Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 42 (2017) 38–58

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Generation of creaky voice for improving the quality of HMM-based speech synthesis

N.P. Narendra *, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

Received 4 January 2016; received in revised form 4 August 2016; accepted 10 August 2016
Available online 7 September 2016

Abstract

This paper aims at developing an HMM-based speech synthesis system capable of generating creaky voice in addition to modal voice. Generation of creaky voice is carried out by addressing two main issues, namely, an automatic prediction of creaky voice and appropriate modelling of the excitation signal of creaky voice. An automatic creaky voice detection method is proposed based on the analysis of variation of epoch parameters for different voicing regions. A neural network classifier is trained using the variances of epoch parameters for detection of creaky regions. A hybrid source model which is an extension of recently developed time-domain deterministic plus noise model is proposed for modelling creaky excitation signal. In the proposed hybrid source model, the pitch-synchronous analysis is performed on the creaky excitation signal of every phone. From the creaky residual frames of every phonetic class, the deterministic and noise components are estimated. The creaky deterministic components of all phonetic classes are stored in the database. The noise components are parameterized in terms of spectral and amplitude envelopes and are modelled by HMMs. During synthesis, the appropriate deterministic component is selected from the database, and the noise component is constructed from the parameters generated from HMMs. The creaky deterministic and noise components are pitch-synchronously overlap-added to produce the creaky excitation signal. Subjective evaluation results indicate that the incorporation of creaky voice has improved the naturalness of the synthetic speech of two male speakers, and the quality is slightly better than the basic time-domain deterministic plus noise model meant for only modal excitation.

Keywords: HMM-based speech synthesis; Detection of creaky voice; Zero-frequency filtering; Epoch parameters; Synthesis of creaky voice; Deterministic plus noise model; Hybrid source modelling

1. Introduction

HMM-based speech synthesis approach has gained a lot of attention in recent years due to its simplicity, flexibility, and reduced memory space (Tokuda et al., 2013). Presently, HMM-based speech synthesis system (HTS) has a lot of potential to be very useful in commercial applications. To produce good quality speech, HTS should be able to produce speech with different voice qualities such as modal and creaky voice. Generally, acoustic feature extraction and voice source modelling approaches followed in HTS are optimized to generate speech particular to modal phonation. But the synthesis of creaky voice in HTS is important as the creaky voice is frequently produced by the speakers used for developing text-to-speech synthesis. The creaky voice is produced by the speakers involuntarily, and the

* Corresponding author at: School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India.
E-mail address: narendrasince1987@gmail.com (N.P. Narendra), ksrao@iitkgp.ac.in (K. Sreenivasa Rao).

http://dx.doi.org/10.1016/j.csl.2016.08.002
0885-2308/

usage of creaky voice has been systematically observed in a variety of speech modes such as read, conversation, and expressive speech (Yanushevskaya et al., 2009). In American English (Kushan and Slifka, 2006) and Finnish (Silén et al., 2009) languages, the creaky voice has been observed as a phrase boundary marker. The creaky voice has been observed to be produced during hesitations (Carlson et al., 2006), and the presence of creaky voice is essential for communicating attitude and affective states (Yanushevskaya et al., 2005). On the whole, the usage of creaky voice at appropriate places will increase the naturalness of synthesized speech (Drugman et al., 2012a; Raitio et al., 2013).

The creaky voice produces dramatically different acoustic characteristics than that of modal voice. The most prominent acoustic features of the creaky voice include: (i) the interval between adjacent glottal flow pulses is long and as a result, little or no superposition of formant oscillations between successive glottal cycles, (ii) occurrence of secondary excitations and (iii) extremely long glottal closed phases (Blomgren et al., 1998; Gobl and Chasaide, 1992). As the acoustic characteristics of creaky and modal voice are significantly different, methods used for modelling the modal voice are not appropriate for modelling the creaky voice. Some of the major issues in the generation of creaky voice under HTS framework include, creaky voice detection, F0 estimation in creaky regions, prediction of creaky voice and source modelling of creaky voice. In this paper, we attempted to generate speech with creaky voice by addressing two issues, namely, creaky voice detection and source modelling of creaky voice.

Creaky voice detection algorithm identifies the regions in speech utterance containing creaky voice quality. In literature, there are very few approaches to automatically detect the creaky regions (Drugman et al., 2012b, 2014; Ishi et al., 2008; Ishi, 2004; Ishi et al., 2005; Kane et al., 2013; Vishnubhotla and Espy-Wilson, 2006), though several methods exist for identifying the broader class, i.e., irregular phonation (Surana and Slifka, 2006; Yoon et al., 2005). In Ishi (2004), the parameters based on autocorrelation of the glottal excitation waveform are extracted for identifying the creaky voice in spontaneous speech. In Vishnubhotla and Espy-Wilson (2006), an extension of the Aperiodicity, Periodicity and Pitch (APP) detector is proposed for automatic detection of irregular phonation including creaky voice. In the first step, irregular frames are separated from periodic frames using the periodicity measure of the APP detector. In the second step, using the dip profile of the Average Magnitude Difference Function (AMDF) in various frequency bands, creaky voiced regions are identified. Ishi et al. (2005, 2008) computed short term power and Intraframe Periodicity (IFP) strength contours from the speech signal for differentiating modal and creaky voiced regions. Interpulse similarity (IPS) measure is used to differentiate unvoiced and creaky regions. In Drugman et al. (2014) and Kane et al. (2013), a creaky voice detection method is proposed using two new acoustic features. The first feature exploits the occurrence of secondary peaks in the linear prediction (LP) residuals of creaky regions (Drugman et al., 2012b), and the second feature captures strong impulse-like peak and long glottal pulse duration properties of creaky regions. Even though several methods try to identify the creaky regions, still there is a necessity for the approach that can accurately detect creaky regions in the speech utterance.

For the generation of speech with creaky voice quality, efficient modelling of the creaky excitation signal is very much necessary. Most of the existing excitation or source modelling approaches are capable of modelling and generating the speech with modal voice quality (Drugman and Dutoit, 2012; Raitio et al., 2011; Zen et al., 2007). Very few researchers have attempted to develop a source model capable of producing speech with creaky voice quality (Csapó and Németh, 2014; Drugman et al., 2012a; Raitio et al., 2013). In Drugman et al. (2012a), a source model capable of producing creaky voice is developed by extending the deterministic plus stochastic model (DSM) of the residual signal (Drugman and Dutoit, 2012). In DSM, the deterministic component is the first eigenvector obtained by Principal Component Analysis (PCA) of residual frames. The stochastic component is obtained by imposing the speaker specific spectrum and amplitude envelopes on white Gaussian noise. In the extended version of DSM, the deterministic and stochastic components are extracted separately from open and closed periods of creaky residual frames. Raitio et al. (2013) synthesized creaky voice by utilizing the GlottHMM F0 tracker (Raitio et al., 2011) and the extension of the DSM-based source model. In Csapó and Németh (2014), two methods are proposed to generate the creaky voice. The first one is a rule-based method which applies pitch halving and amplitude scaling of pitch-synchronous residual frames with random factors to generate creaky excitation. The second one is a data-driven approach where a database of creaky residual frames is developed and during synthesis appropriate creaky residual frames are chosen from the database using unit selection method.

This paper deals with the development of HMM-based speech synthesis system capable of generating the creaky voice. Two main issues involved in the synthesis of creaky voice, namely, detection of creaky voice and source modelling of creaky voice are addressed. An automatic creaky voice detection method is proposed by analysing the variation of epoch parameters for different voicing regions. The epoch parameters are extracted from the speech signal using zero-frequency filtering method. Using the variance of epoch parameters as input parameters, a neural network

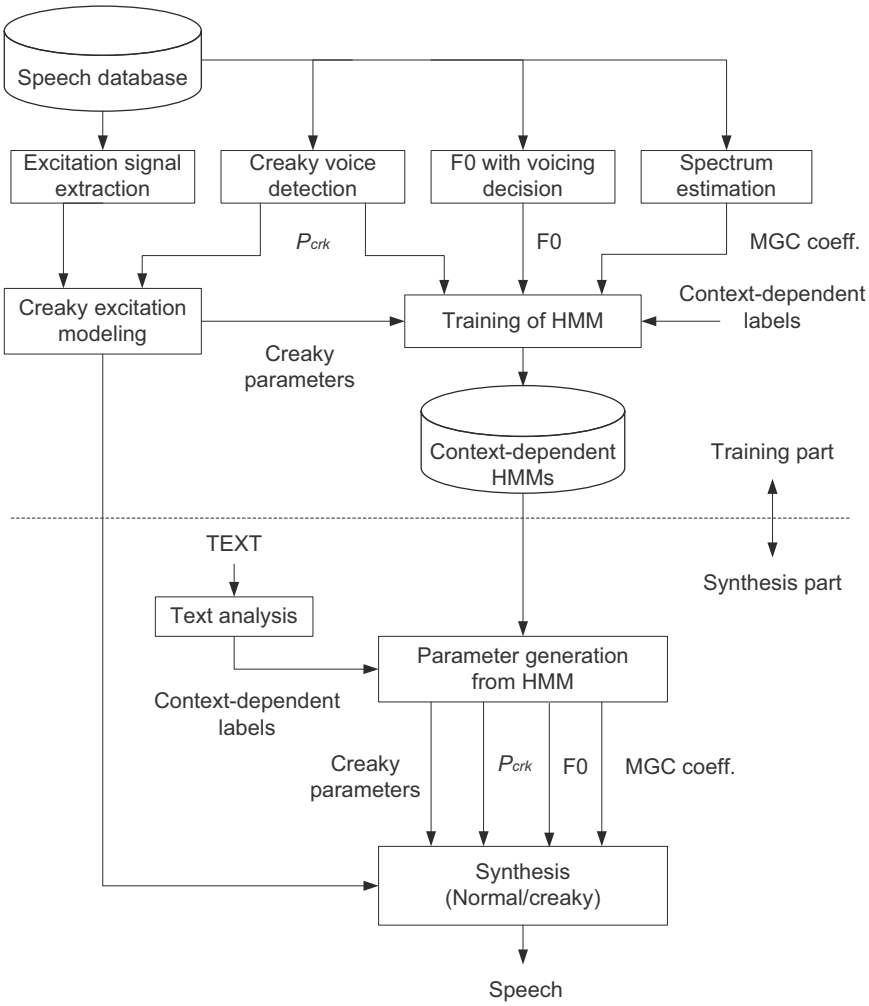

Fig. 1. Block diagram of HMM-based speech synthesis system including additional modules to generate creaky voice.

classifier is trained for identifying the creaky regions. The hybrid source method is proposed for generating creaky excitation signal specific to every phone. The proposed hybrid source model is an extension of recently developed time-domain deterministic plus noise model (Narendra and Rao, 2016). In the proposed hybrid source model, creaky excitation is generated as a combination of creaky deterministic and noise components. Both creaky voice detection method and hybrid source model are incorporated in HTS to synthesize speech with creaky voice.

This paper is organized as follows. Section 2 provides the overview of HMM-based speech synthesis system optimized for generation of speech with creaky voice. In Section 3, the proposed creaky voice detection method is described. The description of the proposed hybrid source model for generating creaky excitation signal is provided Section 4. Finally, Section 5 provides the summary of the contributions of the paper.

2. Overview of HMM-based speech synthesis system

The block diagram of HMM-based speech synthesis system including additional modules to generate creaky voice is shown in Fig. 1. A speech database that is frequently producing creaky voice is considered for developing HTS. From the speech database, creaky regions are identified using the proposed creaky voice detection method (described in Section 3). The proposed creaky voice detection method outputs a creaky probability (P_{crk}) for every frame of speech. For every speech utterance present in the database, the corresponding excitation signal is obtained. Using creaky probability

P_{crk}, creaky regions are identified in the excitation signal. The excitation signal of normal voice (includes modal voiced and unvoiced) is modelled using recently proposed time-domain deterministic plus noise model based hybrid source model (Narendra and Rao, 2016). The excitation signal of creaky voice is modelled using the proposed hybrid source model that is an extension of basic time-domain deterministic plus noise model (explained in Section 4). The proposed hybrid source model extracts creaky parameters and stores the real segments of creaky residual frames in the database. F0 estimation and voicing decision are performed on every utterance present in the speech database. In creaky voice, due to low F0 and highly irregular periodicity, F0 estimation methods either output spurious F0 values or incorrectly determine the region to be unvoiced. In this work, F0 estimation and voicing decision are performed using recently proposed method based on the strength of instants of significant excitation (Narendra and Rao, 2015a). This method is shown to efficiently identify both modal and creaky regions as voiced and extract accurate F0 values. Mel-Generalized Cepstrum (MGC) parameters which represent the spectrum of speech are extracted from every utterance. As suggested in Zen et al. (2008), 34th order MGC coefficients are extracted with the parameter values $\alpha = 0.42$ (Sampling frequency $Fs = 16$ kHz) and $\gamma = -1/3$. The MGC coefficients, F0 with voicing decision, creaky probability, and creaky parameters are modelled using context-dependent HMMs.

During synthesis, input text is converted into a sequence of context-dependent phoneme labels. The contextual information includes a standard list of 53 positional and contextual features provided in the basic HTS toolkit (HMM-based speech synthesis system (HTS)). According to the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs. Then, a sequence of parameters is generated from the sentence HMM. To alleviate the problem of over-smoothing of generated parameters due to statistical averaging, the global variance technique is used (Toda and Tokuda, 2007). Depending on the generated creaky probability, the excitation signal is constructed separately for modal and creaky regions. The excitation signal of creaky regions is constructed from the generated creaky parameters and real segments of residual frames stored in the database. Finally, speech waveform is synthesized using the generated MGC coefficients and the excitation signal.

3. Automatic detection of creaky voice

To generate appropriate creaky excitation, it is essential to have automatic annotation of creaky regions in a given speech corpus. In this section, a description of proposed creaky voice detection method based on epoch parameters is presented (Narendra and Rao, 2015b). First, the epoch parameters are extracted from the speech signal using zero-frequency filtering (ZFF) method (Murty and Yegnanarayana, 2008; Murty et al., 2009). Then, the epoch parameters characterizing the source of excitation are analysed in modal and creaky regions. Using the variance of epoch parameters as input features, a neural network classifier is trained to detect the creaky regions. Fig. 2 provides the block diagram of the proposed creaky voice detection method.

3.1. Zero-frequency filtering method for extracting epoch parameters

In this work, epoch parameters that include the number of epochs in a frame, the strength of excitation of epochs and the epoch interval between successive epochs are extracted by using zero-frequency filtering method (Murty and Yegnanarayana, 2008). To extract the epoch parameters, the information regarding epoch locations should be

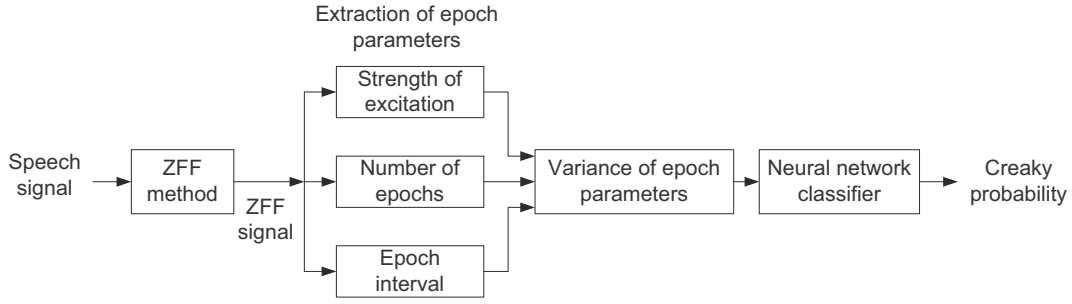

Fig. 2. Block diagram of the proposed creaky voice detection method.

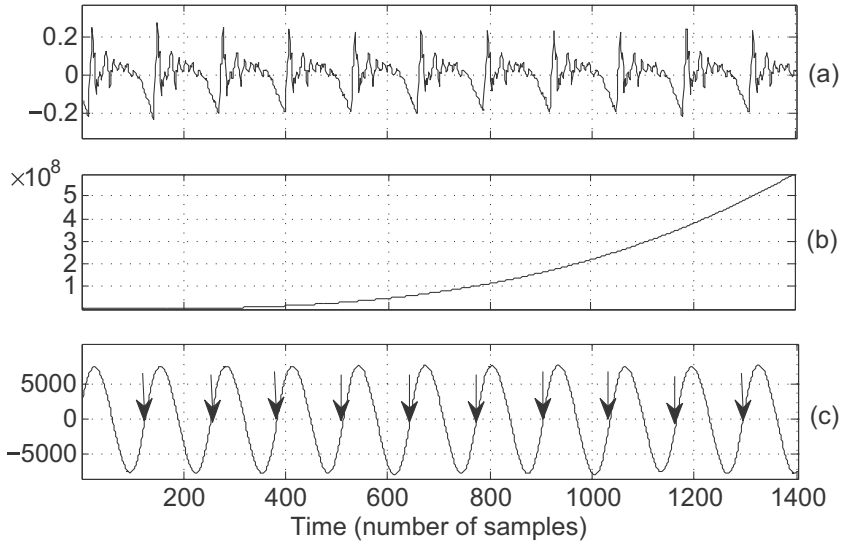

Fig. 3. (a) Speech signal (Fs = 16 kHz). (b) Zero-frequency resonator output. (c) Zero-frequency filtered signal.

obtained. The epoch locations correspond to the instants of impulse-like discontinuities in the excitation signal. The epoch locations are identified from the speech signal by using ZFF method. In ZFF method (Murty and Yegnanarayana, 2008), it is observed that the discontinuity due to impulse-like excitation is reflected across all frequencies including zero-frequency and the resonances due to vocal-tract system are present at frequencies greater than 300 Hz. By designing a resonator at zero frequency, the information around zero frequency is greatly emphasized compared to the vocal tract resonances. As a result, the information regarding impulse-like excitation can be easily obtained from the resulting zero-frequency resonator output. The system function of a zero-frequency resonator is given by

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{1}$$

where $a_1 = -2$ and $a_2 = 1$. The above resonator de-emphasizes the characteristics of vocal tract system. A cascade of two such resonators, given by system function $G(z) = H(z)H(z)$, is used to significantly de-emphasize all the resonances of vocal tract system relative to zero-frequency. Let $s[n]$ denote the input speech signal. The output of cascade of two resonators is given by $x_s[n] = s[n] * g[n]$. The output of zero-frequency resonator $x_s[n]$ decays or grows as a polynomial function of time. Hence, it is difficult to directly detect the effect of discontinuities due to impulse excitation in the filtered output. The characteristics of discontinuities due to impulse excitation are extracted by computing the deviation between zero-frequency filtered output and the local mean. The window length used for computing the local mean is chosen to be around average pitch period. The resulting signal obtained after subtracting the local mean is called zero-frequency filtered signal (Murty and Yegnanarayana, 2008) and is given by

$$y[n] = x_s[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} x_s[n+m] \tag{2}$$

where $2N + 1$ represents the length of window in terms of the number of samples. Fig. 3(a), (b) and (c) shows the segment of speech signal (Sampling frequency Fs = 16 kHz), zero-frequency resonator (ZFR) output and zero-frequency filtered signal, respectively. The locations of epochs are marked by downward arrows in Fig. 3(c). The time instants of negative to positive zero crossings of the ZFF signal are called as the instants of significant excitation or epochs. The strength of excitation is computed as the slope of ZFF signal at each epoch location (Murty et al., 2009). The strength of excitation indicates the rate of closure of the vocal folds in each glottal cycle (Alku et al., 2002). Sharper closure of the vocal folds corresponds to stronger excitation to the vocal-tract system. Epoch interval is computed as the time duration of successive epochs.

3.2. Analysis of epoch parameters

Proposed creaky voice detection is based on the variation of epoch parameters, namely, the number of epochs, the strength of excitation of epochs and the epoch intervals for different voicing regions. The epoch parameters vary with the size of the window used for local mean subtraction. In Narendra and Rao (2015a), the variation of the strength of excitation for different window sizes is analysed for detecting voiced and unvoiced speech regions. In this approach, the variation of epoch parameters for different window sizes is systematically examined for creaky voice detection.

Fig. 4 shows the strength of excitation ((a),(b),(c),(d)), epoch interval ((f),(g),(h),(i)) and number of epochs ((j),(k),(l),(m)) computed from the speech signal (Fs = 16 kHz) with window sizes of 8, 10, 12 and 14 ms for the BDL speaker of CMU Arctic database. The window size is varied in the range of 1.5–2 times the average pitch period of the speaker. The average pitch period of BDL speaker is 5.91 ms. The reason for choosing the above range of window size is provided in detail in Section 3.5. With this range of window size, a high distinction of epoch parameters is observed for modal and creaky regions. The speech signal shown in Fig. 4(e) contains three types of regions, namely, unvoiced, modal and creaky voiced regions. In unvoiced regions, vocal folds do not vibrate and there is no impulse-like excitation. As a result, in unvoiced regions the epochs are located at random instants and the strength of excitation is very low for different window sizes. For modal regions, vocal folds vibration rate is slowly varying and the most significant impulse-like excitation occurs during glottal closure instant (GCI). Hence, in modal regions the epochs are located at regular intervals and the strength of excitation is high for different window sizes. In creaky regions, the vocal folds vibration rate is low and irregular, and in addition to GCI, the impulse-like excitation can occur at the glottal opening instant, following long glottal closed phase. The secondary excitation at glottal opening instant may not occur for all kinds of speakers and styles of creaky voice. The two epochs which provide the locations of glottal opening and closing instants do not occur at equal intervals in the glottal cycles. Hence, the successive epoch intervals in the creaky regions are not equal. The strength of excitation in creaky regions is higher than unvoiced regions but lower than modal regions. On careful analysis of epoch parameters for modal and creaky regions with different window sizes, three main observations can be drawn.

1. **Strength of excitation (Fig. 4(a),(b),(c),(d)):** In creaky regions, in addition to GCI, secondary excitation is also present within a single glottal cycle. The strength of excitation at GCI is relatively high compared to secondary excitation. As a result for different window sizes, the strength of excitation in successive epochs is varying abruptly. In modal regions, slow variation in the strength of excitation can be observed across different window sizes.
2. **Epoch interval (Fig. 4(f),(g),(h),(i)):** In modal regions, successive epoch intervals are almost equal or vary slowly. For different window sizes, variation in epoch intervals is not significant. In creaky regions, due to the presence of secondary excitation, successive epoch intervals are unequal. Hence, the epoch intervals vary significantly for different window sizes.
3. **Number of epochs (Fig. 4(j),(k),(l),(m)):** In creaky regions, the secondary excitations having very low strength are detected for lower window sizes and missed for higher window sizes. As a result, the number of epochs in a frame varies for different window sizes. In modal regions, the secondary excitations are not present. Hence, for different window sizes, the number of epochs in modal voiced regions does not vary significantly.

From the above observations, we can conclude that for different window sizes, the epoch parameters vary significantly in creaky regions compared to modal regions.

3.3. Computation of variance of epoch parameters

To differentiate modal and creaky regions, the variance of epoch parameters is computed for every frame of speech. The procedure for finding the variance of epoch parameters is as follows.

(1) **Variance of the strength of excitation:** For every frame, the strength of excitation obtained from every window size is normalized between 0 and 1 and its variance is calculated. The final variance of the strength of excitation (V_{SE}) of a frame is obtained from the average of variances computed for different window sizes. Variance of the strength of excitation (v_{SEi}) for the i^{th} window size is given by

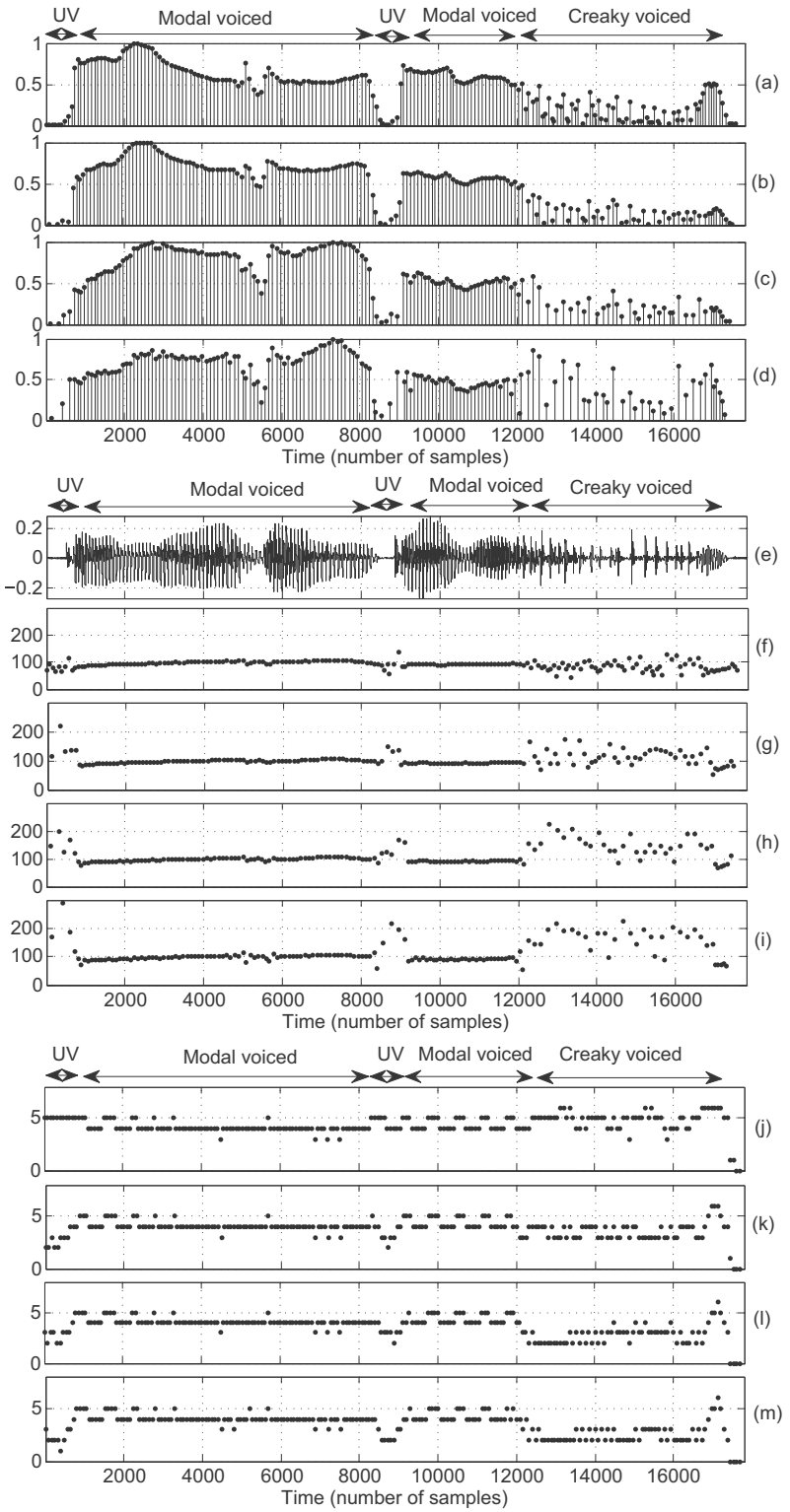

Fig. 4. Illustration of variation of epoch parameters in different voicing regions with different window sizes. Strength of excitation ((a),(b),(c),(d)), epoch interval ((f),(g),(h),(i)) and number of epochs ((j),(k),(l),(m)) computed from the speech signal (Fs = 16 kHz) ((e)) with window sizes of 8, 10, 12 and 14 ms.

$$v_{SEi} = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{SEij} - \mu_{SEi} \right)^2 \tag{3}$$

where N_i is the total number of epochs detected for the i^{th} window size, μ_{SEi} is the mean of strength of excitation computed by considering the strength of excitation of i^{th} window sizes and x_{SEij} is the normalized strength of excitation at j^{th} epoch for the i^{th} window size. The final variance of the strength of excitation (V_{SE}) of a frame is given by (M is the total number of window sizes considered):

$$V_{SE} = \frac{1}{M} \sum_{i=1}^{M} v_{SEi} \tag{4}$$

(2) **Variance of epoch interval:** For every frame, epoch intervals are collected from different window sizes. From the epoch intervals obtained from different window sizes, the variance of epoch interval is computed. Variance of epoch interval (V_{EI}) is given by

$$V_{EI} = \frac{1}{N_{ei} - 1} \sum_{i=1}^{N_{ei}} \left(x_{EIi} - \mu_{EI} \right)^2 \tag{5}$$

where N_{ei} is the total number of epoch intervals of all window sizes, μ_{EI} is the mean epoch interval computed by considering the epoch intervals of all window sizes and x_{EIi} is the i^{th} epoch interval. Here, epoch intervals of all window sizes are considered together for the calculation of variance.

(3) **Variance of the number of epochs:** For every frame, by considering the number of epochs obtained from different window sizes, the variance of the number of epochs is computed. Variance of the number of epochs (V_{NE}) is given by

$$V_{NE} = \frac{1}{M - 1} \sum_{i=1}^{M} \left(x_{NEi} - \mu_{NE} \right)^2 \tag{6}$$

where μ_{NE} is the mean number of epochs obtained by considering the number of epochs of all window sizes and x_{NEi} is the number of epochs calculated for the i^{th} window size. Here, for every window size, only one value of number of epochs is obtained.

All three variances determined from the speech utterance are normalized between 0 and 1. Fig. 5 shows the speech signal (Fs = 16 kHz) and the variances of the strength of excitation, the epoch interval and the number of epochs computed with the window sizes varying from 8 to 14 ms in steps of 1 ms. The procedure for choosing the optimum range of window size is detailed in Section 3.5. From the figure, it can be observed that the variances have high values in creaky and unvoiced regions, and zero or very low values in modal regions. Using voicing detection method at the first stage, unvoiced regions are removed, and epoch parameters are extracted only in voiced regions consisting of modal and creaky regions. For voicing detection, recently proposed method based on the strength of instants of significant excitation (Narendra and Rao, 2015a) is used. In this method, the strength of excitation is efficiently computed by using ZFF method. Here, it is observed that the strength of excitation computed by using ZFF method varies with the window size used for calculating the local mean. For a particular window size, the strength of excitation is observed to be maximum. By choosing the appropriate window size, the strength of excitation is computed from every frame of speech signal. The strength of excitation computed with the optimal window size is high for modal and creaky voiced regions and low for unvoiced regions. Finally, by setting a threshold (emperically determined) on the strength of excitation, the voicing detection is carried out. This method is robust enough to detect both modal and creaky regions as voiced.

3.4. Classification using variance of epoch parameters

Creaky/non-creaky classification can be performed by applying a threshold on the variance of epoch parameters. Instead of using threshold method, neural network classifier is used. The classifier is configured as a feed forward

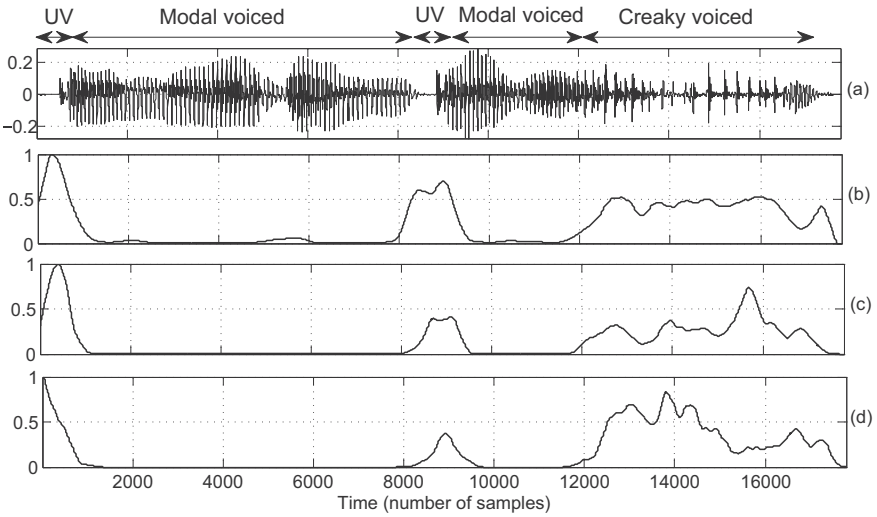

Fig. 5. (a) Speech signal ($Fs = 16$ kHz). Variances of (b) strength of excitation, (c) epoch interval and (d) number of epochs computed with the window sizes varying from 8 to 14 ms in steps of 1 ms.

network consisting of a single hidden layer. All neurons (fixed to 16 in this work) present in the hidden layer utilize a *tanh* transfer function. The output layer consists of a single neuron with a logarithmic sigmoid function suited for a binary decision. The training is performed using a standard error back-propagation algorithm (Bishop, 2006). In this study, the output of neural network classifier is considered as the creaky probability P_{crk}.

For the neural network classifier, the variance of epoch parameters extracted from every speech frame are given as input. The output of neural network classifier is the creaky probability P_{crk}. If P_{crk} is greater than a certain threshold value (α), then the speech frame is considered as creaky, else it is considered as non-creaky.

3.5. Performance evaluation

To evaluate the performance of the proposed method, different speech databases were considered. First, four databases including an American English male speaker (BDL (CMU ARCTIC speech synthesis databases, [Online])), an Indian English male speaker (KSP (CMU ARCTIC speech synthesis databases, [Online])), a Finnish female speaker (HS (Silén et al., 2007)) and a Finnish male speaker (MV (Vainio, 2001)) were considered. All four databases were developed to build text-to-speech synthesis systems. A creaky database was developed in two Indian languages, namely, Hindi and Bengali. In each language, single female and male speakers (native voice talents) were used for recording the speech corpus. For both Hindi and Bengali, the text corpus consists of 100 sentences obtained from children stories. The speakers used for developing the creaky database were research scholars working in the area of speech processing. Before recording, they were briefed about the differences between modal and creaky voice. The speakers did not produce the creaky voice in normal speech. Speakers were asked to utter the sentences in news reading style and were asked to intentionally produce creaky regions at the end of the utterances. The similar methodology of creaky database collection is followed in the previous literature (Silén et al., 2009). As the speech was carefully recorded by the speech experts, unnaturalness was not noticed in the utterances. 100 speech utterances from each of the BDL, KSP, HS, MV, Hindi (H-F1 and H-M1) and Bengali (B-F1 and B-M1) speakers were used in the evaluation. Initially, manual annotation of creaky regions was performed on the speech utterances of every speaker. Manual annotation of creaky regions was performed based on the auditory criterion "a rough quality with the additional sensation of repeating impulses" (Ishi et al., 2008). Also, an inspection of waveforms, spectrograms and F0 contours was also performed to ensure correct annotation. The utterances are arranged in the ascending order of the percentage of creaky regions. The 100 sentences which have the highest percentage of creaky regions are selected. The 100 sentences from every speaker are used for the evaluation of creaky detection methods. All speech utterances were downsampled to a sampling frequency of 16 kHz. Table 1 provides the summary of speech data used for evaluation of creaky detection method. The manual annotation of creaky and non-creaky regions in the speech utterances was utilized for evaluation of creaky detection methods.

Table 1
Summary of speech data used for evaluation of creaky detection method.

Speaker ID	Gender	Language	Creak (%)	Creak duration (s)
BDL	Male	US English	7.6	19.40
KSP	Male	Indian English	2.7	8.34
HS	Female	Finnish	5.8	36.86
MV	Male	Finnish	7.1	31.56
H-F1	Female	Hindi	9.8	24.81
H-M1	Male	Hindi	11.1	26.42
B-F1	Female	Bengali	10.1	20.78
B-M1	Male	Bengali	12.8	22.18

To assess the performance of the proposed method, three standard frame level metrics are used, namely, True Positive Rate (TPR, also called recall), False Positive Rate (FPR), and F1 score. TPR is the proportion of actual creaky frames that are correctly identified. FPR is the percentage of actual non-creaky frames that are wrongly detected as creaky. F1 score is a single metric (bound between 0 and 1) computed using true positives, false positives and false negatives. If the technique is better, then TPR and F1 score are higher, and FPR is lower.

In the neural network classifier, for a given input example, the output is the creaky probability, P_{crk}. The standard binary decision is class 1 (i.e., creaky) if $P_{crk} > \alpha$ (otherwise class 0) and α is typically set to 0.5. For skewed data sets (e.g., creak or laughter) which consists of sparse occurrence of a given class to be detected, this setting may not be optimal. Hence, α is varied in the range [0, 1] and set to the value that maximizes the F1 score on the training set. The threshold setting had very low inter-database sensitivity, as all speakers had their best F1 score for α in the vicinity of 0.3. This kind of optimal threshold setting was followed in Drugman et al. (2014) and Kane et al. (2013). Post processing is carried out on the binary decision (creaky or non-creaky) of the classifier. The detection of creaky regions in very short regions is removed and nearby adjacent creaky regions are merged by performing a 5-point median filtering to the binary decision. For evaluation purpose, the epoch parameters are extracted for a frame length of 32 ms and a frame shift of 10 ms. With a frame length of 32 ms, at least two creaky periods will be present in a frame (assuming the lowest F_0 value of creaky voice is 62.5 Hz). The presence of at least two creaky periods is essential for the computation of epoch parameters. In most of the previous literature on creaky voice detection (Drugman et al., 2012b; Ishi et al., 2008), the evaluation is performed by considering the frame length of 32 ms and a frame shift of 10 ms.

Optimum range of window size: In Fig. 5, epoch parameters are computed by varying the window size from 8 to 14 ms in steps of 1 ms. With this range of window size, a high distinction of epoch parameters is observed for modal and creaky regions. For different speakers, we need to find the optimum range of window size, which results in a higher F1 score and hence better creaky detection. First, by varying the window size from 1 ms to 20 ms in steps of 1 ms, epoch parameters are computed. With one window size as the centre, the epoch parameters of centre window size and the epoch parameters of two window sizes before and after the centre window size (similar to 5-gram model) are considered. For example, if the centre window size is 8 ms, then the epoch parameters of 8, 6, 7, 9 and 10 ms are considered. From the epoch parameters, variances are computed. Using the variance of epoch parameters, a neural network classifier is trained, and the F1 score is computed. Similarly, with every window size as the centre, the variance of epoch parameters are computed and subsequently F1 scores are determined. F1 scores obtained with different window sizes as the centre for speakers BDL and HS are shown in Fig. 6. Average pitch periods of speakers BDL and HS are 5.91 ms and 5.11 ms, respectively. From the figure, we can observe that for a window size range of approximately 1.5–2 times pitch period of the speaker, F1 score is having high values for both BDL and HS speakers. Similar kind of F1 score curves are obtained for other speakers also. Hence in this work, for computing the variance of epoch parameters, the window size is varied from 1.5 to 2 times pitch period of the speaker in steps of 1 ms. Here, instead of considering two window sizes before and after the centre window size, we tried increasing or decreasing the number of window sizes. Increasing the number of window sizes makes the F1 score curve smoother and decreasing the number of window size results in sudden fluctuations in the F1 score curve.

Evaluation of the detection performance was carried out using a leave one speaker out strategy, where the speech data of a given speaker was held out for testing, and the remaining speech data of all speakers was used for training. This procedure was repeated for each speaker. The proposed method is compared with two existing techniques:

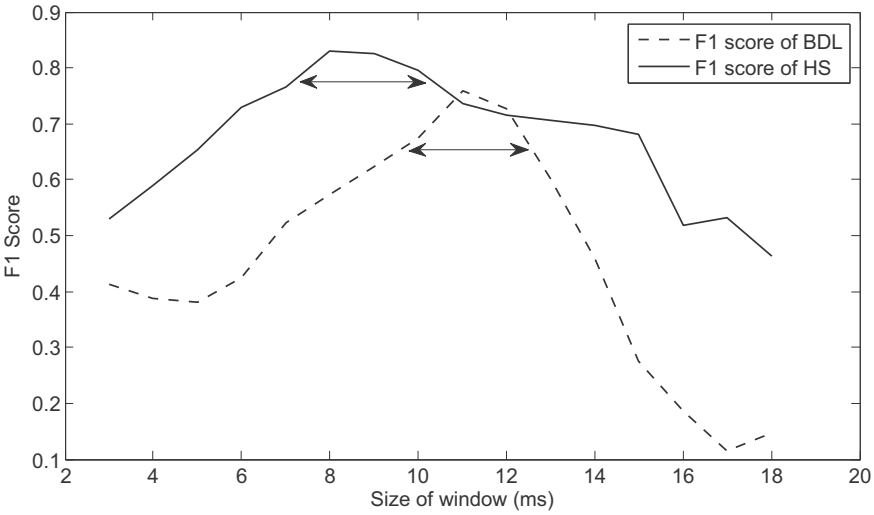

Fig. 6. F1 scores obtained with different window sizes as the centre for speakers BDL and HS. The double arrow indicates the approximate range of 1.5 to 2 times pitch period of the speaker, where superior performance is observed.

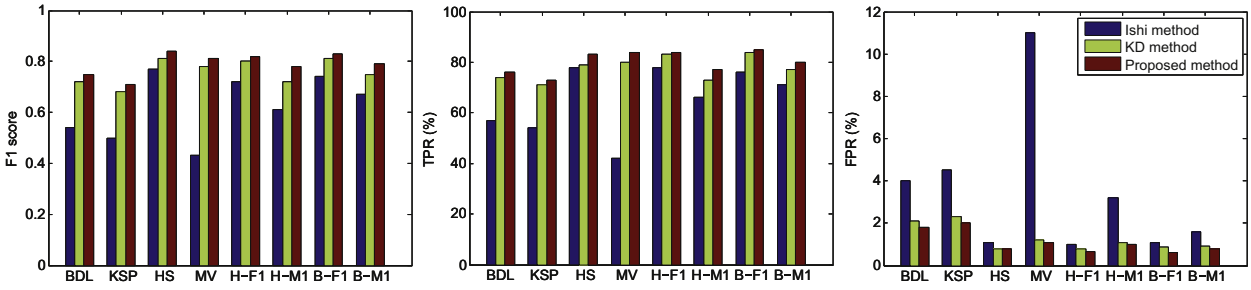

Fig. 7. F1 scores (left), TPR (middle) and FPR (left) values obtained for three creaky detection algorithms on BDL, HS, MV, H-M1, H-F1, B-M1 and B-F1 databases.

(i) **Ishi's method** (Ishi et al., 2008)**:** In this method, first short term power is computed from the speech signal which is bandlimited to 100–1500 Hz. By setting threshold on the short term power, the candidate regions are identified. Then, frame-synchronized periodicity strength measure is computed to discriminate creaky regions from normal voiced regions. Inter-pulse similarity measure is calculated to differentiate between creaky region from non-speech and unvoiced regions. In Ishi's method, the same settings as described in the original publications were used for all speakers (Ishi et al., 2008).

(ii) **Kane–Drugman (KD) method** (Kane et al., 2013)**:** Creaky detection is performed by proposing two new accoustic features from LP residual signal. The first feature exploits the property of occurrence of secondary peaks in creaky regions and the second feature characterizes the presence of strong impulse-like peaks in the residual signal of creaky voice. Two new accoustic features are used train the neural network classifier for detecting the creaky regions. The neural network is configured in the same way as described in Section 3.4. The source codes of Ishi's and KD methods are obtained from Voice Analysis Toolkit (Kane,).

F1 score, TPR and FPR obtained for different speech databases are shown in Fig. 7. From the figure, it can be observed that the proposed method using epoch parameters performs better than the two existing methods, across all databases. Among all results, Ishi's method displays the lowest TPR and F1 score for MV speaker. The main reason for this is that MV speaker has modal regions at low frequency. Intraframe periodicity values extracted from Ishi's method dropped below a threshold value in low-voiced regions. Hence, the modal regions having low-frequency regions were wrongly identified as creaky regions (evident by high values of FPR). Proposed method produced high F1 score

and TPR of MV speaker, as the extracted variance of epoch parameters were independent of the pitch of the speech signal. KD-method performed better than Ishi's method for all speech databases, but its performance is inferior, compared to the proposed method. One-way ANOVA is carried out to investigate whether the performance of proposed creaky detection method is significantly better than the two existing methods. Here, F1 score is treated as the dependent variable and detection method as the independent variable. One-way ANOVA indicated that the creaky detection method had a significant effect on the F1 score [$F = 16.0$, $p < 0.001$] and pair-wise comparisons carried out using Tukeys Honestly Significant Difference (HSD) test showed that the proposed method gave significantly higher F1 scores than both Ishi's method ($p < 0.001$) and KD method ($p < 0.01$).

4. Hybrid source model for generating creaky excitation

In this section, the proposed hybrid source model capable of generating creaky excitation signal is described. First, the basis of the proposed method i.e., the time-domain deterministic plus noise model based hybrid source model (Narendra and Rao, 2016), is described. Time-domain deterministic plus noise model based hybrid source model is a recently proposed method for accurate modelling and generation of excitation signal. In time-domain deterministic plus noise model, the pitch-synchronous residual frames extracted from the excitation signal of a phone are viewed as a combination of deterministic and noise components. The pitch-synchronous residual frame ($e_i(t)$) of i^{th} cycle of a phone is represented as follows:

$$e_i(t) = p(t) + r_i(t) \tag{7}$$

where $p(t)$ and $r_i(t)$ are the deterministic and the noise components, respectively. In this work, it is assumed that the deterministic component is constant for all pitch cycles of a phone and the noise component varies for every pitch cycle of a phone. The duration of all pitch-synchronous residual frames in a phone are not exactly equal, and there exists variation between one cycle to another. Hence, before estimating the deterministic component, the pitch-synchronous residual frames of a phone are normalized to the maximum pitch period of the speaker. The deterministic component ($p(t)$) of a pitch-synchronous residual frame is computed as ensemble average of individual pitch-synchronous residual frames of a phone. The deterministic component is given by:

$$p(t) = \frac{\sum_{i=1}^{N} e_i(t)}{N} \tag{8}$$

The noise component of each pitch-synchronous residual frame is computed by subtracting the deterministic component from the individual pitch-synchronous residual frame of a phone. The noise component ($r_i(t)$) computed for the i^{th} pitch-synchronous residual frame is given by:

$$r_i(t) = e_i(t) - p(t) \tag{9}$$

Using the deterministic component and variable noise components, the excitation signal of a phone can be constructed uniquely. This deterministic plus noise model based source model is used for generating the excitation signal in HTS. During training, the deterministic components estimated from all phones are systematically arranged in the form of a decision tree (Narendra and Rao, 2016). The nodes in the decision tree contain questions related to positional and contextual features of the phone. The decision tree is developed in such a way that every leaf contains a cluster of acoustically similar deterministic components. During synthesis, for the given target unit specification, an appropriate leaf is selected by traversing through the nodes of decision tree. A suitable deterministic component is selected from the leaf based on target and concatenation costs. Clustering the deterministic components in the form of a decision tree helps in selection of appropriate deterministic component during synthesis. The noise components are parameterized in terms of spectral and amplitude envelopes and modelled using HMMs (explained in detail in Section 4.4). The spectrum of noise component is represented using LP coefficients. The amplitude envelope of noise component is parameterized by resampling the overall amplitude envelope into 15 samples. At the time of synthesis, for every phone, a suitable deterministic component is chosen from the leaf of the decision tree. The noise component is obtained by imposing the target spectrum and amplitude envelopes generated from the HMMs. The sum of

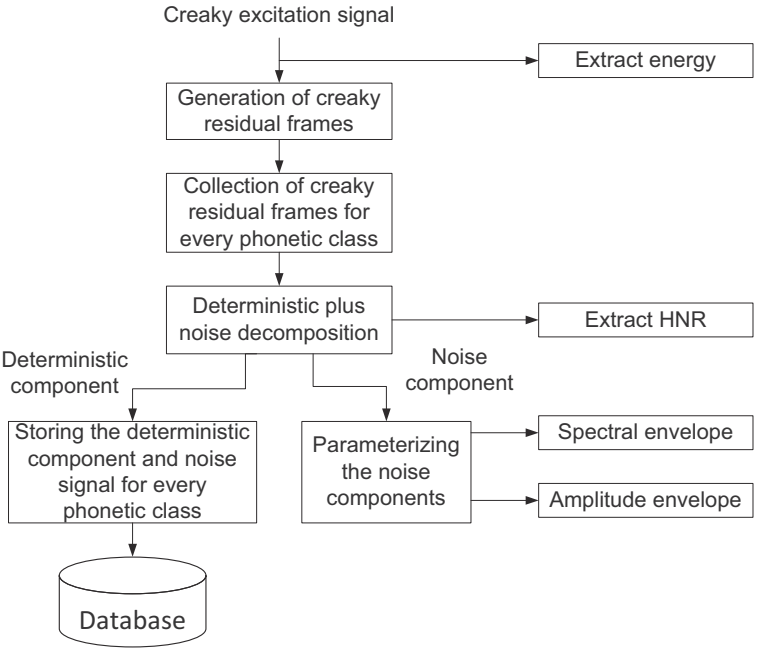

Fig. 8. Flowchart indicating the sequence of steps in creaky excitation modelling.

deterministic and noise components are pitch-synchronously overlap-added to construct the excitation signal of a phone. This source model is capable of modelling and generating the excitation signal with modal voice quality. Hence, the speech synthesis system can synthesize speech with only modal voice quality. But, to produce good quality speech, speech synthesis system should be able to produce both modal and creaky voices at appropriate places in the speech utterance. In this work, a hybrid source model is proposed to generate the excitation signal for creaky voice, in addition to modal voice.

4.1. Overview of proposed hybrid source model

Proposed hybrid source model, which is an extension of basic time-domain deterministic plus noise model (Narendra and Rao, 2016) for generating creaky excitation, is shown in Fig. 8. First, energy is extracted from every frame of the creaky excitation signal. Creaky pitch-synchronous residual frames are extracted from the excitation signal (described in Section 4.2). Here, both glottal closure and secondary excitation instants of creaky residual frames are synchronized. By ensuring GCI at the centre of every frame, GCIs of all creaky residual frames are synchronized. By fixing the interval between secondary excitation instant and GCI to a constant value, synchronization of secondary excitation instants of creaky residual frames is achieved. The procedure for synchronizing the GCIs and secondary excitation instants of creaky residual frames is explained in detail in Section 4.2. Creaky residual frames of every phonetic class are collected from the entire database (detailed in Section 4.3). From the creaky residual frames of every phonetic class, deterministic and noise components are computed. The deterministic components of all phonetic classes are stored in the database. The noise components are parameterized in terms of spectral and amplitude envelopes (described in Section 4.4). Harmonic to noise ratio (HNR) is computed as the ratio of energy of deterministic and noise components. For every phonetic class, along with the deterministic component, a single natural instance of noise signal is also stored. The noise component having spectral and amplitude envelopes close to average spectral and amplitude envelopes of all noise components is considered as the appropriate noise signal of a phonetic class. Energy, HNR, spectral and amplitude envelopes are considered as creaky parameters and modelled under HMM framework. During synthesis, for generating creaky excitation, suitable creaky deterministic component of a phone is chosen from the database, and the noise component is obtained by imposing the target spectral and amplitude envelopes on the natural instance of noise signal (explained in Section 4.5). The issues involved in modelling of creaky excitation signal are explained in the following section.

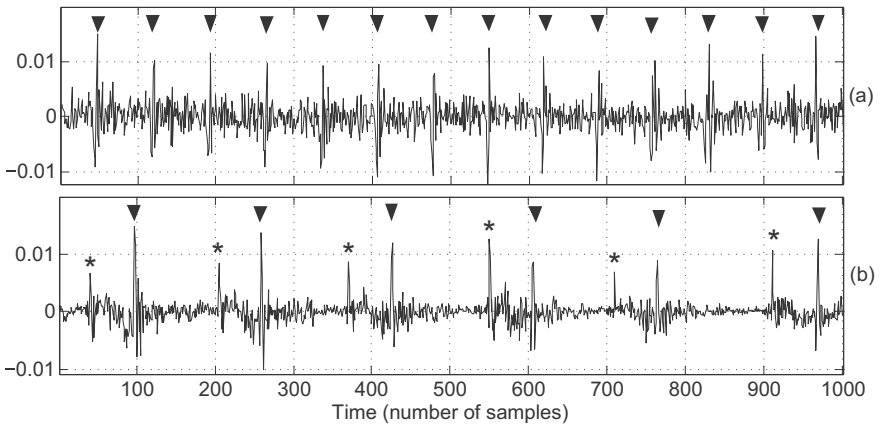

Fig. 9. Excitation signal of (a) modal and (b) creaky voices ($Fs = 16$ kHz). GCI locations and secondary excitations are indicated by downward arrows (▼) and stars (★), respectively.

4.2. Generation of creaky residual frames

For proper source modelling, pitch-synchronous residual frames should be extracted appropriately from the creaky regions. In the creaky voice, in addition to primary excitation at GCI, secondary excitation peaks are also present (Blomgren et al., 1998). Fig. 9 shows the excitation signals of modal and creaky voices ($Fs = 16$ kHz). In the figure, GCI locations and secondary excitations are indicated by downward arrows (▼) and stars (★), respectively. To obtain appropriate creaky pitch-synchronous residual frames, accurate location of GCIs and secondary excitation instants are very essential. In this work, zero-frequency filtering method is used to extract optimal GCI locations in creaky regions. Secondary excitation peaks are detected by identifying the prominent peak between consecutive GCIs. The region around GCI (around ± 1 ms) is not considered to avoid detection in the close vicinity of GCI. Previous studies have reported that the secondary excitation peaks occur due to sharp discontinuities at the glottal opening instant, following a long glottal closed period (Blomgren et al., 1998; Ishi et al., 2010; Kane et al., 2013). In this work, the time interval between the secondary excitation peak and its following GCI is termed as open period, and the time interval between the GCI and its subsequent secondary excitation peak is termed as closed period.

The procedure for extracting the pitch-synchronous residual frames from creaky regions differs from that of modal regions. In modal voice, the amplitude of the excitation signal is maximum at GCI. The pitch-synchronous residual frames are extracted in such way that GCI is at the centre of the frame and the length of the frame is two pitch periods. By ensuring GCI at the centre of the frame results in the accurate estimation of deterministic and noise components. In case of creaky excitation, in addition to GCI, secondary excitation peaks of creaky residual frames should also be synchronized. To synchronize secondary excitation peaks, the open and closed periods of creaky residual frames are analysed. Fig. 10 shows the distribution of open and closed periods of creaky residual frames ($Fs = 16$ kHz) for the male speaker (BDL) obtained from CMU Arctic database (CMU ARCTIC speech synthesis databases, [Online]). From the figure, it can be observed that the closed period varies almost linearly with the pitch period. The open period is narrowly distributed around a single value for different pitch periods. In this work, open periods of creaky residual frames are set to a constant value (3.75 ms for BDL speaker). Constant open period duration of all residual frames ensures synchronization of secondary excitation peaks. Creaky residual frames are pitch-normalized by resampling only the closed period.

4.3. Deterministic plus noise decomposition for every phonetic class

In the basic time-domain deterministic plus noise model based hybrid source model (Narendra and Rao, 2016), deterministic plus noise decomposition is performed by considering the residual frames of every instance of a phone. In case of creaky source modelling, a slightly different approach is followed. The duration of occurrence creaky voiced region is very less in a phone, and subsequently the number of pitch-synchronous residual frames extracted from creaky

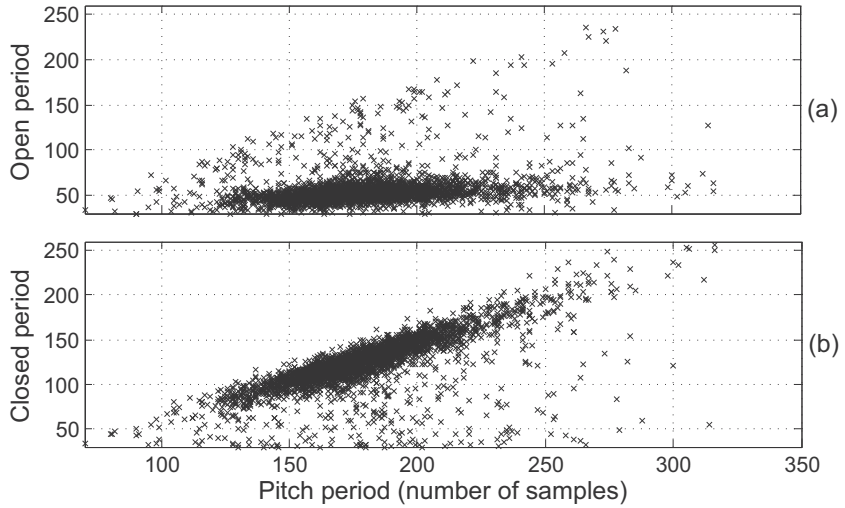

Fig. 10. Distribution of (a) open periods and (b) closed periods of creaky residual frames (Fs = 16 kHz) for the male speaker (BDL) obtained from CMU Arctic database.

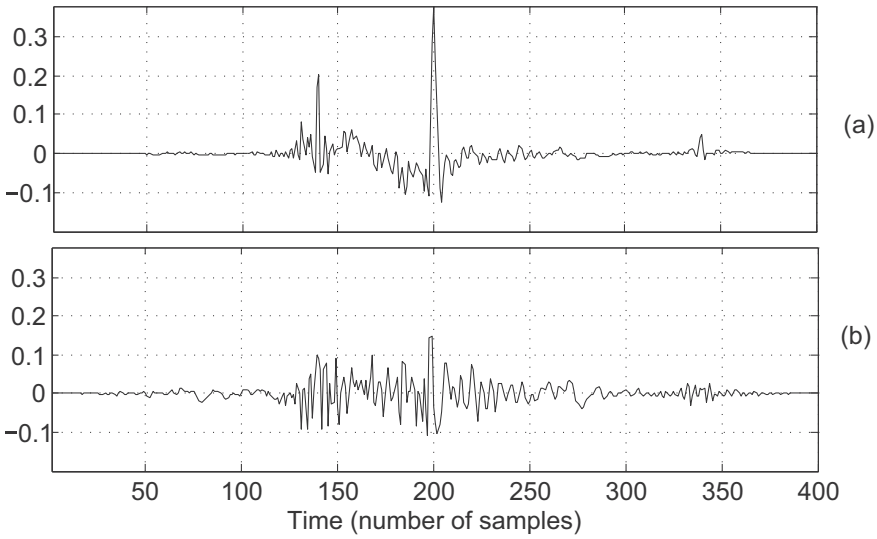

Fig. 11. (a) The deterministic component and (b) a single instance of noise component corresponding to creaky residual frames (Fs = 16 kHz) of phone /aa/.

voiced regions is also very less (about 4–5 frames). On an average, the percentage of creaky regions in the speech corpus varies from 3% to 12% (Drugman et al., 2014). By considering a small number of creaky residual frames, the estimated deterministic component is not accurate. The deterministic component that is computed as ensemble average of residual frames requires relatively more number of residual frames for accurate estimation. Hence, in the context of creaky regions, deterministic plus noise decomposition is not performed on every instance of a phone. Instead, deterministic plus noise decomposition is carried out on the creaky residual frames of every phonetic class. The deterministic component estimated from every phonetic class is stored in the database and the noise components are represented in terms of parameters. In BDL speaker of CMU Arctic database, 28 voiced phonetic classes are present. By considering the creaky residual frames of every phonetic class, the deterministic and noise components are computed. Fig. 11 shows the deterministic component and a single instance of noise component corresponding to creaky residual frames (Fs = 16 kHz) of phone /aa/.

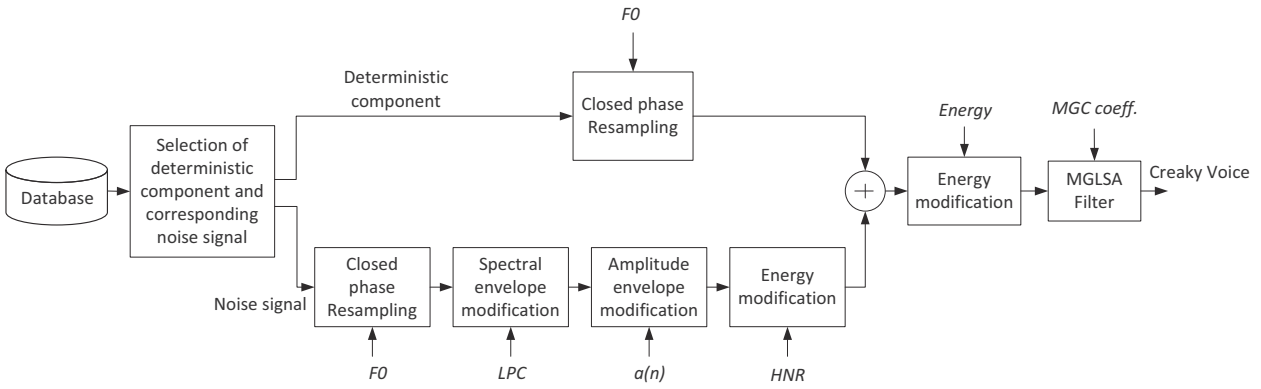

Fig. 12. Block diagram showing different stages in the synthesis of creaky voice. Parameters generated from HMMs are shown in italics.

4.4. Parameterization of noise component

The noise component is parameterized in terms of its spectral and amplitude envelopes. The spectral envelope of the noise component is estimated by using LP coefficients. The order of LPC is chosen based on the sampling frequency. Generally, the order of LPC is chosen to be (Fs/1000) + 4. Thus, according to this, LPC order should be 20 for the sampling frequency of 16 kHz. However, since we are only modelling the noise spectral envelope of excitation and not modelling the format structure of speech, lower LP order can be used. Hence in this work, for the sampling frequency of 16 kHz, the order of LPC is chosen to be 10. The LP coefficients are converted to line spectrum frequency (LSF) coefficients. The LSFs have better quantization properties and result in low spectral distortion than the conventional LP coefficients (Paliwal and Kleijn, 1995; Soong and Juang, 1984). The amplitude envelope ($a(n)$) is obtained by filtering the absolute value of noise component ($u(n)$) with a moving average filter of order $2N+1$. N is chosen to be 8. The amplitude envelope is given by:

$$a(n) = \frac{1}{(2N+1)} \sum_{k=-N}^{N} |u(n-k)|. \tag{10}$$

Normalization of the envelope is performed by setting the maximum value to 1. This method of amplitude envelope estimation was previously performed by Pantazis and Stylianou (2008). Due to smoothening by the moving average filter, the amplitude envelope shows slow variation. The overall shape of the amplitude envelope (whose original sampling frequency is 16 kHz) is represented by a small number of samples. In our case, the amplitude envelope is represented by downsampling it into 15 samples.

HNR, spectral and amplitude envelopes of noise component are computed for every pitch-synchronous residual frame. As it is convenient to model the parameters at frame size of 25 ms with frame shift of 5 ms, the parameters extracted from pitch-synchronous residual frames present in the frame are averaged and assigned as the parameters of that frame. In case of unvoiced speech, except energy of excitation signal, all other excitation parameters are set to zero. We have also examined the excitation parameters of unvoiced regions with the interpolated values instead of zero. But, we have not observed any noticeable difference in the synthesized speech while using the excitation parameters as zero or interpolated values during unvoiced regions. It is also observed that in the unvoiced regions, the excitation parameters including the HNR values are expected to be near 0 dB.

4.5. Speech synthesis using proposed hybrid source model

The block diagram showing different stages in the synthesis of creaky voice is presented in Fig. 12. In the figure, the parameters generated from HMMs are shown in italics. During synthesis, depending on the generated creaky probability, each frame is classified as either creaky or non-creaky. For the creaky frame, depending on the input phone corresponding creaky deterministic component and the natural instance of noise signal are selected from the database.

Pitch period of the deterministic component and the natural instance of noise signal are transformed into target pitch period ($F0$) by resampling only the closed period. The target spectral envelope generated from the HMM is imposed on the noise signal. The target spectral envelope is the all-pole model of noise represented by LSF coefficients. The LSFs are converted to LPCs (a_k). All-pole model of noise signal is evaluated using LPCs (b_k). An IIR filter is constructed from these two all-pole models that filter the noise signal to obtain the desired target spectrum. The transfer function of IIR filter is given by

$$H(z) = \frac{(1 - O(z))}{(1 - G(z))} \tag{11}$$

where $O(z) = \sum_{k=1}^{p} b_k z^{-k}$ and $G(z) = \sum_{k=1}^{p} a_k z^{-k}$ are the FIR filters obtained from the LPCs of noise signal and target spectral envelope, respectively. The target amplitude envelope ($a(n)$) generated by the HMM is imposed on the IIR filtered noise signal. The target amplitude envelope which is represented by 15 samples is upsampled to the required target pitch period. The amplitude envelope of the IIR filtered noise signal is also computed. The target envelope is imposed on the IIR filtered noise signal by compensating the difference between two envelopes. The energy of spectrum and the amplitude envelopes modified noise signal is modified according to the generated HNR and the energy of deterministic component constant throughout the phone. Both deterministic and noise components are pitch-synchronously overlap added, and the gain of the combined signal is matched according to the energy measure generated by the HMM. The resulting excitation signal is given as input to the Mel-Generalized Log Spectrum approximation (MGLSA) filter, controlled by MGC coefficients to obtain creaky voice. For modal voiced regions, the basic time-domain deterministic plus noise model based hybrid source model is used for generating the excitation signal (Narendra and Rao, 2016). For unvoiced speech, white noise whose energy is modified according to the generated energy measure is used as the excitation signal.

4.6. Evaluation

The proposed method is evaluated using two male speakers (BDL and KSP) from CMU Arctic speech database. The main reason for choosing BDL and KSP speakers for evaluation is that they have the highest percentage of creaky regions among all other speakers in the CMU Arctic speech database. Amount of creaky regions present in the BDL and KSP speakers' database are 7.6% and 2.7%, respectively (from Table 1). BDL and KSP speakers' database consists of 1100 phonetically balanced English utterances. The duration of the training set is about 51 and 59 minutes for BDL and KSP speakers, respectively. For every speech utterance, corresponding phonetic transcriptions are available in the CMU Arctic speech database. The proposed hybrid source model is compared with the basic time-domain deterministic plus noise model based source model (Narendra and Rao, 2016) (baseline method) meant for generating only modal voice. In both proposed and baseline methods, same F0 estimations methods are used for extracting pitch values from the speech utterance. In the proposed method, F0 is modelled using creaky probability as one of the features and in the basline method, F0 is modelled without using creaky probabilty information. Hence, there exists variation in the generated F0 trajectories in both baseline and proposed methods. As creaky voice occurs at the end of phrase and utterance, the variation in F0 trajectories between the baseline and proposed methods are generally observed at the end of phrase and utterance.

Subjective evaluation is conducted with 20 research scholars having sufficient speech knowledge for proper assessment of the speech signals. The subjects were given a pilot test about the perception of creaky and non-creaky regions by playing samples of synthesized speech files. The subejcts were briefed about the creaky voice as "a rough quality with the additional sensation of repeating impulses". This auditory criterion for the perception of creaky voice was followed in the previous literature on creaky voice detection methods (Ishi et al., 2008; Kane et al., 2013). Once they were comfortable with judging the synthesized speech, they were allowed to take the tests. The tests were conducted in the laboratory environment by playing the speech signals through headphones. Initially, 100 sentences that were not part of training data were synthesized using the proposed hybrid source model. Among 100 sentences, 20 synthesized speech files that have more than 3% of the creaky region are selected for evaluation. The average length and standard deviation of creaky segments for every utterance of BDL speaker are 0.2218 s and 0.0841 s, respectively. The average length of each of the synthesized speech utterance of BDL speaker is 4.0511 s and the percentage of creaky regions present in each of the utterance is 5.47%. For KSP speaker, the average length and standard deviation

Table 2
Scores for the CMOS test.

Score	Subjective perception
3	Much better
2	Better
1	Slightly better
0	About the same
−1	Slightly worse
−2	Worse
−3	Much worse

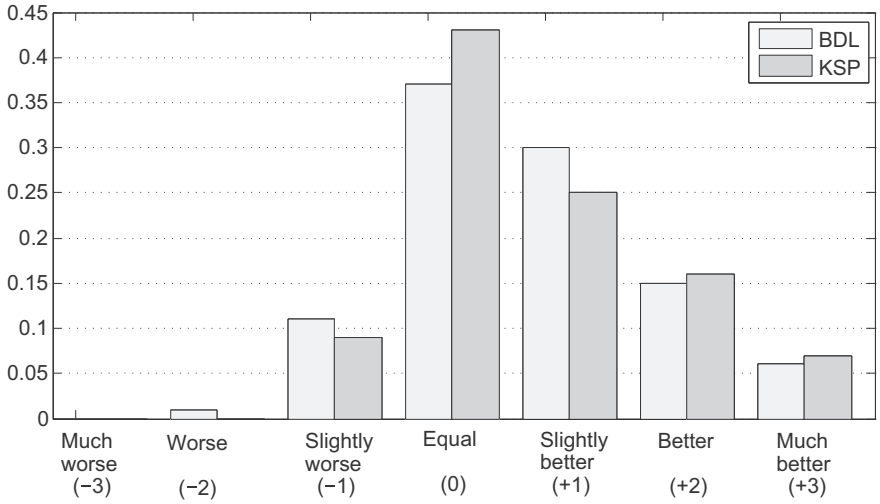

Fig. 13. Results of the CMOS tests obtained by comparing the proposed method with the basic time-domain deterministic plus noise model based source model.

of creaky segments for every utterance are 0.1133 s and 0.0202 s, respectively. The average length of each of the synthesized speech utterance of KSP speaker is 3.1292 s and the percentage of creaky regions present in each of the utterance is 3.62%.

Subjective evaluation is carried out using three measures, namely, comparative mean opinion scores (CMOS), preference tests and mean opinion scores (MOS). In CMOS, subjects were asked to listen to two versions, namely, speech synthesized from the proposed method and the other from the basic time-domain deterministic plus noise model based source model. Subjects were asked to grade overall preference between a pair of synthesized speech on a 7-point scale. The 7-point scale used to rank the preference between a pair of synthesized speech samples is shown in Table 2. A positive score indicates that the proposed method is preferred over other method, negative score implies the opposite and zero indicates that both methods are equivalent. Synthesized speech files are played to the subjects in random sequence to avoid the bias towards any particular method.

The results of CMOS tests are shown in Fig. 13. The figure provides a comparison between the proposed hybrid source model and time-domain deterministic plus noise model based source model according to the CMOS 7-point scale. In most of the cases, the subjects perceived that the proposed method is either *better* or *equivalent* to time-domain deterministic plus noise model based source model. The proposed method was never perceived as *much worse* than the time-domain deterministic plus noise model based source model. The CMOS scores with 95% confidence interval obtained for BDL and KSP are 0.88 ± 0.021 and 0.93 ± 0.018, respectively. As CMOS scores along with 95% confidence interval (BDL varying from 0.859 to 0.901 and KSP varying from 0.948 to 0.912) are above zero, it can be concluded that the performance of the proposed hybrid source model is significantly better than the basic time-domain deterministic plus noise model.

In preference tests, the subjects were asked to give the preference between pair of synthesized speech utterances. The subjects had the option either to prefer one of the synthesized speech utterances or to prefer both as equal. The

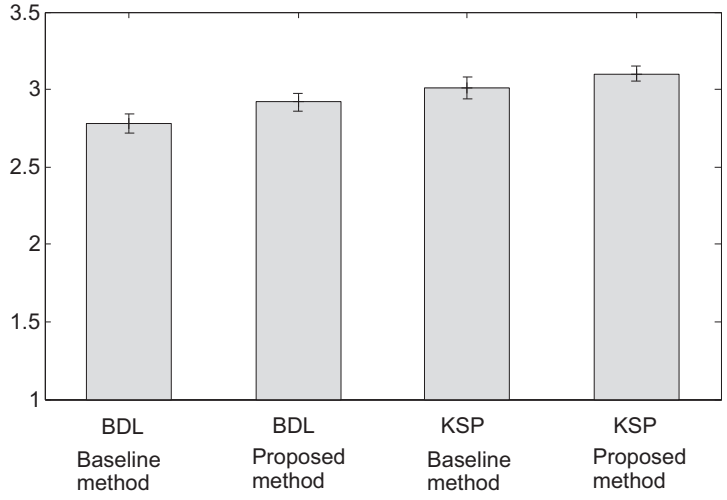

Fig. 14. MOS scores along with 95% conifidence interavls obatined for the proposed method and the basic time-domain deterministic plus noise model based source model (baseline method) for BDL and KSP speakers.

preference scores can be inferred directly from Fig. 13. From figure, it can be observed that around 50% of the ratings indicated preference for the proposed method and around 10% of the ratings preferred basic time-domain deterministic plus noise model based source model compared to the proposed method. Finally, around 40% of the ratings suggested that both methods as equal. These results confirm that the incorporation of creaky excitation has improved the quality of HTS.

In MOS tests, the subjects were asked to rate the naturalness of synthesized speech utterances on a scale of 1 to 5. Here, 1 indicates that the synthesized speech is highly unnatural and 5 indicates that the synthesized speech is completely natural. MOS scores along with 95% confidence intervals obtained for the proposed method and the basic time-domain deterministic plus noise model based source model (baseline method) for BDL and KSP speakers are shown in Fig. 14. For both BDL and KSP speakers, the confidence intervals of the proposed and baseline methods do not overlap. This indicates that the proposed method is statistically better than the baseline method. In addition to this, One-way ANOVA is carried out to investigate whether the performance of the proposed method is significantly better than the baseline method. One-way ANOVA indicated that the proposed method had a significant effect on the ratings [$F = 16.0$, $p < 0.001$] and pair-wise comparisons carried out using Tukey's Honestly Significant Difference (HSD) test showed that the proposed method gave significantly higher ($p < 0.001$) ratings for both BDL and KSP speakers. Fig. 15 shows the segments of synthesized speech utterance of BDL speaker generated using the basic time-domain deterministic plus noise model based source model (baseline) and the proposed hybrid source model. Dotted line indicates the region predicted as creaky. Synthesized speech samples of the proposed and basic time-domain deterministic plus noise model based hybrid source model are made available online at http://www.sit.iitkgp.ernet.in/~ksrao/creakyvoice/crk_hsm.html.

5. Discussion

From the subjective evaluation results, it is observed that the utilization of creaky deterministic component particular to every phonetic class resulted in the improvement of the quality of HTS. For two male speakers used in the evaluation, two separate HTS are developed. As single speaker data is used for developing the synthesis system, speaker normalization is not performed during training and synthesis stages of HTS and hence, mean F_0 information is not utilized in HTS. The proposed creaky voice detection method accurately identified the creaky regions in the speech utterances which prompted generation of creaky voice at appropriate places in the synthesized speech. Voicing detection and F_0 estimation is performed using a recently proposed method based on the strength of excitation. Using this method, most of the creaky regions are detected as voiced and accurate F_0 values are extracted in creaky regions (without F_0 halving and doubling errors). The F_0 values obtained from this method is used for F_0 modelling in both

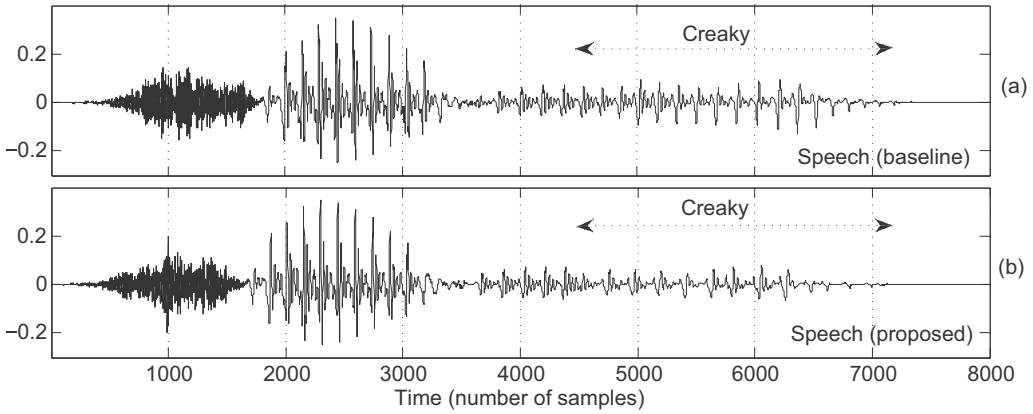

Fig. 15. Segments of synthesized speech utterance of BDL speaker generated using (a) basic time-domain deterministic plus noise model based source model (baseline) and (b) proposed hybrid source model. Dotted line indicates the region predicted as creaky.

proposed and baseline methods. As the creaky probabilty is used during F0 modelling, a slight variation in F0 contours is observed in the creaky regions of the proposed method compared to baseline method. The main reason for the improved quality of synthesis is the usage of efficient creaky voice detection and hybrid source modelling methods. From informal listening tests of different speech utterances containing creaky voice, following observations can be drawn. The creaky regions are mostly occurring at the end of the phrase or utterance where energy contour is tapering down. If the creaky frames of a phone are less than 10% the length of a phone, then the creaky segments are not perceived in most of the cases. As the length of creaky segments is small compared to entire speech utterance, it is difficult to perceive the effectiveness of the creaky segments at the speech utterance level. But, when the speech utterance is observed in small chunks, then the perceptual significance of the creaky regions can be noticed.

As both modal and creaky regions are generated within a phone, there may exist slight distortion and F0 fluctuations at the joining portions of modal and creaky regions in the synthesized speech. At certain places where the creaky regions are wrongly synthesized in place of modal voiced regions, slight unnaturalness can be perceived in the synthesized speech. In our database, most of the creaky regions contain secondary excitation. The hybrid source model is proposed by implicitly assuming that all the creaky residual frames contain secondary excitation. Hence, the proposed method cannot be directly applied on the speakers who do not produce secondary excitation in the creaky regions. The proposed creaky voice detection and hybrid source modelling method is performed by analysing the segments of the creaky voice of neutral speaking syle speaker. The proposed method should be suitably modified to be applied for other speaking styles such as story-telling, interactive conversation and expressive speech.

6. Conclusion

In this paper, HMM-based speech synthesis system was developed for generating creaky voice in addition to modal voice. First, a creaky voice detection method was proposed based on the variation of epoch parameters for different voicing regions. A neural network classifier was trained using the variance of epoch parameters obtained from ZFF method. A hybrid source model was proposed to generate the creaky voice at appropriate places in the synthesized speech. In the proposed hybrid source model, deterministic plus noise decomposition was performed on the creaky residual frames of every phonetic class. During synthesis, the suitable creaky deterministic component was obtained from the database, and the noise component was generated from the parameters generated from HMMs. The creaky deterministic and noise components were pitch-synchronously overlap-added to generate the creaky excitation signal. CMOS, preference and MOS tests indicated that incorporation of creaky excitation has improved the quality of HTS compared to basic time-domain deterministic plus noise model based hybrid source model.

References

Alku, P., Bakstrom, T., Vikman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112 (2), 701–710.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.
Blomgren, M., Chen, Y., Ng, M., Gilbert, H., 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. J. Acoust. Soc. Am. 103 (5), 2649–2658.
Carlson, R., Gustafson, K., Strangert, E., 2006. Prosodic cues for hesitation. In: Proc. of Fonetik, pp. 21–24.
Csapó, T.G., Németh, G., 2014. Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE J. Sel. Top. Sig. Process. 8 (2), 209–220.
Drugman, T., Dutoit, T., 2012. The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. Audio Speech Lang. Process. 20, 968–981.
Drugman, T., Kane, J., Gobl, C., 2012a. Modeling the creaky excitation for parametric speech synthesis. In: Proc. Interspeech.
Drugman, T., Kane, J., Gobl, C., 2012b. Resonator-based creaky voice detection. In: Proc. Interspeech.
Drugman, T., Kane, J., Gobl, C., 2014. Data-driven detection and analysis of the patterns of creaky voice. Comput. Speech Lang. 28 (5), 1117–1138.
Gobl, C., Chasaide, A.N., 1992. Acoustic characteristics of voice quality. Speech Commun. 11, 481–490.
Ishi, C., Sakakibara, K., Ishiguro, H., Hagita, N., 2008. A method for automatic detection of vocal fry. IEEE Trans. Audio Speech Lang. Process. 16 (1), 47–56.
Ishi, C.T., 2004. Analysis of autocorrelation-based parameters for creaky voice detection. In: Proc. International Conference on Speech Prosody, pp. 643–646.
Ishi, C.T., Ishiguro, H., Hagita, N., 2005. Proposal of acoustic measures for automatic detection of vocal fry. In: Proc. Eurospeech, pp. 481–484.
Ishi, C.T., Ishiguro, H., Hagita, N., 2010. Acoustic, electroglottographic and paralinguistic analyses of “Rikimi” in expressive speech. In: Proc. Speech Prosody, pp. 1–4.
Kane, J., Voice analysis toolkit. <https://github.com/jckane/Voice_Analysis_Toolkit>.
Kane, J., Drugman, T., Gobl, C., 2013. Improved automatic detection of creak. Comput. Speech Lang. 27, 1028–1047. Elsevier.
Kushan, S., Slifka, J., 2006. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English. In: Proc. of Speech Prosody, pp. 795–798.
Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1602–1613.
Murty, K.S.R., Yegnanarayana, B., Joseph, M.A., 2009. Characterization of glottal activity from speech signals. IEEE Sig. Process. Lett. 16, 469–472.
Narendra, N.P., Rao, K.S., 2015a. Robust voicing detection and F0 estimation for HMM-based speech synthesis. Circ. Syst. Signal Process. 34 (8), 2597–2619. doi:10.1007/s00034-015-9977-8.
Narendra, N.P., Rao, K.S., 2015b. Automatic detection of creaky voice using epoch parameters. In: Proc. Interspeech, pp. 2347–2351.
Narendra, N.P., Rao, K.S., 2016. Time-domain deterministic plus noise model based hybrid source modeling for HMM-based speech synthesis. Speech Commun. 77 (3), 65–83. http://dx.doi.org/10.1016/j.specom.2015.12.002.
Paliwal, K., Kleijn, W., 1995. Quantization of LPC parameters. In: Kleijn, W., Paliwal, E.K. (Eds.), Speech Coding and Synthesis, Elsevier, New York.
Pantazis, Y., Stylianou, Y., 2008. Improving the modeling of the noise part in the harmonic plus noise model of speech. In: Proc. International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), pp. 4609–4612.
Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., et al., 2011. HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Trans. Audio Speech Lang. Process. 19 (1), 153–165.
Raitio, T., Kane, J., Drugman, T., Gobl, C., 2013. HMM-based synthesis of creaky voice. In: Proc. Interspeech, pp. 2316–2320.
Silén, H., Helander, E., Koppinen, K., Gabbouj, M., 2007. Building a Finnish unit selection TTS system. In: Workshop on Speech Synthesis, pp. 310–315.
Silén, H., Helander, E., Nurminen, J., Gabbouj, M., 2009. Parameterization of vocal fry in HMM-based speech synthesis. In: Proc. Interspeech, pp. 1775–1778.
Soong, F., Juang, B.-H., 1984. Line spectrum pair (LSP) and speech data compression. In: ICASSP, pp. 37–40.
Surana, K., Slifka, J., 2006. Acoustic cues for the classification of regular and irregular phonation. In: Proc. Interspeech, pp. 693–696.
Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inform. Syst. 90 (5), 816–824.
Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K., 2013. Speech synthesis based on hidden Markov models. P. IEEE 101 (5), 1234–1252.
Vainio, M., 2001. Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis (Ph.D. thesis). University of Helsinki, Finland.
Vishnubhotla, S., Espy-Wilson, C., 2006. Automatic detection of irregular phonation in continuous speech. In: Proc. Interspeech, pp. 949–952.
Yanushevskaya, I., Gobl, C., Chasaide, A.N., 2005. Voice quality and f0 cues for affect expression. In: Proc. Interspeech, pp. 1849–1852.
Yanushevskaya, I., Gobl, C., Chasaide, A.N., 2009. Voice parameter dynamics in portrayed emotions. In: Proc. of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA), pp. 21–24.
Yoon, T.-J., Cole, J., Hasegawa-Johnson, M., 2005. Detecting non-modal phonation in telephone speech. In: Proc. Speech Prosody, pp. 33–36.
Zen, H., Toda, T., Nakamura, M., Tokuda, K., 2007. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inform. Syst. E90-D, 325–333.
Zen, H., Toda, T., Tokuda, K., 2008. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. IEICE Trans. Inform. Syst. E91-D (6), 1764–1773.
HMM-based speech synthesis system (HTS). <http://hts.sp.nitech.ac.jp/>.
CMU ARCTIC speech synthesis databases, [Online]. <http://festvox.org/cmu_arctic/>.