the respective emotion database is loaded and selected units from the database to synthesize the speech in the target emotion. A good quality conversational speech is synthesized by Campbell (2006) using phrase unit selection based speech synthesizer from a very large database. Gregar et al. used a blended database by mixing emotion databases of angry, happy and neutral speech for synthesizing speech in the target emotions (Hofer et al. 2005). For achieving this, target cost function is designed to give more penalty to select the units other than the intended emotion. The work done by Fernandez and Ramabhadran (2007) also followed same approach by mixing the units of other emotions to synthesize the target emotional speech. Pitrelli et al. (2006) proposed an unified approach for expressive speech synthesis system by combining corpus driven and prosodic phonology approach. These studies concluded from the listening test that the use of corpus driven approach for conveying good and bad news, and prosodic phonology approach for contrastive emphasis and Yes-No questions. Similar to unit selection approach expressive speech synthesis systems are developed using statistical parametric (HMM) approach also. Junichi et al. trained HMM models for different speaking styles like reading, sad, joyful and rough and synthesized speech in the target styles using the respective trained HMM models (Yamagishi et al. 2003). Some of the synthesized expressive speech samples for happy and angry emotions are available for listen at the following link: http://www.iitg.ernet.in/stud/dgovind/emotionsynthesis.htm.

2.3 Expressive speech synthesis by implicit approach

The implicit control based expressive speech system controls the expressivity by interpolation between two statistical models trained on the different expressive databases. The expressive speech synthesized by the interpolation and adaptation of HMM models are examples of implicit control. HMM based speech synthesizers offer various adaptation techniques to adapt the average style model to a specific style. Miyanaga et al. (2004) proposed an HMM based style synthesis system using a style control vector estimated for each style. During the synthesis the style control vector associated with the target style transforms the mean vectors of the neutral HMM models. The adaptation techniques provides flexibility to build the statistical models with a few minutes of data if an average model is available. As the speech synthesized using speaker adaptation are found to be more robust than speaker dependent case, these adaptation techniques can be used for synthesizing various styles also (Yamagishi et al. 2007). Apart from the adaptation techniques, HMM speech synthesis systems provide flexibility to synthesize various speaking styles or emotions by HMM interpolation or multiple regression of emotion vectors (Barra-Chicote et al. 2010; Tachibana et al. 2005; Nose et al. 2007). In spite of all these advantages for HMM based speech synthesis systems the notable disadvantage is the inherent over-smoothing of the spectral and excitation parameters by the HMM models (Barra-Chicote et al. 2010). This over-smoothing causes the reduced naturalness in the synthesized emotions. However, the perceptual studies presented by Barra-Chicote et al. shows that the emotional speech synthesized using HMM based speech synthesis system and unit selection based speech synthesis system provides almost similar emotion identification rates (Barra-Chicote et al. 2010).

The present work focuses on the development of expressive speech synthesis systems based on explicit control of prosodic features. Here the issue will be framing of prosodic rules by the analysis of each expression in the database and incorporating them into the neutral speech.

3 Issues in expressive speech synthesis by explicit control

The ESS by explicit control is achieved by transforming the neutral speech by a signal processing approach according to the prosodic rules framed for the target expression. The various issues in the ESS by explicit control approach are the following:

- Synthesizing a good quality neutral speech
- Analysis and estimation of expressive parameters
- Incorporation of expressive parameters

The various issues and approaches for the development of neutral speech synthesizers are presented in Sect. 4. Based on this review, the speech synthesized either from a unit selection concatenative system or HMM based statistical parametric speech synthesis system is of good intelligibility and reasonably natural. Therefore any of the two systems can be used as the neutral speech synthesizer for the present work.

The analysis and estimation of expression specific parameters of various emotions are performed on an expressive database. Section 5 reviews various existing expressive databases used for the expressive speech analysis. Expression specific parameters for each expression are analyzed with respect to the neutral expression. In this stage, the issues will be the accurate estimation of parameters to analyze its variation across various expressions. Therefore the choice of signal processing tools that accurately estimates the expressive parameters are important for analyzing the expressive parameters. Section 5 also reviews studies made on various expressive parameters in expressive speech analysis. Finally, the outcome of this study will be a set of rules on expressive parameters which can modify the parameters of neutral speech to synthesize the expressive speech.

The final stage in the ESS by explicit control is the incorporation of the rules for each expression on the parameters of neutral speech to obtain the speech in the target expression. This is typically achieved by a signal processing method. The issue in incorporating these expressive rules is to introduce minimum perceptual distortion without affecting the naturalness in the synthesized expressions. Section 6 reviews various methods for incorporating expressive parameters.

4 Review of text to speech synthesis

The front end text to speech synthesis system serves as the NSS. The parameters of the neutral speech synthesized by the TTS system are modified according to target expression to generate the speech in the target expression. Every TTS has a front end text processing block, which converts the text to be synthesized to an abstract linguistic specifications. These abstract linguistic specifications could be a sequence of phonemes or any sub-word unit and also it could be annotated with the prosodic information (Klatt 1987; Clark et al. 2007; King 2011). The text processing stage generally includes the text normalization, phrasification and lexical analysis modules. The role of the text processing module is to provide a unique contextual description about the sound units across the entire utterance. This abstract linguistic representation drives the waveform generation module to synthesize the speech in accordance with the text given. For the waveform generation from the abstract linguistic representation, there are four approaches namely,

- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis
- Statistical parametric speech synthesis

4.1 Articulatory speech synthesis

The objective of articulatory speech synthesis is to model various articulatory processes involved during speech production and use that knowledge to synthesize good quality speech sounds. Various stages in the development of articulatory speech synthesizers are the following:

- Articulatory data acquisition
- Geometric modeling of vocal tract
- Acoustic modeling for the synthesis

In articulatory data acquisition, the positions of various articulators are studied during the production of various sound units. In articulatory data acquisition, snapshots of the speech production organs are taken during the production of various sound units. The articulatory data is acquired using various sensors like fiberscope (Gauffin and Sundberge 1978), x-ray (Mermelstein 1973; Maeda 1979; Beautemps et al. 2001) and Magnetic Resonance Imaging (MRI) (Narayanan et al. 1995). The popular commercially available devices for measuring articulation are developed using electro magnetic articulography (EMA) and electro palatography (EPG). After acquiring the articulatory data, geometric models are built for the vocal tract which is in turn used for the acoustic synthesis of the sound units. There are a number of 2D and 3D models proposed in the literature for the accurate geometric modeling of vocal tract system using the available articulatory data (Palo 2006; Heinz and Stevens 1964; Engwall 1999). After generating a complete geometrical model from the articulatory data for various sound units, these parameters in terms of the area functions have to be mapped into acoustic parameters for the speech synthesis which is the final stage of the articulatory speech synthesis. The source filter theory of speech production proposed by Fant (1960) is the basis for the speech synthesis from the acoustic parameters. The VT tube acoustics is obtained by solving Websters horn equations for the sound pressure. The area functions of the geometrical models are mapped to simple 2D circular cross-sectional areas for applying Websters Horn equations. Once the acoustic parameters of VT tube is obtained, the electrical analogue circuit can be designed for the synthesis (Dunn 1950). The acoustic parameters can also be simulated by estimating the digital filter coefficients (Kelly and Lochbaum 1962; Badin and Fant 1984). The source information (pitch and intensity) required for the acoustic synthesis can be directly computed for each sound unit from the recorded data. Palo (2006) provides a detailed review of articulatory speech synthesis in his MSc Thesis.

Even though articulatory synthesis is based on physical theory, construction of geometrical models and their mapping makes it computationally complex. Also improper coarticulation modeling causes degradations in the synthesized speech. Computational complexity and reduced naturalness makes the articulatory speech synthesis approach less popular compared to other existing approaches for speech synthesis.

4.2 Formant speech synthesis

Formant speech synthesizers are example of the speech synthesizers using a speech production model. The formant speech synthesizer is developed based on source filter theory of speech production (Fant 1960). From a historic perspective Dudley's (1939) channel vocoder developed in 1939 is a primitive form of formant synthesizer. In Dudley (1939), the distribution of the formant energies and voicing are adjusted by an expert human to synthesize speech like waveform. Formant speech synthesis involves the simulation of formant frequencies, formant amplitudes and glottal source characteristics for each sound unit. The vocal tract is simulated using a set of resonators connected in cascade or parallel. The popular technique for the formant synthesis is developed by Klatt (1980). The parameters corresponds to formants and voicing source are tuned manually for synthesizing a good quality speech. After the development of Klatt formant synthesizer, Fant and Liljencrants came up with an improved parametric glottal model to provide a better shape for the glottal waves used in the Klatt synthesizer (Fant et al. 1985). As the formant synthesizers provide flexibility to vary the voice qualities of the synthesized speech by varying the control parameters of the source and the system, formant synthesizers are used in emotive speech synthesis applications (Cahn 1989). Recent development in the formant speech synthesis is the data driven formant synthesis (Carlson et al. 2002). Here formant parameters stored in the units library are selected and set as the control parameters for the formant synthesizer.

Even though formant synthesizers provide flexibility for varying voice qualities in the synthesized speech, increases complexity due to large number of control parameters. This necessity of setting the control parameters for speech synthesizer increases the time required to build speech synthesizer with good intelligibility and improved naturalness. Even though formant synthesized speech is observed intelligible, but sounds unnatural which is its main drawback.

4.3 Concatenative speech synthesis

The basic idea in concatenative speech synthesis is synthesis by joining the segments of the natural speech waveform that are stored in the database (Hunt and Black 1996; Clark et al. 2007). These segments can be words, subword units like phonemes, diphones and syllables. The widely used concatenative speech synthesis works on the principle of *unit selection*. The popular unit selection speech synthesis systems are *unisyn*, *clunits* and *multisyn* (Olive 1977; Black and Campbell 1995; Hunt and Black 1996; Clark et al. 2007). These systems differ to each other in terms of the type of unit, database and unit selection criteria used for synthesizing the speech. The unisyn unit selection system uses diphones as the basic units for concatenation. A diphone is defined from the stable middle region of one phone to the stable middle region of another phone. Unisyn attempts for building a diphone synthesizer by storing fixed size diphone units obtained from natural recordings stored in the repository (Olive 1977; Courbon and Emerald 1982). However, due to the availability of only single example of diphone units representing all the phonetic contexts, causes unnaturalness in the synthesized speech. To overcome this problem, *clunits* speech synthesis system is proposed in Black and Campbell (1995), Hunt and Black (1996). In clunits, the basic units of concatenation are mono phones. Here a large phonetically labeled database of 4–5 hours of continuous speech is used as unit inventory. The similarly sounding phonemes of different phonetic contexts in the entire database are clustered for the same phoneme class. During the synthesis, according to the phonetic context, the appropriate cluster of the same phone class are picked from clustered database and optimum units are selected for the reduced temporal and spectral discontinuities (join cost) at the concatenation points by the efficient Viterbi search algorithm. In multisyn based speech synthesis systems, use diphones as the basic unit of concatenation. Here the diphones required for the concatenation are selected from a large diphone labeled database of 4–5 hours of continuous speech recordings. The two cost functions are used for the optimum unit selection by the viterbi search are join cost and the target cost functions. The join cost computes the penalty cost in terms of the spectral and temporal distortion while joining two units and target cost gives the penalty cost of the unit with respect to the target diphone context. Figure 3 shows the example of unit selection for the text "two". For each target unit, the example candidate units are listed from the database. The variable size of the candidate units shown in Fig. 3 indicate that the size of the listed candidate units from the database

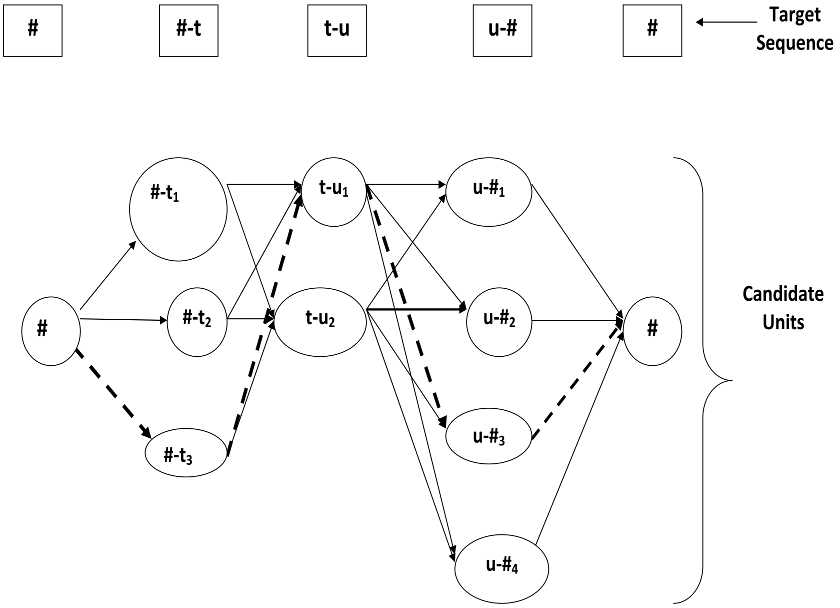

Fig. 3 Unit selection in concatenative speech synthesis system: the *bold-dotted lines* indicate the optimum path of the diphone units to be concatenated for the text "two"

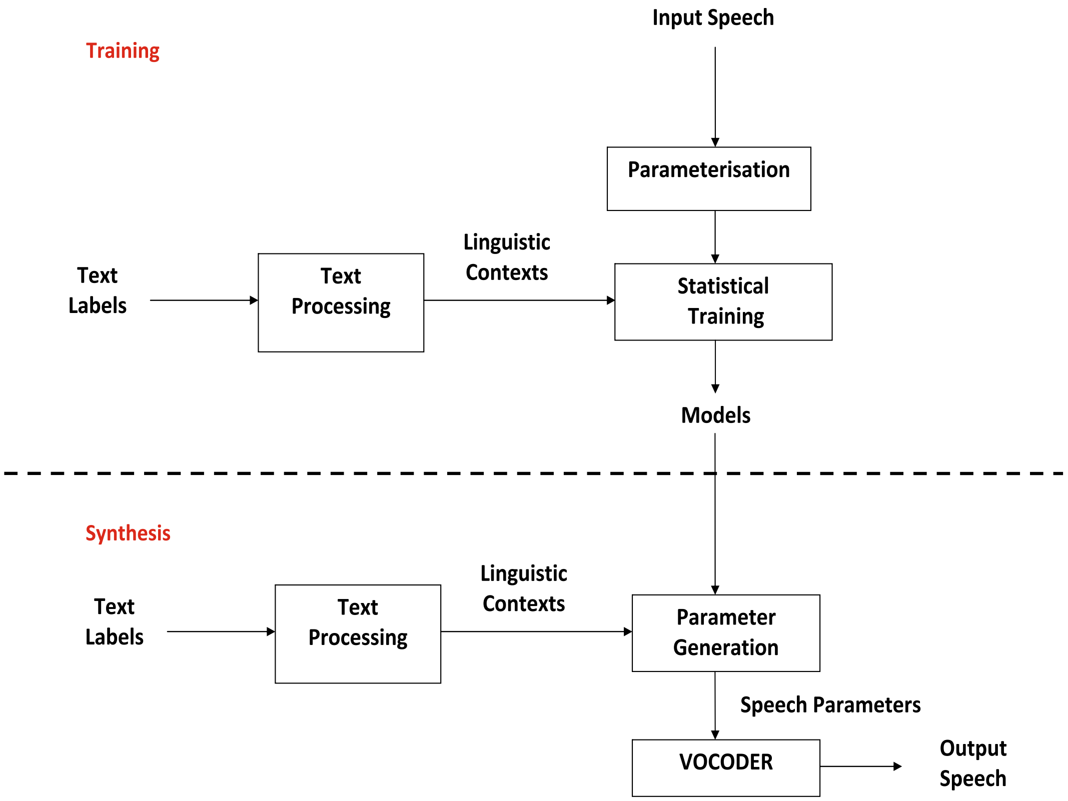

Fig. 4 Statistical parametric speech synthesis: the block diagram showing training and synthesis phases in building a statistical parametric speech synthesizer (King 2011)

need not be fixed size units. Also variable number of candidate units per target unit indicate that, the number of example units available in the database are different for each target unit having specific linguistic context. The path in bold dotted line indicates the optimum path obtained by the Viterbi search algorithm which has total minimum sum of target and join costs. As unit selection speech synthesis approach requires a large repository of labeled database, it demands higher memory requirements. The unit selection systems also suffer from the spectral and temporal mismatches at the concatenation points to some extent. Signal processing techniques, like Time Domain Pitch Synchronous Overlap Add (TD-PSOLA), are used to smooth the discontinuities at the concatenation points (Moulines and Charpentier 1990; Taylor 2009). Even though there are advances in the articulatory and formant synthesis approaches, unit selection based speech synthesis approach remain as the mostly used speech synthesis approach.

4.4 Statistical parametric speech synthesis

Statistical parametric speech synthesizers follow a model based approach for speech synthesis. In contrast to concatenative systems, instead of storing the units here the models corresponding to each unit will be stored in the repository. In the model based approach, the speech is parameterized and uses statical methods to build models for those speech parameters, hence the name *statistical parametric speech synthesis* (King 2011). In statistical parametric speech synthesis, the statistical parametric models are built using HMM models.[1] Hence statistical parametric speech synthesis is also known as HMM-based speech synthesis. The schematic block diagram of statistical parametric speech synthesis is given in Fig. 4. The HMM based speech synthesis is proposed in Yoshimura et al. (1999). The increased popularity of the HMM in speech recognition and availability of efficient learning algorithms (Forward-Backward algorithm, EM algorithm), computationally efficient search algorithms (Viterbi search) and parameter tying methods by decision tree clustering, are the motivation behind the development of HMM based speech synthesis systems (King 2011; Zen et al. 2007).

In HMM based speech synthesis, the speech in the database is parameterized into system and excitation source

[1] In contrast with the HMM based speech recognition, HMM based speech synthesis uses Hidden Semi Markov Models (HSMM) for representing the speech parameters for each sound unit (King 2011). The terminology of HMM models used in this chapter refers to HSMM models that is used for the speech synthesis.

components. The context dependent HMM models are built by training the HMMs simultaneously with source and system components for all the training data set. The HMM parameters during training are estimated using maximum likelihood criterion.

The speech parameters used for HMM training include, value of F_0 and 5 parameters for spectral envelope of the aperiodic excitation as the excitation parameters and 40 to 60 parameters are used for the spectral envelope (mel cepstral coefficients) (Zen et al. 2009; Yoshimura 1999). For the natural synthesis of speech dynamic features (delta and delta-delta coefficients) of both F_0 and spectrum are also used for modeling. These parameters are extracted typically at 5 ms frame rate. Like in speech recognition HMM models are trained with labeled speech data. Unlike the speech recognition case, here full context labels are used for the training. Use of these full contexts for the HMM modeling increases the complexity as compared to speech recognition case where simple context models like triphone HMM models are used. For the model complexity control, model parameter tying techniques are adopted where model parameters shared among models having similar contexts. Decision tree based clustering techniques are commonly used for parameter tying for HMM based speech synthesis. These parameter tying is responsible for retrieving models corresponding to the unseen contexts (for which there were no examples in the training data) during the synthesis. To synthesize a sentence, the text processing block generates the context dependent phoneme sequence. The corresponding stored HMM models are then retrieved (decision tree based clustering is used to find the model parameters for the phonemes with the unseen context factors) and concatenated to form the sentence HMM. From the sentence HMMs the speech parameters are generated. The number of frames of speech parameters to be generated is determined by the explicit duration model. The speech parameter trajectories are generated based on maximum likelihood parameter generation algorithm (MLPG) using the model parameters for dynamic features (Tokuda et al. 1995). The generated spectral and F_0 parameters are used for vocoding to synthesize the speech. The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) vocoder is generally used for synthesizing the speech in HMM based speech synthesis systems (Kawahara et al. 1999). Some HMM based systems also use MLSA (MeL Spectral Approximation) algorithm for synthesizing the speech (Imai 1983).

Statistical parametric speech synthesis offers more flexibility to adjust the speech characteristics as compared to unit selection synthesis systems. Due to these parameter flexibility, speaker independent speech synthesis systems are developed using various adaptation techniques (speaking style adaption, speaker adaption) using limited training data. Since statistical parametric speech synthesizers use vocoders for synthesizing the speech, synthesized speech sounds a little unnatural as compared to the speech synthesized using concatenative speech synthesizers. However, different techniques are being developed for improving the quality of the synthesized speech in statistical parametric synthesizer by incorporating glottal source parameters (Cabral et al. 2011), articulatory parameters (Ling et al. 2011), etc.

Despite the development in articulatory and formant speech synthesis in the recent years, currently the unit selection based concatenative and HMM based speech synthesis systems are the mostly used approaches in the area of speech synthesis. Hence to develop a good quality neutral speech synthesizer in the context of expressive speech synthesis, we can use either unit selection based speech synthesis system or HMM based speech synthesis system.

5 Analysis and estimation of expressive parameters

5.1 Expressive speech database

As analysis and estimation of expressive parameters are performed on an expressive speech database to frame the explicit rules for the ESS, the development of expressive speech database is a crucial step for the present work. Hence the present section starts with the review of expressive speech databases used for the ESS. Very few works related to ESS used commonly available database for ESS. Most of the works are based on the data collected by their own and publicly unavailable. These expressive databases differ by the language, type of expressions considered, type of text materials used, number of speakers and so on Schroder (2009).

In the literature two types of expressive data are collected. One is the expressive data simulated by actors (Burkhardt et al. 2005; Pitrelli et al. 2006; Fairbanks and Hoaglin 1939; Banks and Hoaglin 1941; Whiteside 1998) and the second is the spontaneous expressive data collected from a real life scenario (Williams and Stevens 1972; Jhonstone and Scherer 1999; Campbell 2004; Ishii and Campbell 2002). Most of the ESS systems described in Sect. 2 used expressions simulated by actors. Angry, happy, sad, fear and disgust are the commonly used emotions for the analysis in the case of simulated emotions. Williams and Stevens (1972) compared the spontaneous fear and sorrow emotions obtained from the radio announcer recording of the Hindenburg aircraft disaster, with the same sentences simulated by professional artists. This work concluded that, emotion specific parameters estimated from simulated emotion speech data is comparable with that of the real life emotion speech data. Johnston collected multimodal spontaneous data from the subjects by making them to participate in a competitive computer game (Jhonstone and Scherer 1999). The various instants of the game are manipulated in order to obtain various emotional responses from the subjects. Speech, electroglottogram (EGG) and electromyogram (EMG) are collected for tense, neutral, irritated, happy, depressed, bored and anxious expressions. Despite the practical difficulties in inducing the emotions in speaker, the combined analysis of acoustic features (from speech) and physiological features (EGG and EMG) gave a clearer indications of emotional states of the speakers. JST/CREST database collected by Campbell (2004) consists of natural telephonic conversation of various social interactions. The databases of spontaneous expressions are used for synthesizing expressive speech by unit selection approach.

If the goal of ESS system is to deploy in cartoon animations, call center applications or any other commercial applications the simulated expressions can be used for analysis. Since actors are well trained to produce emotions effectively, the use of these simulated emotions by them are recommended. The expressive data collected from non-professionals can also be used. If the application of ESS system is to deploy a dialogue system where the machine should interact more naturally with the user, the analysis of spontaneous expressions is better. Theune et al. (2006) described an interesting application of ESS aimed at children story telling. The database collected for this application is recordings of story narrated by professional artists. Johnson et al. (2002) discussed the ESS for military applications to simulate the shouted commands, shouted conversation, normal spoken commands and normal spoken conversation for animated characters. The data set used for training is the recordings of simulated shouting and normal commands.

5.1.1 Berlin emotional speech database

Burkhardt et al. (2005) described the development of acted emotion speech corpus in German language. The database is created with 10 professional actors (5 Males and 5 Females) of 10 emotionally unbiased sentences in six target emotions (Angry, Happy, Fear, Boredom, Sad and Disgust). Each file is recorded at 48 kHz sampling rate is downsampled to 16 kHz sampling rate with 16 bits per sample resolution. The listening test conducted on the recorded emotions gave more than 80 % emotion recognition rate by the listeners. Each recorded speech file is annotated at the word and syllable levels.

5.1.2 LDC emotional prosody speech transcripts database

The data collected in this database are grouped into distance/dominance category and emotional state category (Liberman et al. 2002). In the distance category the speakers have to give data by imagining whether speaker is speaking in a close room environment with a single listener or speaking with one or more people or speaking to someone who is standing far way of a room. In the emotional category, speakers have to give data in 14 different emotions with each of the emotions are well defined with a context. The data is collected from 8 professional actors (three males and five females). The actors were asked to speak semantically neutral English phrases with dates and numbers in a given category (emotional or distance category). 14 emotional states are hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt. Each speaker is given a script card in which the emotion category and phrases to be spoken are written and they are allowed to utter the phrase until the speaker is satisfied about the emotional category conveyed. The data is recorded in 22.5 kHz with 16 bits per sample bit resolution.

5.2 Studies on the analysis of expressive parameters

The following parameters are used for the expressive speech analysis,

- Prosodic parameters
- Excitation source parameters
- Vocal tract parameters

5.2.1 Studies on prosodic parameters

The typical prosodic features used for expressive analysis are parameters of F_0 contour, duration (sentence duration, syllable duration etc.) and intensity (Fairbanks and Hoaglin 1939; Williams and Stevens 1972; Bulut and Narayanan 2008; Hashizawa et al. 2004). F_0 is the average rate at which vocal folds vibrate for voiced sounds. The F_0 contour refers to the variation of F_0 with respect to time. The characteristics of these F_0 can be considered as the prosodic parameters. The duration parameters can be the total duration of the utterance or duration of the sound units like phones, syllables or words etc. The intensity parameter of the prosody is measure of loudness in the utterance. Fairbanks and Hoaglin (1939) studied the F_0 characteristics of five expressions (anger, fear, indifference, grief and contempt) simulated by actors. They found that the expressions can be classified on the basis of F_{0range} (absolute difference between F_{0min} value and the F_{0max} value), F_{0avg}, overall F_0 inflections (variations in F_0 values) and F_0 slope of F_0 contour. Based on the analysis of F_0 characteristics, the indifference expression showed lowest F_{0avg} and narrowest F_{0range}. Fear expression exhibited widest F_{0range} and highest F_{0avg}. The angry expression shows highest F_0 inflection.

Studies on duration characteristics by Fairbanks et al. on the same simulated expressions showed that the duration features like speech rate (number of words per minute), variation in number of pauses, length of the pauses and ratio of pause duration to total phonation time affects expressions (Banks and Hoaglin 1941). For instance, anger, indifference and fear expressions showed higher speaking rate and the expressions grief and contempt showed lower speaking rates (Banks and Hoaglin 1941). Williams explored the effect of prosodic parameters on actor simulated emotional expressions such as neutral, sorrow, angry and fear. The prosodic parameters considered for the study are $F0_{med}$ (median of F_0 values), $F0_{range}$ and speech rate. According to this study, the angry expressions found to have increased $F0_{med}$ and $F0_{range}$ and sorrow expressions showed reduced $F0_{med}$ and reduced $F0_{range}$ (Williams and Stevens 1972). The duration of the utterance spoken in fear expressions found to be longer than that of the anger expression.

The intensity parameter computed as the average spectral energy, found to be higher for anger for some syllables as compared to neutral speech (Williams and Stevens 1972). The proposed prosodic parameters for the simulated expression case is compared with various emotion contexts of the broadcast recordings during the Hindenburgh aircraft crash disaster. The study concluded that the characteristics of F_0 contour alone gives the indication of the emotional state of the speaker (Williams and Stevens 1972). Vroomen et al. showed that emotions can be accurately expressed by manipulating F_0 contour and sentence duration in a rule based manner. The seven expressions (neutral, joy, anger, boredom, indignation, sadness and fear) of two sentences in Dutch recorded by an actor are considered as the expressive data for their work. Based on perceptual and acoustic analysis, the F_0 contour of each expression is modeled using Dutch intonation rules (Vroomen et al. 1993). During the synthesis, the F_0 contour for each expression is generated and used to manipulate the F_0 contour of the neutral speech. The duration is modified by the linear compression according to the optimum modification factors obtained from the analysis stage. Various prosodic parameters by Murray and Arnott (1993) are $F0_{avg}$, $F0_{range}$, F_0 changes (F_0 inflections, F_0 inclination, F_0 declination), intensity and speech rate. Six expressions (anger, sadness, happiness, fear and disgust) along with the neutral expression are considered for the analysis. According to the analysis the fear expressions and disgust expressions showed the highest and lowest speech rate, respectively. The $F0_{avg}$ was highest for anger and lowest for disgust. $F0_{range}$ was narrower for the sad expressions. Anger and happiness got the highest intensity whereas disgust showed lower intensity. Among the F_0 changes, anger showed abrupt F_0 changes in stressed vowels and happiness showed smooth and upward F_0 inflections. Sadness and disgust showed downward deflections whereas fear expression showed normal F_0 changes. The rule based expressive speech synthesis system developed by Murray and Arnott (1995) modifies the prosodic parameters at the phoneme level according to the prosodic analysis described in Murray and Arnott (1993). The rules are set for each prosodic parameter in the HAMLET neutral formant speech synthesizer for synthesizing the speech in the target expression (Murray and Arnott 1995). Hashizawa et al. considered $F0_{max}$, speech rate and F_0 of the pitch accented syllables for the analysis. The analysis showed that, the $F0_{max}$ is the highest for anger emotion, F_0 and pitch accents are enhanced for joy and $F0_{max}$ is minimum for sad emotion (Hashizawa et al. 2004). Tao et al. (2006) used $F0_{avg}$, $F0_{topline}$, $F0_{baseline}$, syllable duration and intensity as the prosodic parameters. The $F0_{topline}$ is the mean of the line connecting the peaks in F_0 contour and $F0_{baseline}$ is the mean of the line connecting the valleys in F_0 contour. According to Tao et al., the $F0_{avg}$ and $F0_{topline}$ provide more classification abilities for five emotions (neutral, anger, happiness, fear and sadness). Murtaza et al. showed the significance of $F0_{range}$ than $F0_{avg}$ in classifying four emotions (neutral, anger, happy and sad) of two sentences from two speakers.

5.2.2 Studies on excitation parameters

Excitation parameters refer to the parameters representing the characteristics of the excitation source. The excitation source parameters are analyzed at subsegmental and segmental levels. The parameters computed within 2 to 3 pitch periods (10–20 ms) of speech are termed as the segmental parameters. The parameters such as jitter and shimmer are examples of excitation parameters estimated at the segmental level. The parameters estimated within a pitch period of speech are subsegmental parameters. The glottal flow parameters like open quotient (OQ), Return quotient (RQ) and speech quotient (SQ) are examples of excitation parameters at the subsegmental level.

Jitter is the average change of F_0 from one cycle to another, where as, shimmer is the change in the excitation strength from one cycle to another. Whiteside (1998) has shown the significance of jitter and shimmer in discriminating various emotions. Seven expressions (neutral, cold anger, hot anger, happiness, sadness, interest and elation) of 5 sentences from two speakers are used for this study. The prosodic parameters like mean intensity, standard deviation of intensity and $F0_{avg}$ are also used along with mean of shimmer and jitter for the analysis of seven expressions. According to the analysis done in this work, hot anger possessed highest mean jitter and mean shimmer and sadness showed minimum mean jitter and mean shimmer. Using these five parameters, the expression discrimination accuracy was found to be 88.9 % and 85.7 % for the two speakers, respectively (Whiteside 1998). Jhonston et al. performed the expressive analysis on both spontaneously recorded expressions and simulated expressions. The first part of the paper used multimodal (speech, EGG, EMG) data of spontaneous emotions collected from subjects by exposing them to different instants of a manipulated computer game (Jhonstone and Scherer 1999). The participants were asked to pronounce the sentences to be recorded and asked them to choose one expression from the list of expressions (irritated, disappointed, surprised, relieved, helpless and alarmed). Based on the choice of expressions made by the participants at different situations of the game, the recorded expressions are categorized into low coping, high coping, obstructive and constructive responses. The parameters used for the analysis are glottal slope, F_{0range}, heart period, respiratory period and respiratory depth. According to the analysis, the glottal slope obtained from the EGG and the heart rate tend to be higher for obstructive situations and the low coping situations are characterized by the longer respiratory cycle. The second part of the study consisted of glottal analysis on EGG data of seven expressions (tense, neutral, happy, irritated, depressed, boredom and anxious) of 5 digit strings, short phrases and sustained vowel /a/. The expressive data of these seven expressions were collected from eight speakers. The speakers were asked to imagine the emotions for recording the expressive data. The excitation parameters used for the analysis are mean jitter, closing quotient (glottal closing time of the glottis as a percentage of pitch period, T_0). According to the analysis, mean jitter was highest for happy and anxious expressions and lowest for boredom and depressed expressions. The depressed and boredom expression showed the higher values of closing quotient and anxious expression showed lowest values for closing quotient. Cabral et al. used jitter, shimmer and glottal flow parameters like OQ, RQ and SQ as the excitation parameters for synthesizing emotion (Campell et al. 2006). In this work, the excitation parameters are extracted from seven emotions (angry, happy, fear, boredom, neutral, sad and disgust) of German emotional speech database. According to the excitation parameter analysis presented in this work, happy and fear expressions tend to show decrease in OQ. The breathy quality of the anger expressions are confirmed by the decrease in SQ and RQ compared to other expressions. Along with prosodic parameters Tao et al. (2006) used jitter to analyze five emotional expressions like neutral, anger, happiness, fear and sadness for the task of neutral to expressive speech conversion. According to this study, the happiness expression tend to have highest jitter and sadness showed lowest jitter.

5.2.3 *Studies on vocal tract parameters*

Formant frequencies (F_1, F_2, F_3, F_4 and F_5) and bandwidth associated with each formant form important characteristics of the vocal tract system. Mean F_1, mean F_2 and F_1 bandwidth are the vocal tract cues reviewed by Scherer (1986). The acoustic characteristics of around 14 expressions are reviewed in Scherer (1986). Compared to other emotions, lower mean F_1 was observed for happy and elation expressions and higher mean F_1 was observed for other expressions. Whereas lower mean F_2 was observed for all the expressions other than happy and elation. The expressions, hot anger, cold anger, disgust and fear, tend to show narrower F_1 bandwidth (Scherer 1986). The parameters considered for these articulatory stimulus are F_1 mean, F_2 mean and corresponding formant bandwidths. Ishii et al. used a subset of spontaneous expressions collected in JST (Japan Science & Technology) CREST (Core Research for Evolutional Science and Technology) ESP (Expressive Speech Processing) project. The spontaneous expressive data is collected by recording subject's daily spoken conversations using mini recording devices and wearing head mounted close speaking microphones. After recording the data, speakers were asked to label the expressions based on their mood at various times in the conversations such as neutral, worried, contend, happy, bright, sad, angry, tension, energy ("Energy" is categorized based on the global intensity of the speech) etc. By analyzing the average F_3 and average F_4 parameters of words in bright, energy and tension expressions, Ishii et al. found that the average F_4 is higher for bright expressions than the expressions labeled with energy. There was no correlation observed for average F_3 values indicating inconsistency of F_3 parameter for the same expressions. Erickson et al. (2008) studied the effect of formant frequencies on spontaneous sad emotions. The spontaneous sad emotional data is collected in two sessions through the telephonic conversions with the subject. The spontaneous sad emotions were evoked by asking about the sad demise of the subject's mother. Lowering of F_2, F_3 and F_4 was observed for the sad emotions when compared with the non emotional data.

Table 1 presents the summary of the review of the studies made on expressive parameters. The columns given in the table represent the contributors, choice and type of expressive data used in their work, expressive parameters considered and the important findings of their work.

5.3 Estimation of expressive parameters

This section reviews various studies made on estimation of prosodic, excitation and VT parameters from speech.

5.3.1 *Estimation of prosodic parameters*

As most of the works related to expressive speech synthesis use prosodic parameters as the expression dependent parameters, it is essential to accurately estimate these prosodic parameters for the expressive speech analysis. The features of F_0 contour, speech rate and intensity are the prosodic parameters reviewed in this section.

F_0 or pitch is the fundamental frequency of vibration of the vocal folds during the production of voiced sounds.

Table 1 Summary of various studies about expressive parameters

Author	Expressions	Type of expressive data used	Expressive parameters explored	Findings
Williams and Stevens (1972)	Anger, sorrow, fear and neutral	Simulated by actors	F_0 mean, F_0 range, speech rate and Energy	1. F_0 contour as the indicator of different emotional states 2. F_0 parameters of simulated and real emotions are similar
Scherer (1986)	Happy, cold anger, hot anger, anxiety, disgust and sad	Simulated by actors	F_0 parameters, F_1 mean, F_2 Mean and Formant Band width	F_0 parameters along with VT parameters represents the acoustic properties of emotion
Murray and Arnott (1993)	Angry, happy, sad, fear and disgust	Simulated by actors	Speech rate, F_0 mean, F_0 range and Intensity	Emphasizes the role of prosodic parameters in synthesizing emotions
Whiteside (1998)	Cold anger, hot anger, happy, sad, interest and elation	5 Short sentences simulated by two speakers	Mean of overall jitter, Mean of overall shimmer	Significance of jitter and shimmer in discriminating the emotions
Jhonstone and Scherer (1999)	Tense, neutral, irritated, happy, depressed, bored, anxious	data (EGG, EMG and speech) collected during computer game events	Jitter, Glottal closing Time	EGG signal gives emotion dependent characteristics
Ishii and Campbell (2002)	Neutral, worried, bored, polite, depressed, angry	Natural telephonic conversation recorded	F_0 parameters, F_3 mean, F_4 mean	F_4 influences different voice qualities
Hashizawa et al. (2004)	Angry, happy and sad	Isolated Words by professional announcers	Speech rate, F_0 max and Pitch Accent	1. F_0 max is higher for anger 2. For happy both accents and F_0 will be enhanced 3. F_0 and accents were suppressed for Sad
Cabral and Oliveira (2006a)	Angry, happy, sad, fear, surprise, boredom, disgust	Simulated by actors	Jitter, shimmer, glottal wave parameters (OQ, SQ, RQ)	Better recognition rates obtained for happy, angry and fear
Bulut and Narayanan (2008)	Angry, happy, sad and neutral	Simulated by professional and non-professional actors	F_0 mean, F_0 range, F_0 stylization characteristics	Changes in F_0 range significantly changes perceived emotions

Since vocal folds vibrate only during the production of voiced sounds, F_0 is defined only for voiced sounds. F_0 is undefined or represented by random values during the production of unvoiced sounds such as fricatives. Representation of F_0 values versus the time instants at which they are calculated is termed as F_0 contour or pitch contour. In order to derive F_0 contour, F_0 values have to be accurately estimated from the speech. F_0 estimation techniques described in the literature are broadly classified into block processing based approach and event based approach (Hess 1983; Yegnanarayana and Murty 2009). Block processing approach computes the average F_0 from block of speech segment where as event based approach accurately determines the instantaneous F_0 by processing entire speech utterance. Most of the earlier works employ block processing approach for estimating F_0 for expressive speech analysis (Williams and Stevens 1972; Scherer 1986; Murray and Arnott 1993; Whiteside 1998; Jhonstone and Scherer 1999; Ishii and Campbell 2002; Hashizawa et al. 2004; Bulut and Narayanan 2008). Auto correlation (Hess 1983; Deller et al. 1993), cepstral analysis Deller et al. (1993), simplified inverse filtering (SIFT) (Markel 1972) and average magnitude difference function (AMDF) (Ross et al. 1974) are the popular methods for estimating F_0 by block processing. A robust method by exploiting the properties of Hilbert envelope (HE) of LP residual for reliably estimating average F_0 in adverse conditions is proposed by Prasanna and Yegnanarayana (2004).

In order to accurately estimate all the instantaneous F_0 values for the entire speech utterance, the event based approach is used. The instantaneous pitch period is defined as the interval between glottal closing instant of one cycle to the next. As the discontinuities related to pitch occurs at the instants of glottal closure where the maximum excitation of the vocal tract occurs, accurate determination of these instants of significant excitation or epochs are essential for computing the instantaneous F_0. The epochs or instants of significant excitation can be defined as instants of glottal closure incase of voiced speech or onset of burst or frication incase of unvoiced speech (Murty and Yegnanarayana 2008; Yegnanarayana and Murty 2009). The interaction of vocal tract in the produced speech makes the estimation of epochs location a challenging task. There are several methods proposed in the literature to estimate the epochs location accurately. The epoch estimation using group delay (GD) functions (Smits and Yegnanarayana 1995), Dynamic programming based projected phase-slope algorithm (DYPSA) (Naylor et al. 2007), HE based method and Zero Frequency Filtering (ZFF) (Murty and Yegnanarayana 2008) of speech based methods are the popular existing methods for the epoch estimation. The instantaneous pitch period is computed as the interval between successive epochs location (Rao and Yegnanarayana 2006b). The instantaneous pitch period is also termed as the epoch interval (Rao and Yegnanarayana 2006b). The instantaneous F_0 is computed by scaling the reciprocal of epoch interval with F_s. The representation of instantaneous F_0 values at the corresponding epochs location gives the instantaneous F_0 contour of the utterance. The significant F_0 parameters derived from the F_0 contour are F_{0avg}, F_{0max}, F_{0min}, F_{0range}. These F_0 parameters can be computed using the following equations,

$$F_{0avg} = \frac{1}{N} \cdot \sum_{i=1}^{N} F_{0i} \tag{1}$$

$$F_{0max} = \max\{F_{0i}, i = 1, 2, \ldots, N\} \tag{2}$$

$$F_{0min} = \min\{F_{0i}, i = 1, 2, \ldots, N\} \tag{3}$$

$$F_{0range} = F_{0max} - F_{0min} \tag{4}$$

Various duration parameters that are used for expressive analysis are the speech rates at sentence, syllable and phoneme levels and number of pauses. Unlike the F_0 parameters, the estimation of duration parameters are mostly measured directly from the database. Fairbanks et al. computed the speech rate by counting the number of words uttered per second for analyzing emotional expressions (Banks and Hoaglin 1941). Burkhardt et al. (2005) used syllable duration as the prosodic parameter for the analysis of the expressions given in German emotional speech database. The syllable boundaries are labeled manually by listening, analyzing spectrograms and simultaneous EGG recordings. Cabral et al. used sentence duration of each utterance as the duration parameter for analyzing various expressions. Murray and Arnott (1993) analyzed duration at the phoneme level. The duration of phonemes are estimated directly from the phone boundary labeling of the utterances. The automatic phone boundary marking can also be done using HMM based force alignment techniques (Clark et al. 2007). Other prosodic parameter used for emotion analysis is the intensity. The intensity is measured by computing the energy of the utterance.

5.3.2 Estimation of excitation parameters

According to the review of excitation parameters given in Sect. 5.2.2, various excitation parameters used for expressive speech analysis are the segmental parameters like jitter and shimmer, and subsegmental parameters related to glottal flow and strength of excitation. This section reviews the method employed to estimate these parameters.

Shimmer is a measure of strength of excitation (Murty and Yegnanarayana 2009) of the glottal wave which is defined as the change of strength of excitation pulses from one cycle to another. One of the method to characterize the glottal activity is from the LP residual obtained by the LP analysis of speech (Makhoul 1975). During the glottal activity, the LP residual has high energy region and during non glottal activity region, LP residual shows noisy characteristics. Analysis of the excitation source based on LP residual depends on the accuracy of LP analysis. Murty and Yegnanarayana (2008) described a method to compute the glottal activity and strength of excitation in speech based on the ZFF of speech. Since the rate of vibration of the vocal folds is proportional to the glottal air flow, the excitation strength can be found by measuring the sharpness with which glottal closure occurs. This can be computed by measuring the slope of the ZFF signal around the epochs location. Now shimmer can be measured as the change of excitation pulse between successive epochs location. Farrus and Hernando (2009) computed shimmer as the variation of the peak to peak amplitude values in consecutive pitch period and then proposed shimmer measurements in the various levels for speaker verification. The average shimmer measurement described in Farrus and Hernando (2009) is given by (5)

$$S_{avg} = \frac{1}{N-1} \sum_{i=1}^{N} |A_i - A_{i+1}| \tag{5}$$

where A_i is peak to peak amplitude in the ith pitch period and N is the total number of pitch periods.

Jitter is estimated by measuring the average change in pitch period from one pitch cycle to another. Jitter is derived from the instantaneous F_0 contour. Method to estimate the jitter in different levels for speaker verification task is described in Farrus and Hernando (2009)

$$J_{avg} = \frac{1}{N-1} \sum_{i=1}^{N} |T_{0i} - T_{0i+1}| \tag{6}$$

where T_{0i} is ith pitch period and N is the total number of pitch periods.

These shimmer and jitter measurements can be used to find the variation of target expressions with respect to the neutral speech.

There are several methods discussed in the literature to estimate glottal waveform parameters directly from speech.

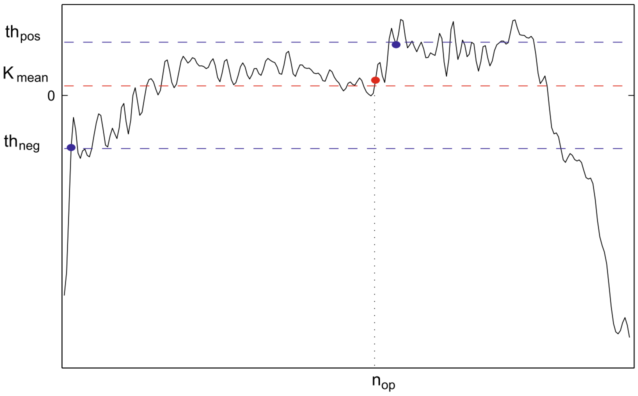

Fig. 5 Locating the instant of glottal opening in a short time segment of LP residual (figure used with the permission of J.P. Cabral)

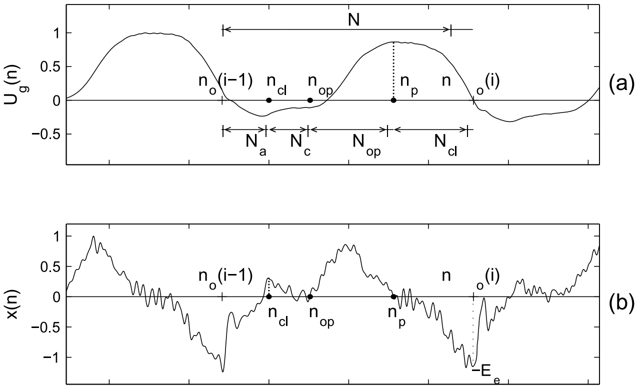

Fig. 6 Representation of glottal phases in a (**a**) glottal cycle and (**b**) in its derivative for a short time segment of LP residual (figure used with the permission of J.P. Cabral)

Fant et al. (1985) developed LF (Liljencrants and Fant) model to uniquely represent glottal flow derivative for a given pitch period. The LF model is a four parameter model developed based on the glottal closure instants and glottal closure discontinuity points. The four parameters are frequency, amplitude, growth constant of sinusoid and recovery time constant. Cabral et al. estimated the expression dependent glottal flow parameters from the LP residual. These parameters are estimated by integrating the LP residual. These glottal flow parameters are measured by estimating the following time instants first:

- *Glottal closure instants*, n_o: By estimating the instants of glottal closure from the LP residual.
- *Closed phase instants*, n_{cl}: It is the instant at which closed phase of the glottis starts. n_{cl} is calculated by finding the instant of the first peak after the zero crossing.
- *Glottal Opening instants*, n_{op}: The time instant of the opening phase is calculated by setting a positive (th_{pos}) and negative threshold (th_{neg}) to the short time signal. where th_{pos} is the 75 % of the maximum value of the signal energy and $th_{neg} = -th_{pos}$. The first point of the positive growing part that intercept with the negative threshold and the last point of the growing part of the opening phase that intercepts the positive threshold are calculated (Cabral 2006). The average signal amplitude value, k_{mean} between the two estimated points is calculated. The last positive crossing point of the signal over the k_{mean} axis is estimated to be n_{op}. The calculation of n_{op} is shown in Fig. 5.
- *Maximum of glottal flow*, n_p: The zero crossing of DC value between n_{op} and end of the short time signal.

The estimated time instants are used to compute the duration of the glottal cycle phases, Return phase (N_a), Peak flow duration (N_e), Closed phase (N_c), Opening phase (N_{op}) and Closing phase, (N_{cl})

$$N_a = n_{cl} \tag{7}$$

$$N_e = N - n_{op} \tag{8}$$

$$N_c = N - N_a - N_e \tag{9}$$

$$N_{op} = n_p - n_{op} \tag{10}$$

$$N_{cl} = N - n_p \tag{11}$$

where N is the total duration of the glottal cycle. The duration of these phases in a glottal cycle is pictorially represented in Fig. 6. The glottal flow parameters like OQ, RQ and SQ are calculated as given by the following equations,

$$OQ = \frac{N_a + N_e}{N} \tag{12}$$

$$RQ = \frac{N_a}{N} \tag{13}$$

$$SQ = \frac{N_{op}}{N_{cl}} \tag{14}$$

The other important excitation source parameter is the strength of excitation. The strength of excitation is a sub-segmental feature which is the strength with which the vocal folds are vibrating during the production of voiced speech (Murty and Yegnanarayana 2009). One of the methods to compute the excitation strength is from the LP residual obtained by the LP analysis of speech (Murthy and Yegnanarayana 1999). In the glottal activity region, the LP residual has high energy and during non-glottal activity region, LP residual shows low energy noisy characteristics (Murthy and Yegnanarayana 1999). The strength of excitation is computed by computing the energy of the residual samples in the region around the glottal closure instants. Murty and Yegnanarayana (2009) described a method to compute the glottal activity and strength of excitation in speech using the ZFF based epoch extraction. Since rate of vibration of the vocal folds is proportional to the glottal air flow, the excitation strength can be found by measuring the sharpness with which glottal closure occurs (Murty and Yegnanarayana 2009). This can be computed by measuring the slope of the ZFFS around the epochs location.

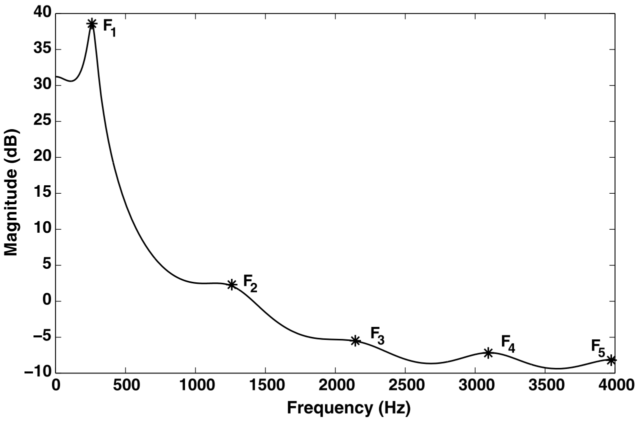

Fig. 7 The formant estimation from LP spectrum: this figure shows the linear prediction spectrum and Formant locations (indicated by '*') obtained from the peaks in the LP spectrum

5.3.3 *Estimation of vocal tract parameters*

There are many methods discussed in the literature to estimate the formants from speech. In Schafer and Rabiner (1970) used a method to extract formants by picking peaks from smoothed log-spectra obtained by cepstral analysis. Figure 7 indicates how the formants are located by picking the peaks in the LP spectrum. The spurious peaks in the log magnitude spectrum causes wrong identification of the formants and is the main disadvantage of the approach. Formant extraction by the linear prediction analysis solved the issue of spurious peaks appearing in the LP spectrum (Atal and Hanauer 1971). But here also peak picking wrongly estimates the formants in the case of merged peaks in the LP spectrum. Yegnanarayana (1978) showed that the differentiated linear prediction phase spectrum can clearly resolve the merged peaks because of the additive nature of the phase of cascaded digital resonators. The Formants can be extracted by picking peaks of the differentiated LP phase spectrum. To further resolve the spectral peaks in the smoothed log spectra, properties of the group delay function of minimum phase signals are utilized in Murthy and Yegnanarayana (1991). The significance of measuring the free resonances of the vocal tract to analyze various regions like consonant vowel transitions is described in Yegnanarayana and Veldhuis (1998). The speech produced during the closed phase of the vocal folds are mainly due to vocal tract. At this time the vocal tract tube is closed at one end and hence the resonances produced are free from the vocal fold vibrations and glottal air flow that occur during the opening phase of the glottis. The formant parameters are extracted from the analysis segments taken around the glottal closure instants. The pole zero model of these analysis frames are determined. This can be represented mathematically as follows (Yegnanarayana and Veldhuis 1998):

$$r(n) = \sum_{k=1}^{p} A_k \rho_k^n e^{i\theta_k n} = \sum_{l=1}^{p/2} \rho_l^n \left(A_l e^{i\theta_l n} + \overline{A}_l e^{-i\theta_l n} \right) \tag{15}$$

where n is the discrete time index and $p/2$ is the number of formants. The variable θ_k is the kth formant frequency such that $-\pi < \theta_k \leq +\pi$. As $r(n)$ is real, it can be represented as complex conjugate pair. The factor A_k is the formant amplitude and ρ_k is the formant damping factor, where $0 < \rho_k \leq 1$. It can be seen that the formant frequency F_k and bandwidth B_k can be computed from (16) and (17) as given in Yegnanarayana and Veldhuis (1998)

$$F_k = \frac{F_s}{2\pi}\theta_k \tag{16}$$

$$B_k = -\frac{F_s}{\pi}\ln(\rho_k) \tag{17}$$

6 Incorporation of expressive parameters

The expressive parameters have to be incorporated into the neutral speech according to rules set in the expressive analysis stage for effective synthesis of the speech in the target expression. The incorporation of expressive parameters are performed at the prosodic, excitation and vocal tract levels. This section reviews methods used to incorporate the expressive parameters at each level.

6.1 Methods to incorporate prosodic parameters

The expression specific F_0, duration and intensity parameters can be incorporated by prosody modification algorithms. Manipulation of F_0, duration and intensity of the given speech without affecting the perceptual quality is termed as prosody modification (Rao and Yegnanarayana 2006b; Mourlines and Laroche 1995). There are several methods discussed in the literature for prosody modification (Rao and Yegnanarayana 2006b; Hardam 1990; Mourlines and Laroche 1995). The approaches like Over Lap Add (OLA), Synchronous Over Lap Add (SOLA) and Pitch Synchronous Over Lap Add (PSOLA) operate directly on the speech waveform to modify the prosodic parameters (Mourlines and Laroche 1995). The OLA and SOLA methods are mainly used for time scale modification of the given speech signal (Hardam 1990). The duration modification here is achieved by overlap adding the analysis frames having time scaled length chosen from the crosscorrelation with the actual frames of the given speech signal. Development of the PSOLA allowed both time scale and pitch scale modification by using pitch marks as the anchor points (Rao and Yegnanarayana 2006b; Mourlines and Laroche

1995). In PSOLA method, in general, the pitch modification is achieved by placing the analysis windows around the modified pitch marks and adding the overlap regions. In the timescale modification, first resampling the actual pitch mark locations according to the desired timescale and then the analysis frames around the actual pitch mark is copied and overlap added to the new pitch locations closest to the original location. The resulting signal obtained will be duration modified according to the desired modification factor.

Depending on the domain in which PSOLA is applied there are Time Domain PSOLA (TD-PSOLA), Frequency Domain (FD-PSOLA) and Linear Prediction PSOLA (LP-PSOLA) (Mourlines and Laroche 1995). In FD-PSOLA prior to overlap add, the spacing between the pitch and harmonics of excitation signal obtained by the source-filter decomposition are modified according to the desired pitch modification factors by resampling in the frequency domain (Mourlines and Laroche 1995). Unlike TD-PSOLA, LP-PSOLA operates on the LP residual of the speech signal to be prosody modified. As the LP residual samples are less correlated than speech samples, the overlap-adding of residual analysis frames give less distortion. H. Kawahara developed a method to manipulate the speech parameters like F_0, speech rate and vocal tract length using speech representation and transformation using adaptive interpolation of weighted spectrum (STRAIGHT) (Kawahara 1997). Here a pitch adaptive speech analysis is carried out for speech parameter manipulations. The instantaneous F_0 estimation method developed as part the work uses Gabor filters for finer frequency resolutions. The manipulated speech is obtained by reconstructing the smooth time frequency representation using bilinear transformations. Smoothing is done to remove the pitch periodicity effects in the time-frequency surface representation of the original speech. Muralishankar et al. proposed F_0 modification method using discrete time cosine transformation (DCT) of pitch synchronous residual frames (Muralishankar et al. 2004). Here the DCT coefficients are estimated from the pitch synchronous residual frames obtained after the LP analysis. According to the pitch modification factors the DCT coefficients are either truncated (increase in F_0) or padded with zeros (decrease in F_0). For instance, if N_1 is the number of DCT coefficients in residual frame, N_2 point IDCT is taken where N_2 is N_1 divided by the F_0 modification factor. $N_2 - N_1$ trailing end DCT coefficients are removed for increasing F_0 and $N_1 - N_2$ zeros are padded to decrease F_0. Each frame of speech is then synthesized by the LP filtering with the LP coefficients computed.

A method of prosody modification by accurately determining the epochs location is proposed in Rao and Yegnanarayana (2006b), Smits and Yegnanarayana (1995), Rao and Yegnanarayana (2003). This type of prosody modification is generally known as epoch based prosody modification. The steps involved in the epoch based prosody modification are the following:

- Finding the accurate epochs location
- Modifying the epochs location according to the desired prosodic parameters
- Reconstruct the prosody modified speech

6.1.1 Estimating epochs location

As described earlier epochs in speech can be defined as the instants of glottal closure in voiced speech and onset of burst or frication in unvoiced case (Smits and Yegnanarayana 1995; Murty and Yegnanarayana 2008). Due to the effect of vocal tract, the accurate estimation of epochs location from speech is a challenging task. There are many methods proposed for the estimation of the epochs location from speech (Naylor et al. 2007; Smits and Yegnanarayana 1995; Murty and Yegnanarayana 2008). Smits and Yegnanarayana (1995) proposed group delay (GD) based approach to estimate the epochs location from LP residual of speech. In GD method, GD function is computed from the LP residual by considering in blocks of about 1–2 pitch periods length with shift of every sample. If $e(n)$ is the LP residual, the Fourier transform of $e(n)$ and its time weighted function are given by (18) and (19)

$$E(\omega) = FT\big[e(n)\big] = E_R + jE_I \tag{18}$$

$$F(\omega) = FT\big[ne(n)\big] = F_R + jF_I \tag{19}$$

The computation of GD function $\tau(\omega)$ is given in (20)

$$\tau(\omega) = -\phi'(\omega) = \frac{E_R F_R + E_I F_I}{E_R^2 + E_I^2} \tag{20}$$

After removing isolated peaks from $\tau(\omega)$ using a 5 point median filter, average value of the GD function is computed for each block of residual with a shift of one sample. The average GD function obtained for every sample shift is known as the phase slope function. The epochs are estimated as the zero crossings of the phase slope function. Later, the robustness of GD based method against various degradation are studied by Satyanarayana et al. in Murthy and Yegnanarayana (1999). As the GD function is computed from LP residual frames for every sample shift, the GD based epochs estimation is a computationally complex task (Smits and Yegnanarayana 1995). To reduce the time complexity in estimating the epochs from LP residual, a two stage processing approach is proposed by Rao et al. (2007). In the first stage, the approximate epochs location are estimated from the HE of LP residual and in the next phase, the GD function calculated around the approximate epochs location obtained from the HE of LP residual. Dynamic programming based projected phase slope algorithm (DYPSA) is proposed by Naylor et al. (2007) found to provide better epochs estimation accuracy than GD method.

Recently, a simple, fast and accurate method for estimating epochs from speech is proposed by Murty and Yegnanarayana (2008). In ZFF method, the speech is passed through the cascade of two zero frequency resonators (ZFR). The ZFR output $y(n)$ is given by (21)

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{21}$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$ and $x(n)$ is difference speech obtained by the successive difference of samples in the speech signal which is given by $x(n) = s(n) - s(n-1)$. The variations in the ZFR output due to epochs are obtained by subtracting local mean from the ZFR output. This local mean subtracted ZFR output is termed as the zero frequency filtered signal (ZFFS). The local mean substraction from ZFR output can be expressed as,

$$\hat{y}(n) = y(n) - \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{22}$$

Here $2N + 1$ corresponds to the size of window used for computing the local mean, which is typically the average pitch period computed over a long segment of speech. Thus, the epochs location will be the positive zero crossings of the ZFFS. The accuracy of the epochs estimated using ZFF method is more as compared to DYPSA and GD method (Murty and Yegnanarayana 2008). In ZFF method, as the epochs are estimated directly from speech without computing LP residual, the method is found to be computationally fast as compared to DYPSA and GD methods (Murty and Yegnanarayana 2008).

6.1.2 *Modifying epochs location for prosody modification*

Rao and Yegnanarayana (2006b) described a method to modify the epochs location according to the desired prosody modification factors. In this method the modified epochs location are obtained by deriving the epoch intervals. The epoch intervals are derived as the difference between successive epochs location. The epoch interval plot is then generated by interpolating the epoch intervals of successive epochs location. In the case of F_0 modification, this epoch interval plot obtained for the entire utterance is scaled according to desired pitch modification factor. The epoch interval plot is resampled according to the desired duration modification factor in case of duration modification. Modified epochs location are obtained from the resampled and/or scaled interpolated epoch interval plot. For instance, if A is the starting sample index of the modified epoch interval plot, the modified epoch location B is computed by adding, the modified epoch interval number of samples at Ath location in the modified epoch interval plot, to Ath time index. Similarly, modified epoch location sample index C is obtained by adding modified epoch interval number of samples at Bth location in the modified epoch interval to the location sample index B.

6.1.3 *Reconstructing the prosody modified speech*

The prosody modified speech is synthesized by generating the modified residual waveform. For generating the modified LP residual, the modified epochs location that are nearest to the original epochs location are found. Perceptually relevant (20 % epoch interval region starting from the epoch) residual samples starting from original epochs location are copied to the new modified epochs location. The perceptually relevant residual samples in the epoch interval refers to human listening in TTS and speech enhancement tasks. The remaining residual samples in the original epoch interval other than the samples in the perceptually relevant region, are resampled to fill up the modified epoch interval. Resampling is used mainly to avoid the spectral discontinuities introduced due to truncation of epoch intervals (in case of raising F_0) and replication of samples (in case of lowering F_0). This way the prosody modified LP residual is reconstructed. In pitch modification, as the duration of the utterance remains same, the LPCs computed from the original speech signal are excited by the modified LP residual to synthesize the pitch modified speech. In duration modification, since overall duration of the utterances are changing, LPCs of original speech are updated for very frame shift according to duration modification factors. These new LPCs are excited by the modified LP residual to synthesize the duration modified speech. The perceptual quality of the synthesized files for various pitch and duration modification factors are evaluated using perceptual tests. 25 research scholars of the lab participated in the perception experiments. Here, the subjects were asked to evaluate the prosody modified speech files based on the distortion present in the speech file. The filenames of each method are coded to avoid the biasing towards a particular method. The subjects were asked to judge their opinion scores on a five point scale where score 1 corresponds to poor quality with objectionable distortion and score 5 is excellent quality with no perceptual distortion. For moderate pitch and duration modification factors, the epoch based residual modification and LP-PSOLA based approaches provide almost equal mean opinion scores (MOS). As compared to LP-PSOLA approach, epoch based prosody modification approach had higher MOS scores for the case of extreme modification factors (modification factors greater than 2 and less than 0.5).

Figure 8 plots the pitch modification example by increasing the pitch of the original speech segment by factor of 1.5. As we can observe that there are around 13 pitch cycles in Fig. 8(c) which is 1.5 times than the number of pitch cycles (8 pitch cycles) in the LP residual segment of the original signal as shown in Fig. 8(b).

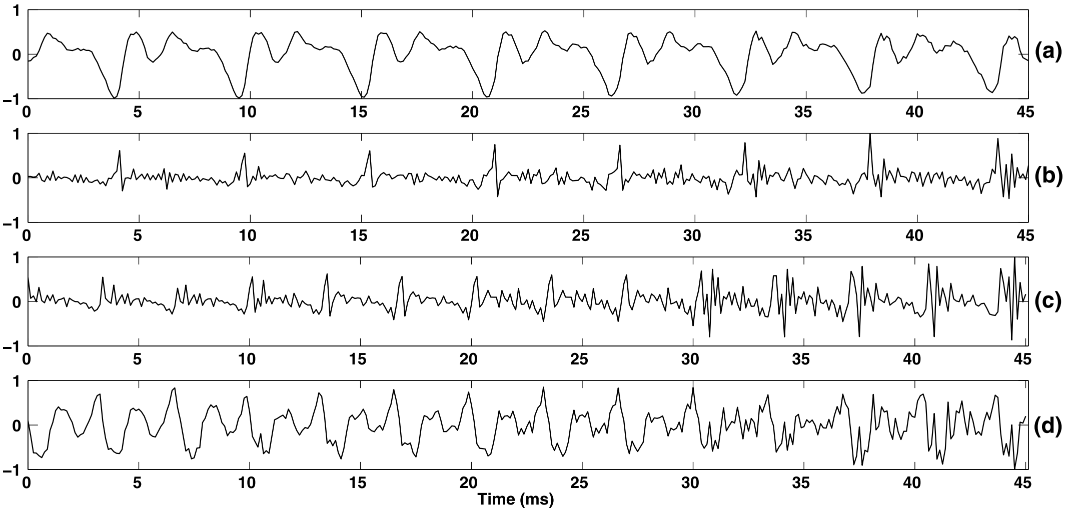

Fig. 8 Pitch Modification: (**a**) Longer Segment of a voiced speech, (**b**) its LP residual, (**c**) modified LP residual by increasing the pitch by 1.5 times and (**d**) reconstructed pitch modified speech

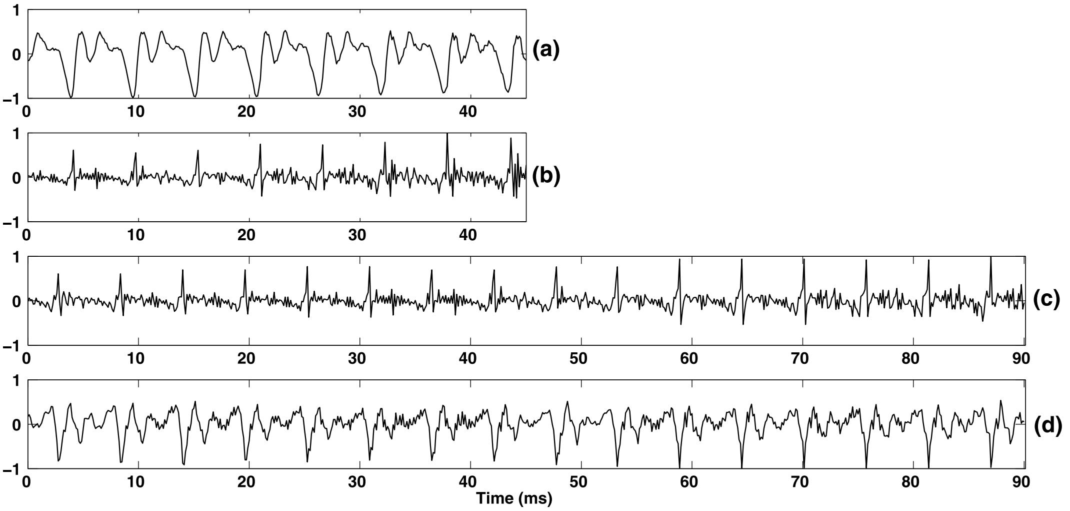

Fig. 9 Duration Modification: (**a**) Longer Segment of a voiced speech, (**b**) its LP residual, (**c**) modified LP residual by increasing the duration by 2 times and (**d**) reconstructed duration modified speech

Figure 9 plots the duration modification example by increasing the duration of the original speech segment by factor of 2. As we can observe that the overall duration of the signal is doubled compared to the original duration of the signal. Also it has to be observed that the pitch intervals in the duration modified speech remain unaltered as in original speech. Figure 9(c) has duration that is 2 times that of the LP residual segment of the original signal as shown in Fig. 9(b).

Some of the neutral to emotion converted samples for various emotions in German emotion speech database are available in the following link: http://www.iitg.ac.in/eee/emstlab/demos/demo4.php. The neutral to emotion conversion is achieved by residual prosody modification, where the prosodic parameters are fixed scaling factors. This demo is a part of the work done by the authors in Govind et al. (2011).

6.2 Methods to incorporate excitation parameters

The rules related to the excitation parameters like jitter, shimmer and glottal flow parameters have to be incorporated to effectively convey the expressive information in the synthesized speech. Cabral (2006) described one method to incorporate the jitter into the neutral speech by adding a random value to the pitch period. Here the time index of the synthesis pitch marks are randomly varied (according to jitter modification factor) for incorporating the voice quality that is related to jitter. Also the shimmer is incorporated by

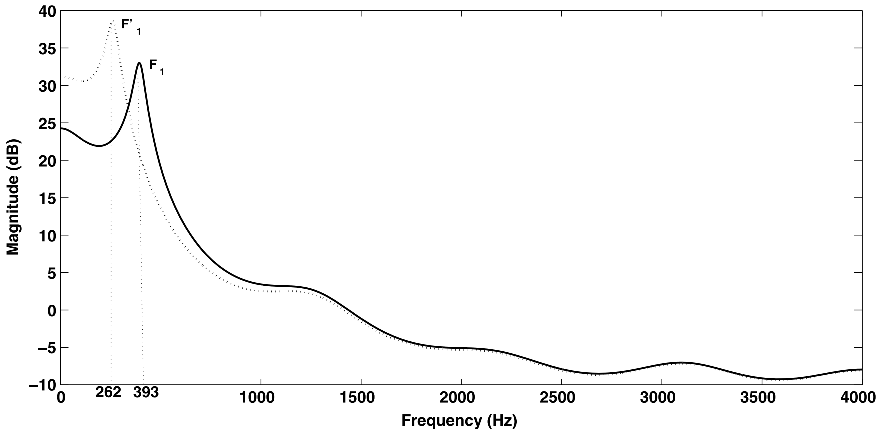

Fig. 10 The formant frequency modification: this plot demonstrate the shifting of the first formant (F_1) by a factor of 1.5 times the actual formants locations (*dotted plot*)

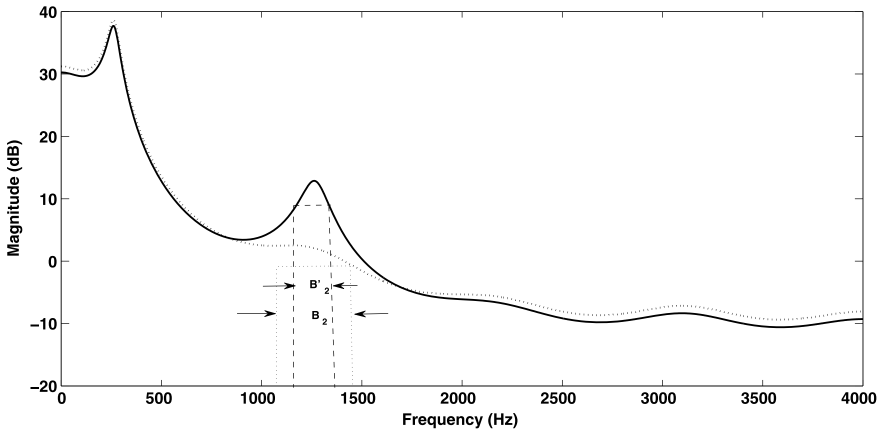

Fig. 11 The formant frequency modification: this plot demonstrate the bandwidth scaling corresponds to the second formant (F_2) by a factor of 0.25 times the actual formant bandwidth (*dotted plot*)

scaling the energy envelope of the short time signal by a random number (whose variation is according to shimmer modification factor). The glottal flow parameters like OQ, SQ and RQ can be modified by scaling the time indices used to estimate them (Campell et al. 2006).

Ruinskiy and Lavner (2008) brought another method to incorporate shimmer and jitter for simulating hoarse voice quality in a given speech. Here the jitter is introduced by relative stretching and shortening of pitch cycles. The jitter modification factor for the modification is retrieved from a jitter bank which stored the trends in the jitter values for every 2 to 4 pitch cycles. The residual samples in the given pitch cycle is resampled to modify jitter. Similarly, the shimmer modification factors are also stored across consecutive pitch cycles. The shimmer is introduced by multiplying each pitch cycle by a window function with varying peak amplitudes according to the shimmer factor (Ruinskiy and Lavner 2008).

6.3 Methods to incorporate vocal tract parameters

Even though there are little works discussed in the literature towards incorporating VT parameters for ESS, there are some works done towards voice conversion (Rao and Yegnanarayana 2006a; Rao 2010). Rao and Yegnanarayana (2006a) achieved vocal tract modification by the linear transformation of formant locations and bandwidth in the z-plane. The formant frequencies are modified by shifting the polar angle that the corresponding conjugate pair of poles make with respect to positive real axis of the z-plane. Formant frequency shifting of the LP spectrum is demonstrated in Fig. 10. In Fig. 10, the first formant of the modified LP spectrum, F_1', located at 393 Hz (shown by the thicker plot) is shifted 1.5 times the formant value of the original LP spectrum, F_1 which is at 262 Hz (shown by the dotted plot).

The formant bandwidth is modified by scaling the magnitude of the conjugate pair of poles for the corresponding formant. The bandwidth scaling corresponding to F_2 in the LP spectrum is demonstrated in Fig. 11. In Fig. 11, the bandwidth of F_2, B_2' scaled 0.25 times (thick plot) the original F_2 bandwidth, B_2 (dotted plot).

In a recent work on voice conversion, Rao (2010) achieved the VT modification by deriving mapping functions using a feed-forward neural networks (FFNN). Here, the line spectral frequencies (LSF) derived from LPCs are used to represent vocal tract characteristics. The mapping functions that represent the relation between the VT characteristics of source and target speakers are derived by feeding the time aligned LSFs of both source and target speakers to the FFNN using a database containing 500 Hindi utterances (Rao 2010). The dynamic time warping (DTW) is used to time align the LSFs of source and target speakers. Joseph et al. (2010) used multi-layer FFNN to map the vocal tract parameters of throat microphone to close speaking microphone to enhance throat microphone speech. Here, the objective was to bring a nonlinear relationship between source and system features of throat microphone data and speech from close speaking microphone for the enhancement of throat microphone data. The LP cepstral coefficients (LPCC) derived from LPCs are used to characterize the vocal tract parameters.

7 Evaluation of the synthesized expressions

Most of the works in ESS reviewed in this paper use subjective tests for evaluating the perceptual quality and level of expressiveness in the synthesized expressive speech. The ESS works described in the papers (Cahn 1989; Murray and Arnott 1995; Murray and Arnott 1993; Cabral and Oliveira 2006a) generate the stimuli for subjective test by pooling original neutral, original target and synthesized expressive speech. In these studies, subjects were asked to forcefully categorize into any of the expressive categories. Due to lack of reference expressive model (in terms of expressive contexts or reference prosody) for the subjects, the reliability of the perception results were questioned in van Santen et al. (2003). Following the perception experiments conducted in Johnson et al. (2002), House et al. (1999), Theune et al. (2006) used comparative subjective studies for evaluating the expressive fragments synthesized for story telling applications. Here, the subjects were presented with both original reference fragments and synthesized fragments. The subjects judged the quality of the synthesized story fragments by comparing the original neutral and original climax fragments of the story. The higher quality synthesized files were given a score of 5 and poor quality synthesized files were given a mean opinion score of 1. For neutral to emotion speech conversion applications, the synthesized emotions can be evaluated based on the comparison subjective tests. Where the subjects were asked to compare the synthesized expressive speech with the original neutral and target expressions to provide comparison mean opinion scores (CMOS). Higher score is given for a synthesized file if the synthesized file is close to the target expression and a low CMOS is obtained if synthesized speech sounds like the neutral speech. Steidl et al. proposed a method for evaluating the synthesized emotional speech by emotion identification. The emotion specific acoustic features used in the INTERSPEECH 2010 paralinguistic challenge baseline, are used to develop the emotion identification system Steidl et al. (2012). The level of expressiveness present in the synthesized expressive speech is determined by measuring emotion recognition rate of the synthesized expressions.

8 Summary of the works related to ESS by explicit control and discussions

As the unit selection based neutral speech synthesis provides more naturalness compared to statistical parametric speech synthesis systems, unit selection based speech synthesis systems can be used as NSS system for the ESS by explicit control. From the studies on expressive parameters, most of the works use prosodic parameters as common expressive parameters. Also, many of the works use the expression dependent excitation parameters and vocal tract parameters as the supplementary features with prosodic parameters. Hence it is necessary to accurately estimate and analyze these prosodic parameters for various expressions. It has been observed that almost all the studies use conventional methods for estimating prosodic parameters for expressive speech analysis. As expressions are characterized by the presence of prosodic variations that are much more than that of the neutral expressions, the accuracy of the estimated parameters using conventional methods have to be verified for different expressions. The recently developed ZFF method which provides best accurate estimates of various prosodic parameters as compared to other existing methods. This can be used for accurate speech analysis of various expressions. Most of the expressive speech systems use conventional PSOLA based methods to incorporate the expression specific prosodic variations for ESS. Since the epoch based prosody modification provides improved perceptual quality for moderate prosody modification factors than PSOLA based methods, epoch based prosody modification can be applied for ESS. The GD method of epochs estimation increases the computational complexity of the existing epoch based prosody modification. Hence more accurate and computationally faster ZFF epochs can be used instead of epochs estimated using GD method for epoch based prosody modification. Based on the review of incorporation of prosody parameters for ESS, the ESS achieved in most of the studies are by modifying the prosody parameters for fixed scale factors. However, fixed scaling of the prosodic parameters will not capture the dynamics of the prosody due to various expressions. Hence the prosody modification methods that incorporate the time varying dynamics of the prosodic parameters have to be used for ESS.

To evaluate the epoch estimation performance across various expressions, five expressions (Neutral, Angry, Happy, Boredom and Fear) of German emotional speech database having simultaneous EGG recordings are used. The epochs

Table 2 Epoch estimation performance of conventional ZFF algorithm for different emotional speech signals taken from the German database Burkhardt et al. (2005)

Emotion	IDR (%)	MR (%)	FAR (%)	IDA (ms)
ZFF				
Neutral	99.12	00.08	00.79	0.3194
Angry	87.93	00.41	11.66	0.4115
Happy	90.66	00.33	09.02	0.3858
Boredom	98.75	00.04	01.20	0.3495
Fear	94.90	00.13	04.97	0.2774

estimated from EGG recordings of these expressions, are used as the reference epochs location for epoch estimation performance analysis. The performance measures as used in Naylor et al. (2007) are used for evaluating the epoch estimation performance. The identification rate (IDR), miss rate (MR), false alarm rate (FAR) and identification accuracy (IDA) are the various measures used for evaluating epoch estimation performance. Table 2 shows the epoch estimation performance for each expression as given in Govind and Prasanna (2012). Here, we can observe that, even though the ZFF method gives a reliable epoch estimate for neutral speech, significantly degrades in the case of expressive speech. For instance, there is a significant drop in IDR for angry emotions as compared to neutral speech. The reason may be the rapid source and prosodic feature variations in emotional speech as compared to the neutral speech. Hence a refined epoch estimation method has to be proposed for reliable epoch estimation from emotional speech.

In another way, since the degradations in epoch estimation performance in emotional speech is due to effect of emotions ZFF signal. For instance, the increase in false alarm rate is due to additional zero crossing occurring in the ZFF signal which is treated as the epochs location. However, this additional zero crossing incurred in the ZFF signal can be studied to explores some emotion specific activities at the source level. Also, another epoch performance measure, IDA provides variation of the genuinely (without missing and false estimations) estimated epochs with respect to the actual location of reference epochs obtained from EGG. Here the IDA provides an absolute gross level variation of the estimated epochs with respect to the reference epochs. This difference in absolute gross level variation for each emotion can be observed in Table 2 for each emotion. Hence more careful analysis has to be done to track the variation in estimated epochs for various emotions. The IDA, if properly analyzed can be used as an objective measure for neutral to emotion conversion along with the conventional subjective evaluations for judging the level of expressiveness induced in the synthesized speech. The other expressive parameters like F_{0Avg}, sentence duration and strength of excitation have to be estimated and analyzed across various expressions. A dynamic prosody modification method has to be devised to incorporate dynamic variations of these parameters in the neutral speech to synthesize the speech in the target expression.

The unit selection based speech synthesis system can be used as NSS to synthesize neutral speech from the text. The proposed prosody modification method can then be used to incorporate the prosodic variations according to the target expression. Modification of the prosody parameters can be made according to the scale factors derived for the desired target expression in the expressive speech analysis stage. The resulting prosody modified speech will sound like from the target expression.

Acknowledgement The work done in this paper is funded by the on going UK-India Education Research Initiative (UKIERI) project titled “study of source features for speech synthesis and speaker recognition” between IIT Guwahati, IIIT Hyderabad and University of Edinburgh. The present work is also supported from the ongoing DIT funded project on the Development of text to speech synthesis systems in Assamese and Manipuri languages.

References

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, *50*, 637–655.

Badin, P., & Fant, G. (1984). *Notes on vocal tract computation* (Tech. rep.). STL-QPSR.

Banks, G. F., & Hoaglin, L. W. (1941). An experimental study of duration characteristics of voice during the expression of emotion. *Speech Monographs*, *8*, 85–90.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., & Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, *52*(5), 394–404.

Beautemps, D., Badin, P., & Bailly, G. (2001). Linear degrees of freedom in speech production: analysis of cineradio- and labio-films data for a reference subject, and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America*, *109*(5), 2165–2180.

Black, A. W., & Campbell, N. (1995). Optimising selection of units from speech database. In *Proc. EUROSPEECH*.

Bulut, M., & Narayanan, S. (2008). On the robustness of overall f0-only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America*, *123*, 4547–4558.

Burkhardt, F., & Sendilmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant synthesis. In *Proc. ISCA workshop on speech & emotion* (pp. 151–156).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlemeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proc. INTERSPEECH* (pp. 1517–1520).

Cabral, J. P. (2006). *Transforming prosody and voice quality to generate emotions in speech*. Master’s thesis, L2F-Spoken Language Systems Lab, Lisboa, Portugal.

Cabral, J. P., & Oliveira, L. C. (2006a). Emo voice: a system to generate emotions in speech. In *Proc. INTERSPEECH*.

Cabral, J. P., & Oliveira, L. C. (2006b). Pitch-synchronous time-scaling for prosodic and voice quality transformations. In *Proc. INTERSPEECH*.

Cabral, J. P., Renals, S., Yamagishi, J., & Richmond, K. (2011). Hmm-based speech synthesiser using the lf-model of the glottal source. In *Proc. ICASSP*.

Cahn, J. E. (1989). Generation of affect in synthesized speech. In *Proc. American voice I/O society*.

Campbell, N. (2004). Developments in corpus-based speech synthesis: approaching natural conversational speech. *IEICE Transactions*, *87*, 497–500.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 1171–1179.

Campell, N., Hamza, W., Hog, H., & Tao, J. (2006). Editorial special section on expressive speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 1097–1098.

Carlson, R., Sigvardson, T., & Sjölander, A. (2002). *Data-driven formant synthesis* (Tech. rep.). TMH-QPSR.

Clark, R. A. J., Richmond, K., & King, S. (2007). Multisyn: open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, *49*, 317–330.

Courbon, J. L., & Emerald, F. (1982). A text to speech machine by synthesis from diphones. In *Proc. ICASSP*.

Deller, J. R., Proakis, J. G., & Hanson, J. H. L. (1993). *Discrete-time processing of speech signals*. New York: McMillan.

Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). Emotions and voice quality: experiments with sinusoidal modeling. In *Proc. ITRW VOQUAL'03* (pp. 127–132).

Dudley, H. (1939). *The vocoder* (Tech. rep.). Bell laboratories.

Dunn, H. K. (1950). The calculation of vowel resonances, and an electrical vocal tract. *The Journal of the Acoustical Society of America*, *22*, 740–753.

Engwall, O. (1999). Modeling of the vocal tract in three dimensions. In *Proc. EUROSPEECH*.

Erickson, D. (2005). Expressive speech: production, perception and application to speech synthesis. *Acoustical Science and Technology*, *26*(4), 317–325.

Erickson, D., Schochi, T., Menezes, C., Kawahara, H., & Sakakibara, K.-I. (2008). Some non-f0 cues to emotional speech: an experiment with morphing. In *Proc. speech prosody* (pp. 677–680).

Fairbanks, G., & Hoaglin, L. W. (1939). An experimental study of pitch characteristics of voice during the expression of emotion. *Speech Monographs*, *6*, 87–104.

Fant, G. (1960). *Acoustic theory of speech production*. s-Gravenhage: Moutan & Co.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four parameter model of glottal flow. *STL-QPSR*, *26*(4), 1–13.

Farrus, M., & Hernando, J. (2009). Using jitter and shimmer in speaker verification. *IET Signal Processing*, *3*(4), 247–257.

Fernandez, R., & Ramabhadran, B. (2007). Automatic exploration of corpus specific properties for expressive text-to-speech: a case study in emphasis. In *Proc. ISCA workshop on speech synthesis* (pp. 34–39).

Gauffin, J., & Sundberge, J. (1978). Pharyngeal constrictions. *Phonetica*, *35*, 157–168.

Govind, D., & Prasanna, S. R. M. (2012). Epoch extraction from emotional speech. In *Proc. SPCOM*.

Govind, D., Prasanna, S. R. M., & Yegnanarayana, B. (2011). Neutral to target emotion conversion using source and suprasegmental information. In *Proc. INTERSPEECH*.

Hardam, E. (1990). High quality time scale modification of speech signals using fast synchronized overlap add algorithms. In *Proc. IEEE*.

Hashizawa, Y., Hamzah, S. T. M. D., & Ohyama, G. (2004). On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of the emotion. In *Proc. speech prosody* (pp. 655–658).

Heinz, J. M., & Stevens, K. N. (1964). On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *The Journal of the Acoustical Society of America*, *36*(5), 1037–1038.

Hess, W. (1983). *Pitch determination of speech signals*. Berlin: Springer.

Hofer, G., Richmond, K., & Clark, R. (2005). Informed blending of databases for emotional speech synthesis. In *Proc. INTERSPEECH*.

House, D., Bell, L., Gustafson, K., & Johansson, L. (1999). Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program. In *Proc. EUROSPEECH* (pp. 1843–1846).

Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP* (pp. 373–376).

Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M. (2000). A speech synthesis system for assisting communications. In *ISCA workshop on speech & emotion* (pp. 167–172).

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. ICASSP* (pp. 93–96).

Ishii, C. T., & Campbell, N. (2002). Analysis of acoustic-prosodic features of spontaneous expressive speech. In *Proc. Ist international congress of phonetics and phonology*, Kobe, Japan (pp. 85–88).

Jhonstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proc. int. congr. phoetic sciences*, San Fransisco (pp. 2029–2031).

Johnson, W. L., Narayanan, S., Whitney, R., Das, R., Bulut, M., & Labore, C. (2002). Limited domain synthesis of expressive military speech for animated characters. In *Proc. IEEE speech synthesis workshop*, Santa Monica, CA.

Joseph, M. A., Reddy, M. H., & Yegnanarayana, B. (2010). Speaker-dependent mapping of source and system features for enhancement of throat microphone speech. In *Proc. INTERSPEECH 2010* (pp. 985–988).

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proc. ICASSP* (pp. 1303–1306).

Kawahara, H., Masuda-Katsuse, I., & de Cheveigne, A. (1999). Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based f0 extraction: possible role of repetitive structure in sounds. *Speech Communication*, *27*(3–4), 187–207.

Kelly, J., & Lochbaum, C. (1962). Speech synthesis. In *Proc. international congress on acoustics*.

King, S. (2011). An introduction to statistical parametric speech synthesis. *Sadhana*, *36*(5), 837–852.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, *67*, 971–995.

Klatt, D. H. (1987). Review of text to speech conversion for English. *The Journal of the Acoustical Society of America*, *82*, 737–793.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). LDC emotional prosody speech transcripts database. University of Pennsylvania, Linguistic data consortium.

Ling, Z.-H., Richmond, K., & Yamagishi, J. (2011). Feature-space transform tying in unified acoustic-articulatory modelling of articulatory control of hmm-based speech synthesis. In *Proc. INTERSPEECH*.

Maeda, S. (1979). An articulatory model of the tongue based on a statistical analysis. *The Journal of the Acoustical Society of America*, *65*, S22–S22.

Makhoul, J. (1975). Linear prediction: a tutorial review. *Proceedings of the IEEE*, *63*, 561–580.

Markel, J. D. (1972). The sift algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics*, *AU-20*, 367–377.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, *53*, 1070–1082.

Miyanaga, K., Masuko, T., & Kobayashi, T. (2004). A style control techniques for hmm-based speech synthesis. In *Proc. ICSLP*.

Montero, J. M., Gutierrez-Arriola, J., Colas, J., Enriquez, E., & Pardo, J. M. (1999). Analysis and modelling of emotional speech in Spanish. In *Proc. ICPhS* (pp. 671–674).

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 452–467.

Mourlines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, *16*, 175–205.

Muralishankar, R., Ramakrishnan, A. G., & Prathibha, P. (2004). Modification pitch using dct in the source domain. *Speech Communication*, *42*, 143–154.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*, 1097–1108.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion by rule in synthetic speech. *Speech Communication*, *16*, 369–390.

Murthy, H. A., & Yegnanarayana, B. (1991). Formant extraction from group delay function. *Speech Communication*, *10*(3), 209–221.

Murthy, P. S., & Yegnanarayana, B. (1999). Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals. *IEEE Transactions on Speech and Audio Processing*, *7*(6), 609–619.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(8), 1602–1614.

Murty, K. S. R., & Yegnanarayana, B. (2009). Characterization of glottal activity from speech signals. *IEEE Signal Processing Letters*, *16*(6), 469–472.

Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *98*(3), 1325–1347.

Naylor, P. A., Kounoudes, A., Gudnason, J., & Brookes, M. (2007). Estimation of glottal closure instants in voiced speech using DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 34–43.

Nose, T., Yamagishi, J., & Kobayashi, T. (2007). A style control technique for hmm-based expressive speech synthesis. *IEICE Transactions on Information and Systems E*, *90-D*(9), 1406–1413.

Olive, J. P. (1977). Rule synthesis of speech from dyadic units. In *Proc. ICASSP*.

Palo, P. (2006). *A review of articulatory speech synthesis*. Master's thesis, Helsinki University of Technology.

Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., & Picheny, M. A. (2006). The ibm expressive text to speech synthesis system for American English. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 1099–1109.

Prasanna, S. R. M., & Yegnanarayana, B. (2004). Extraction of pitch in adverse conditions. In *Proc. ICASSP*, Montreal, Canada.

Rao, K. S. (2010). Voice conversion by mapping the speaker-specific features using pitch synchronous approach. *Computer Speech & Language*, *24*(3), 474–494.

Rao, K. S., & Yegnanarayana, B. (2003). Prosodic manipulation using instants of significant excitation. In *IEEE int. conf. multimedia and expo*.

Rao, K. S., & Yegnanarayana, B. (2006a). Voice conversion by prosody and vocal tract modification. In *Proc. ICIT*, Bhubaneswar.

Rao, K. S., & Yegnanarayana, B. (2006b). Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 972–980.

Rao, K. S., Prasanna, S. R. M., & Yegnanarayana, B. (2007). Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. *IEEE Signal Processing Letters*, *14*, 762–765.

Ross, M., Shaffer, H. L., Cohen, A., Freudberg, R., & Manley, H. J. (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-22*, 353–362.

Ruinskiy, D., & Lavner, Y. (2008). Stochastic models of pitch jitter and amplitude shimmer for voice modification. In *Proc. IEEE* (pp. 489–493).

Schafer, R. W., & Rabiner, L. R. (1970). System for automatic formant analysis of voiced speech. *The Journal of the Acoustical Society of America*, *47*, 634–648.

Scherer, K. R. (1986). Vocal affect expressions: a review and a model for future research. *Psychological Bulletin*, *99*, 143–165.

Schröder, M. (2001). Emotional speech synthesis—a review. In *Proc. EUROSPEECH* (pp. 561–564).

Schroder, M. (2009). Expressive speech synthesis: past, present and possible futures. *Affective Information Processing*, *2*, 111–126.

Smits, R., & Yegnanarayana, B. (1995). Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *4*, 325–333.

Steidl, S., Polzehl, T., Bunnell, H. T., Dou, Y., Muthukumar, P. K., Perry, D., Prahallad, K., Vaughn, C., Black, A. W., & Metze, F. (2012). Emotion identification for evaluation of synthesized emotional speech. In *Proc. speech prosody*.

Tachibana, M., Yamagishi, J., Masuko, T., & Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems E*, *88-D*(3), 1092–1099.

Tao, J., Kang, Y., & Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 1145–1154.

Taylor, P. (2009). *Text to speech synthesis*. Cambridge: Cambridge University Press.

Theune, M., Meijs, K., Heylen, D., & Ordelman, R. (2006). Generating expressive speech for story telling applications. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1099–1108.

Tokuda, K., Kobayashi, T., & Imai, S. (1995). Speech parameter generation from hmm using dynamic features. In *Proc. ICASSP* (pp. 660–663).

van Santen, J., Black, L., Cohen, G., Kain, A., Klabbers, E., Mishra, T., de Villiers, J., & Niu, X. (2003). Applications of computer generated expressive speech for communication disorders. In *Proc. EUROSPEECH* (pp. 1657–1660).

Vroomen, J., Collier, R., & Mozziconacci, S. J. L. (1993). Duration and intonation in emotional speech. In *Proc. EUROSPEECH* (pp. 577–580).

Whiteside, S. P. (1998). Simulated emotions: an acoustic study of voice and perturbation measures. In *Proc. ICSLP*, Sydney, Australia (pp. 699–703).

Williams, C. E., & Stevens, K. (1972). Emotions and speech: some acoustic correlates. *The Journal of the Acoustical Society of America*, *52*, 1238–1250.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2003). Modeling of various speaking styles and emotions for hmm-based speech synthesis. In *Proc. EUROSPEECH* (pp. 2461–2464).

Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., & Nakano, Y. (2007). Model adaptation approach to speech synthesis with diverse voices and styles. In *Proc. ICASSP* (pp. 1233–1236).

Yegnanarayana, B. (1978). Formant extraction from linear-prediction spectra. *The Journal of the Acoustical Society of America*, *63*(5), 1638–1641.

Yegnanarayana, B., & Murty, K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 614–625.

Yegnanarayana, B., & Veldhuis, R. N. J. (1998). Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, *6*(4), 313–327.

Yoshimura, T. (1999). *Simultaneous modeling of phonetic and prosodic parameters and characteristic conversion for hmm-based text-to-speech systems*. Ph.D. thesis, Nagoya Institute of Technology.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH*.

Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of nitech HMM-based speech synthesis system for the Blizzard challenge 2005. *IEICE Transactions on Information and Systems E*, *90-D*, 325–333.

Zen, H., Tokuda, K., & Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*, 1039–1064.

Zovato, E., Pacchiotti, A., Quazza, S., & Sandri, S. (2004). Towards emotional speech synthesis: a rule based approach. In *Proc. ISCA SSW5* (pp. 219–222).