Incorporation of Excitation Source and Duration Variations in Speech Synthesized at Different Speaking Rates

Sri Harish Reddy M.[1] *and B. Yegnanarayana*[1]

[1]International Institute of Information Technology, Hyderabad, India.
sriharsham@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

The effect of speaking rate on the excitation source is examined using instantaneous fundamental frequency (F_0) and perceived loudness (η). The instantaneous F_0 and η seem to increase in the case of normal to fast speech, where as they are speaker-specific for the case of normal to slow speech. The study on duration variations of voiced, unvoiced and silence segments show that the duration changes are not uniform when speaking rate is varied. These observed variations in the excitation source and durations are incorporated in the epoch-based duration modification method. Perceptual studies show that these variations are significant for the perception of speaking rate.

Index Terms: speaking rate, duration modification, excitation source feature.

1. Introduction

Speaking rate variation is an important prosodic feature present in spontaneous speech, and it plays a vital role in human speech communication. It can be used by the speaker to convey paralinguistic and extralinguistic information. Variations of the speaking rate are also significant for naturalness. Because of large variability of the speaking rate in spontaneous speech, analyzing these variations helps to improve the performance of speech systems in practical scenario.

When humans vary speaking rate, they tend to vary the duration of the sound units. There are several approaches in literature for duration modification. Some of these approaches use sinusoidal model, pitch synchronous overlap and add (PSOLA), and phase vocoders to modify the duration [1, 2]. These methods modify the speech signal directly to achieve the desired duration modification, which may produce some spectral and phase distortions. Modification of the linear prediction (LP) residual to achieve the desired duration modification was proposed in [3]. Modification of duration in the residual domain will reduce the spectral and phase distortions [3]. All the above mentioned methods modify the duration of the speech signals uniformly. In general, it is observed that all speech regions are not uniformly modified with changes in the speaking rate. There are a few methods suggested in the literature to perform nonuniform duration modification [2, 4, 5]. The assumption in most nonuniform duration modification methods is that compression and expansion do not occur during sounds which are not voiced, and they occur during voicing due to changes in the speaking rates. To perform nonuniform duration modification, voicing probability derived from sinusoidal pitch estimate was used in [2], information from voicing onset time was used in [5].

It was shown in several studies that, along with duration, various other acoustic features also vary when speaking rate is changed [6]. These changes occur at subsegmental (less than pitch period), segmental, and suprasegmental level. For producing natural sounding synthetic speech at different speaking rates, not only the overall duration, but the subsegmental, segmental, and suprasegmental level variations also need to be captured and incorporated in synthesis. This paper attempts to incorporate some of the variations at subsegmental level into a nonuniform duration modification algorithm. The nonuniform duration modification algorithm is based on the epoch-based uniform duration modification approach proposed in [5].

This paper is organized as follows: Section 2 describes extraction of some of the excitation source features and their variation with speaking rate. In section 3, analysis of variations in duration of different sound units is presented. A nonuniform duration modification method, which incorporates duration variations and variations of the excitation source features described in section 2 and section 3, is described in section 4. Evaluation of the proposed method is given in section 5. Section 6 summarizes the paper, and identifies the issues that need further study.

2. Variations in excitation source

In order to analyze the effect of speaking rate on the excitation source, variations in two properties of the excitation source is studied. (i) Instantenous F_0, which is proportional to the reciprocal of time interval between two glottal closure instants (epochs), and (ii) perceived loudness, which is proportional to the total acoustic energy emitted from the excitation source.

2.1. Speech material

In order to analyze the variations of instantaneous F_0 and perceived loudness with speaking rate, naturally spoken utterances at various speaking rates (slow, normal and fast) were recorded. The speech database used in the present study consists of 10 English sentences chosen from TIMIT dataset [7]. Each sentence was uttered by 25 male speakers at three different speaking rates, namely, fast, normal, and slow. The speakers were aged between 20 and 30 years. All the speakers spoke Indian English, and the native language of each speaker was one among Telugu, Hindi, Kannada, and Tamil. A total of 750 utterances (250 utterances for each speaking rate) were collected. The speech signals were sampled at 8 kHz.

2.2. Extraction of instantaneous F_0 and study of variation of instantenous F_0

During production of voiced speech, the excitation to the vocal tract system can be approximated by a sequence of impulses of varying amplitudes. The effect of discontinuity due to the impulse-like excitation is reflected across all the frequen-

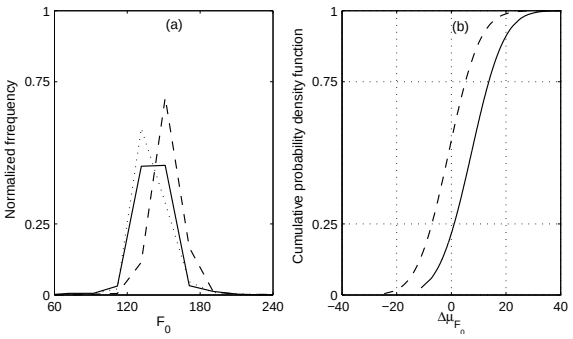

Figure 1: (a) Distributions of instantaneous F_0 for a male speaker. The solid ('—'), the dashed ('- - -'), and the dotted ('· · · ') lines correspond to normal, fast and slow speaking rates, respectively. (b) Cumulative probability density functions computed from $\Delta\mu_{F_0}$ values. The solid ('—') line correspond to normal to fast conversion, and dashed ('- - -') line correspond to normal to slow conversion.

cies, including around the zero frequency [8]. Therefore the zero-frequency filtering helps in emphasizing the characteristics of the excitation [8]. It was shown in [8] that the instants of positive-to-negative zero crossings of the zero-frequency filtered signal correspond the epochs in voiced speech. The instantaneous F_0 at each epoch is derived by computing the reciprocal of the time interval between the current epoch and the next epoch [8].

Fig. 1 (a) shows the distributions of the instantaneous F_0 of a male speaker, chosen at random from the list of 25 speakers. It is observed that the distributions of the instantaneous F_0 does discriminate between fast, normal and slow utterances of the speaker. The general trend observed across all the speakers is that instantaneous F_0 of fast speech is higher than that for normal and slow speech. In order to study the effect of speaking rate on the instantaneous F_0, difference between mean of the instantaneous F_0 of a fast utterance and the corresponding normal utterance ($\Delta\mu_{F_0}$) of the same sentence and uttered by the same speaker is computed for all 250 pairs. Fig. 1 (b) shows the cumulative probability density function (CPDF) computed from a Gaussian probability density function constructed from $\Delta\mu_{F_0}$ values. From the CPDF corresponding to normal to fast conversion (shown by solid line ('—') in Fig. 1), the percentage of sentences for which μ_{F_0} increases when speaking rate is increased is 80 %. Whereas, from observing the CPDF corresponding to normal to slow conversion (shown by dashed line ('- - -') in Fig. 1), shows that in 50 % of the cases $\Delta\mu_{F_0}$ is decreased and in 50 % of the cases $\Delta\mu_{F_0}$ is increased. This observation has implication in synthesis of speech at different speaking rates, because unlike duration, where increase (decrease) in the speaking rate results in a corresponding decrease (increase) in the duration, the instantaneous F_0 doesn't follow any specific trend. For synthesizing speech at different speaking rates, the instantaneous F_0 has to be modified suitably.

2.3. Variation of perceived loudness

The effect of speaking rate on perceived loudness was investigated by performing a perceptual loudness tests of speech at different speaking rates. The perceptual loudness tests were conducted using six listeners in the age group of 20-23 years, in a laboratory environment by playing the speech signals through headphones.. For perceptual evaluation, a subset of the database

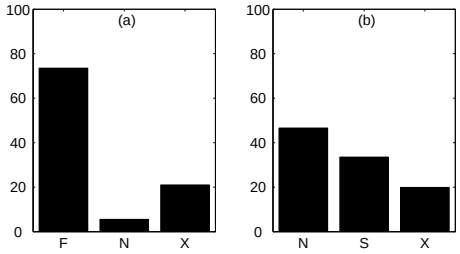

Figure 2: Evaluation of the effect of speaking rate on the perception of loudness. Results of comparison of (a) fast speech and normal speech, and (b) normal speech and slow speech.

(described in Sec. 2.1) was chosen. The subset contains speech utterances spoken by 6 male speakers (i.e., 60 utterances at each speaking rate). Two speech files, one at a normal speaking rate and the other at a fast speaking rate, were played out in succession. Both the utterances contained the same sentence, and were spoken by the same speaker. The listeners were asked to choose 'F' (or 'N'), if they perceived the utterance spoken at fast (or normal) speaking rate as the louder of the two. If the listeners were unable to distinguish between the loudness of the two utterances in the pair, they were asked to choose 'X'. Sixty pairs of utterances containing fast and normal utterances were used for listening. The same procedure was used to compare the loudness of utterances spoken at normal and slow speaking rates. In this case, the listeners were asked to choose 'S' (or 'N') if they perceived the utterance spoken at slow (or normal) speaking rate as louder. Figs. 2(a) and (b) show the results of perceptual evaluation.

From Figs. 2(a) and (b), it is observed that the listeners were able to distinguish between the loudness of fast and slow speech when compared with normal speech (percentages of 'X' in Figs. 2(a) and (b) are 21 and 19, respectively). In the case of fast vs normal speech, a significant number of utterances of fast speech have been marked as louder (73 % of the cases as shown in Fig. 2(a)). By contrast, the loudness scores of normal and slow utterances in the case of normal vs slow speech are close to each other (43 % and 33 %, respectively in Fig. 2(b)). This study shows that fast speech sounds louder than normal speech in most cases, whereas the evidence is insufficient to show that normal speech is louder than slow speech. This implies that perceptually, fast speech is louder than normal speech. Also a change in the perception of loudness takes place when the speaking rate is changed (because of the low percentage of 'X' in Figs. 2(a) and (b)).

3. Variation of durations of voiced, unvoiced and silence

3.1. Identification of voiced, unvoiced and silence regions

During production of voiced speech, vibration of the vocal folds is prominent with high strength of excitation. In the absence of vocal fold vibration, the vocal-tract system can be considered to be excited by random noise, as in the case of fricatives. The energy of the random noise excitation is distributed both in time and frequency domains. While the energy of an impulse is distributed uniformly in the frequency domain, and it is highly concentrated in the time-domain. As a result, the filtered signal exhibits significantly lower amplitude for random noise excitation compared to the impulse-like excitation. It was shown that the energy of the zero-frequency filtered signal described

Table 1: Slopes (m) of the straight lines that fit the clusters shown in Fig. 3, and root mean squared error (RMSE) of the fits to illustrate the variation in duration modification of different units.

Speech Segment	Normal to fast		Normal to slow	
	m	RMSE	m	RMSE
voiced	0.68	0.057	0.44	0.145
unvoiced	0.69	0.148	0.63	0.4902
silence	1.434	0.175	2.39	0.6482

in Sec. 2.2 can be used to detect the regions of significant vocal folds vibration (voiced regions).

Nonvoiced regions consist of silence and unvoiced regions of speech. Unvoiced regions typically have a higher spectral energy than silence regions at higher frequencies (around 3000 Hz). This information is used to separate silence regions and unvoiced regions in the nonvoiced regions. A resonator located at 3000 Hz and having a bandwidth of 100 Hz is used to filter the speech signals. The output signal will have frequencies only in the regions around 3000 Hz. The system function of a two pole resonator having center frequency at 3000 Hz and bandwidth of 100 Hz is given by

$$H(z) = \frac{1}{1 - 2e^{-\frac{\pi}{8}} cos(\frac{5\pi}{4})z^{-1} + e^{-\frac{\pi}{4}} z^{-2}}. \qquad (1)$$

A high frequency signal ($x_{uv}[n]$), is obtained by filtering the speech signal through $H(z)$. The energy of $x_{uv}[n]$ over a 20 ms window represented by $v_{uv}[n]$, is used to discriminate unvoiced and silence regions. A binary unvoiced-silence decision is computed as

$$d_{uv}[n] = \begin{cases} 1, & \text{if } v_{uv}[n] > vt \\ 0, & \text{otherwise,} \end{cases} \qquad (2)$$

where vt is the threshold. The value of the threshold is the maximum of average energies of starting 200 ms and final 200 ms of $v_{uv}[n]$.

3.2. Duration variations

For each utterance in the database, the durations of voiced, unvoiced and silence regions are obtained. The measure used to analyze the duration variations of these units when speaking rate is changed from normal to nonnormal (fast or slow) is the duration modification factor. The duration modification factor of a unit is defined as $\alpha = \frac{t_f}{t_i}$, where t_f and t_i are the initial and final durations, when the speaking rate is changed from normal to non-normal (fast or slow) speaking rate. The duration modification factors of voiced, unvoiced, silence and whole sentence are represented as α_v, α_u, α_p and α_S, respectively. If α_r represents any one of voiced, unvoiced and silence regions, it is analyzed with reference to the α_S to identify the relationship between α of a region and the utterance. Fig. 3 shows the scatter plots of α_r *vs* α_S in both normal to fast and normal to slow cases. In order to identify the relationship between α of a region and the utterance, a polynomial curve fitting algorithm was used to find a straight line that best fits the clusters shown in Fig. 3. The slopes of the lines that fit the cluster and the root mean square error (RMSE) of the fit are given in Table 1

Table 1 shows that the error of the fit is highest for silence regions and lowest for voiced regions in both normal to fast and normal to slow cases. This implies that variability in the way in which voiced duration is modified is less than that of unvoiced and silence regions. Also, error of the fit is more for normal to slow case than for normal to fast case. This implies that variability in duration modification of regions during normal to fast conversion is less than for normal to slow conversion. If the slope of the line that fits the cluster is approximately equal to 1, then α_r is equal to α_s. Likewise, if the slope of the line is less than 1, then α_r is less than α_s, and vice-versa. Observing the slopes of lines that fit the clusters, the slopes of lines corresponding to voiced and unvoiced regions are less than 1 and that of silence region are more than 1. This implies that in both normal to fast and normal to slow conversions, the duration modification factor of voiced and unvoiced regions is less than the duration modification factor of the utterance, and duration modification factor of silence regions is higher than than duration modification factor of the utterance. This analysis showed that humans rely more on variations of silence regions, than on variations of voiced and unvoiced regions during the production of speech at different speaking rates.

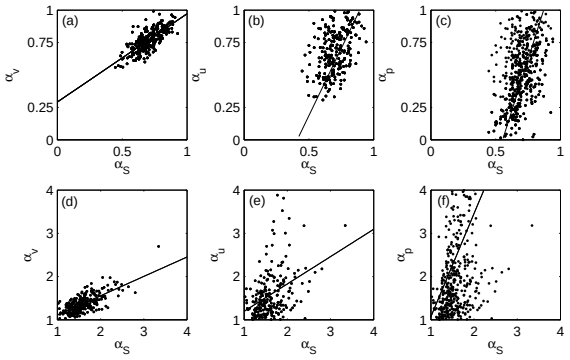

Figure 3: Scatter plots of duration modification factors of a region *vs* the corresponding utterance. (a), (b) and (c) show the scatter plots of α_v *vs* α_S, α_u *vs* α_S, and α_p *vs* α_S, respectively in normal to fast conversion. Similarly, (d), (e) and (f) show the scatter plots in normal to slow conversion. In each plot, straight line that fits the cluster is shown by a solid line.

4. Synthesis of speech at different speaking rate

It was shown in section 3.2 that nonuniform duration modification occurs when humans produced speech at different speaking rates by analyzing the variation in the durations of voiced, unvoiced and silence regions. In order to incorporate these variations, prior information about duration modification factors of voiced, unvoiced and silence regions is required when the task is to modify the duration of an utterance by a required factor. The information about the factor with which the duration of a unit has to be modified, when the duration modification factor of an utterance is available, is obtained from the lines that fits the clusters shown in Fig. 3. When the line is in the form of $y = mx + c$, as 'm' and 'c' are available, abscissa denotes the duration modification factor of an utterance and ordinate denotes the duration modification factor of the unit. A nonuniform duration modification method which uses the information of duration modification factors of a unit when duration modification of an utterance is available, is proposed in this section. The nonuniform duration modification method is based on epoch-based duration modification method proposed in [3].

Table 2: Mean opinion scores and AB ranking test (in %).

method \ α	0.5	0.7	0.9	1.1	1.3	1.5	AB-test
U	2.0	3.0	3.0	3.0	3.0	3.0	12.5
NU	2.6	3.3	3.6	3.0	3.3	3.3	50.0
$NU + F_0$	2.6	3.0	4.0	3.3	3.0	3.0	37.5

4.1. Incorporation of duration variations and excitation source variations

The proposed nonuniform duration modification method is as follows: Linear prediction residual of the speech signal is computed by performing a 12^{th} order linear prediction (LP) analysis using a 20 ms frame size and 5 ms frame shift. Identification of the boundaries of voiced, unvoiced and silence regions is performed using the method described in Sec. 3.1. Modification of the duration of the LP residual is performed with duration modification factor equal to voiced duration modification factor using epoch-based method [3]. This step takes care of the duration modification of voiced regions. Residual corresponding to silence and unvoiced regions are resampled to match the required duration, and silence and unvoiced regions in scaled residual signal are replaced by the resampled residual signal. This step takes care of the required duration modification of silence and unvoiced region. The filter coefficients (LPCs) are updated depending on the length of the modified LP residual. Speech for the desired duration modification can be synthesized by exciting the all-pole filter using the modified LP residual.

It was shown in Sec. 2 that when speaking rate is changed, along with modification of duration, the excitation source also changes. The variations of the instantaneous F_0 is incorporated in the proposed nonuniform duration modification method using the epoch-based pitch modification algorithm, proposed in [3].

5. Evaluation of synthetic speech at different speaking rates

The performance of the proposed method for synthesis of speech at different speaking rates is compared with epoch-based duration modification method using perceptual evaluation. Perceptual evaluation was carried out by conducting subjective tests with 10 research scholars in the age group of 21-35 years. Two sentences were chosen to perform the test. Speech signals were derived for the duration modification factors from 0.5 to 1.5 in the steps of 0.2. For each modification factor, three types of speech signals were derived. They are: speech signals using (i) uniform duration modification (U), (ii) nonuniform duration modification (NU), and (iii) nonuniform duration modification with instantaneous F_0 variations incorporated ($NU + F_0$). In the $NU + F_0$ method, the instantaneous F_0 is modified by constant factor, 1.2 and 0.8 in for the cases of normal to fast and normal to slow conversions, respectively. The tests were conducted in the laboratory environment by playing the speech signals through headphones. Two types of tests: (i) mean opinion score (MOS) test in which the listener evaluates on a score from 1 (worst) to 5 (best) based on the quality and perceptual distortion, and (ii) AB ranking test where the listener has to choose an utterance from the presented utterances.

The mean opinion scores (MOS) for duration modification factor is shown in Tab. 2. It can be observed from perceptual evaluation that speech synthesized from NU method has higher scores than U for almost all the duration modification factors (expect for α=1.1). Similarly, $NU + F_0$ has higher scores than U for some values of α and equal for the rest. AB ranking test (shown in last column of Tab. 2), a significant number of listeners have chosen speech signals synthesized using NU and $NU + F_0$ methods (50 % + 37.5 %). It was shown in Sec. 2 that instantaneous F_0 varies significantly with speaking rate, but the difference between scores of NU and $NU + F_0$ is not as expected. The quality of $NU + F_0$ can be improved by modifying the instantaneous F_0 speaker-specifically. Also, a study of the temporal variations of the instantaneous F_0 and incorporating these variations in synthesis of speaking rate may improve the quality.

6. Conclusions

The effect of speaking rate on the excitation source and durations of voiced, unvoiced and silence segments is studied. The features related to excitation source such as instantaneous F_0 and perceived loudness showed a consistent trend across all the speakers when the speaking rate is increased. Whereas, when speaking rate is decreased the trend is not consistent. The study on duration variations showed that variability is less in voiced duration modification than unvoiced and silence duration modification. Also voiced duration modification factor is less than unvoiced and silence segments. A nonuniform duration modification method is proposed which is based in the epoch-based uniform duration modification method, which uses the needed information about the duration modification factors of voiced, unvoiced and silence regions from scatter plots of α_r *vs* α_S. The variations of instantaneous F_0 have been incorporated in the proposed nonuniform duration modification method. Perceptual evaluation results showed that speech signals synthesized from the proposed methods have higher scores the epoch-based duration modification method. The quality of synthesis of speech at different speaking rates can be improved further by incorporating variations in the perceived loudness and temporal variations of the instantaneous F_0.

7. References

[1] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175–206, 1995.

[2] T. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. on Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar 1992.

[3] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[4] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *IEEE Int. Conf. on Advanced Learning Technologies*, Athens, Greece, Jul. 2003, pp. 165–169.

[5] K. S. Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Commun.*, vol. 51, pp. 1263–1269, 2009.

[6] T. Gay, "Effects of speaking rate on vowel formant movements," *J. Acoust. Soc. Am.*, vol. 63, no. 1, pp. 223–230, Jan. 1978.

[7] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom," Linguistic Data Consortium, Philadelphia, PA, USA, 1993.

[8] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.