INTRODUCING ROUGHNESS IN INDIVIDUALITY TRANSFORMATION THROUGH JITTER MODELING AND MODIFICATION

Ashish Verma

IBM India Research Lab
Indian Institute of Technology
New Delhi, India
vashish@in.ibm.com

Arun Kumar

Center for Applied Research in Electronics
Indian Institute of Technology
New Delhi, India
arunkm@care.iitd.ernet.in

ABSTRACT

Individuality transformation is a process to modify the speech signal in a person's voice so that it sounds as if spoken by another person. In most individuality transformation methods, pitch transformation is performed through a simple scaling considering the global pitch characteristics of the source and target speakers without considering the short-term pitch variation or jitter. In this paper we present a novel method to model and modify jitter in the speech signal to introduce a handle on roughness in the process of individuality transformation. The proposed method is based upon computing the average intensity in a band around the fundamental frequency in the spectrum of a speaker's mean normalized pitch contour. The validity of the proposed method to model jitter has been established by subjective tests for perceived roughness in the speaker's voice. It is also shown that modification of jitter by the proposed method results in an improved subjective rating for individuality transformation.

1. INTRODUCTION

The process of individuality transformation mainly consists of transformations of the spectral envelope [1, 2, 3, 4, 5], and prosodic features like pitch and speaking rate [6, 7, 8]. While an enormous amount of research has been performed for spectral envelope transformation, other features like roughness, prosody etc. have received a relatively limited attention. However, it has been observed that these features also play a very important role in the perceived speaker's identity. Jitter, *i.e.*, period to period pitch variation, is known to be an important factor in the voice individuality [9, 10]. It affects perceived roughness and hoarseness in the speech signal [11, 12]. However, there is very little evidence in the literature to modify this aspect of the speech signal for individuality transformation.

Generally, simple scaling based methods have been used to transform the pitch value of one speaker to that of another speaker. These methods either statistically normalize the pitch values or find a mapping for the pitch contour as a whole [1, 4, 5, 7, 8]. In this paper, we propose a method to model and modify jitter of one speaker to that of another speaker in the context of individuality transformation. We define jitter of a speaker as the average intensity in a band around fundamental frequency in the spectrum of the mean normalized pitch contour. Other researchers have also defined jitter to quantify the period to period pitch variation in different ways [13, 14]. We also present a two-step approach for pitch modification which accounts for the jitter difference between the source and the target speakers and also compensates for the global pitch mean and variance differences of the two speakers.

It may be noted that individuality transformation is an efficient method for personalized speech synthesis as most of the current state of the art text-to-speech synthesis systems are based on concatenative technology. The process of individuality transformation renders personalized speech synthesis without requiring a large speech corpus in the target speaker's voice. It transforms the speech signal in a given speaker's voice so that it sounds as if spoken by the target speaker. Recently, the concept of voice fonts has been proposed for independent representation of a speaker's individuality and its transformation to another speaker's individuality [3, 4, 6]. Personalized speech synthesis is an integral component of various applications like, multimedia mail, distance learning, very low bit rate speech coding, personal assistant, speech to speech translation etc.

The rest of the paper is organized as follows. We present the concept of voice fonts in Section 2. The proposed approach for jitter modeling and modification is described in Section 3. We describe the experiments performed in Section 4 and the corresponding results in Section 5. We conclude in Section 6.

2. VOICE FONTS

Voice fonts represent a speaker's individuality as a set of "descriptors" of the speaker's voice, *viz.*, spectral envelope, pitch and speaking rate [3, 4, 6]. This representation is independent for a speaker and is not dependent on any source-target pair. The spectral envelope descriptor in voice fonts is represented at the phonemic class level [3] or at the articulatory class level [4]. Continuous speech utterances are recorded in the speaker's voice whose voice font is to be created. All the speech utterances are phonetically aligned by using the acoustic models of a speech recognition system through Viterbi alignment. Then the spectral envelope for all the frames corresponding to a phone or an articulatory class is modeled through a Gaussian Mixture Model (GMM). The speaking rate, is represented as average durations of voiced, unvoiced and silence acoustic categories. The value of these durations is computed from the aligned speech corpus of the speaker [6]. Pitch is represented for the speaker as mean and standard deviation of the pitch values across all the voiced frames in the corpus. The process to create voice fonts for a speaker is depicted in Fig. 1. At the time of individuality transformation the voice font descriptors for the source and the target speaker are used in an integrated framework to modify different aspects of the speech signal. In-

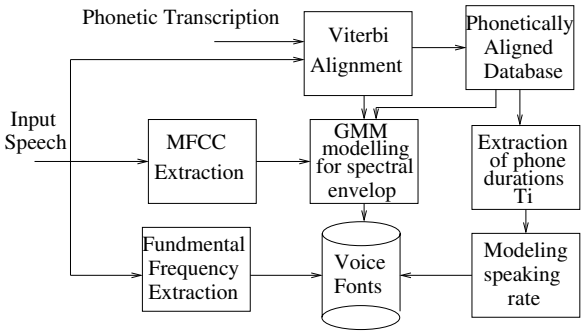

Fig. 1. *Creation of a speaker's voice font [6]*

dividuality transformation using voice fonts is depicted in Fig. 2. More details about voice fonts descriptors and invidiuality transformation using these descriptors can be found in [6].

In this paper we augment the set of descriptors for voice fonts by adding another descriptor for speaker individuality, *viz.*, jitter. Jitter is an important aspect of the voice quality and it affects the individuality of speech signals [9, 10]. We aim to model the jitter for a speaker and incorporate it in the overall individuality transformation process.

3. PROPOSED APPROACH

The proposed approach is motivated by the fact that for a given voiced segment, if the pitch is sampled at a high sampling rate, then jitter, or pitch variations from one pitch period to another, will reflect around the mean fundamental frequency in the spectrum of the sampled pitch sequence. The sampling rate of the pitch should be high enough to capture the fastest pitch variation for any speaker. We describe this approach in detail in the following subsections.

3.1. Jitter Modeling

We model jitter for a speaker as the average intensity in a band around the mean fundamental frequency in the spectrum of the mean normalized pitch contour. Firstly, the pitch values are computed for all the voiced frames present in the speaker's corpus. We use an autocorrelation based pitch detector similar to the pitch detector described in ITU-T standard G.729. The mean pitch value, $\bar{P}$, and its standard deviation, σ, for the speaker are computed from the set of pitch values thus obtained. Now for every voiced *segment*, (*i.e.*, the complete voiced region between two unvoiced regions), v, in the speech corpus, we compute the pitch values at a uniform sampling interval, τ_s, which is set to one millisecond. This sampling rate for pitch values is chosen so that it can capture short term pitch variations upto 500 Hz which is sufficient for the human pitch range. These pitch values are then refined using the mean pitch value to remove any halving or doubling of the pitch. The pitch values so obtained are normalized with the mean pitch value, $\bar{P}$, and a normalized pitch contour, $C_v(n)$, is obtained for the voiced segment.

$$C_v(n) = \{\hat{P}_1, \hat{P}_2, .., \hat{P}_n, .., \hat{P}_L\}$$

where $\hat{P}_i$, $1 \leq i \leq L$ are the mean normalized pitch values and L is the number of pitch samples in the voiced segment. Now for each of the pitch contours, a K-point ($K \geq L$) Fast Fourier Transform (FFT), $S_v(k)$, is computed corresponding to the normalized pitch contour, $C_v(n)$. An overall pitch spectrum, $\bar{S}(k)$, for the speaker is then obtained by taking an average of the individual spectrums:

$$\bar{S}(k) = \frac{1}{V}\sum_{v=1}^{V} S_v(k) \tag{1}$$

where V is the number of voiced segments present in the speaker's speech corpus. Jitter, J, for the speaker is computed as the average intensity in a band around the mean fundamental frequency of the speaker.

$$J = 10 * \log_{10}\left(\frac{1}{k_u - k_l + 1}\sum_{k=k_l}^{k_u} |\bar{S}(k)|^2\right) \tag{2}$$

where $k_l (= \lceil K\tau_s/(\bar{P} + m\sigma\rceil)$ and $k_u (= \lceil K\tau_s/(\bar{P} - m\sigma\rceil)$ are FFT indices which correspond to the pitch limits ($\bar{P} \pm m\sigma$) and m is a parameter to control the frequency range over which the intensity is averaged. It can be chosen depending upon the distribution of the pitch values. For example, for a Gaussian distribution, $m = 2$ will capture about 95% of speaker's pitch range.

3.2. Jitter Modification

The modification of jitter is performed as part of the *pitch and speaking rate conversion* module shown in Fig. 2. The algorithm for jitter modification is as follows:

- *Step 1*: For a given voiced segment, v, of length L, compute the pitch values, P_i^s, $1 \leq i \leq L$, at a regular interval (at sampling interval τ_s).
- *Step 2*: Compute corresponding target pitch values, P_i^t, to compensate for the global mean and variance

$$P_i^t = \frac{\bar{P}^s - P_i^s}{\sigma^s} * \sigma^t + \bar{P}^t \quad 1 \leq i \leq L \tag{3}$$

  where $\bar{P}^s$ and σ^s are the mean pitch and standard deviation for the source speaker respectively and $\bar{P}^t$ and σ^t are the corresponding values for the target speaker.
- *Step 3*: Obtain the mean normalized pitch contour, $C_v(n)$, for P_i^t, $1 \leq i \leq L$, and the corresponding spectrum, $S_v(k)$
- *Step 4*: Compute jitter, $\tilde{J}^t$ for the voiced segment by using $S_v(k)$ in place of $\bar{S}(k)$ in (2). Choose k_l and k_u as $\lceil K\tau_s/(\bar{P}^t + m\sigma^t)\rceil$ and $\lceil K\tau_s/(\bar{P}^t - m\sigma^t)\rceil$ respectively.
- *Step 5*: Pass the transformed pitch sequence, P_i^t, through a band pass filter, centered around $1/\bar{P}^t$ and having a passband corresponding to $1/(\bar{P}^t \pm m\sigma^t)$. The gain of the filter is adjusted to compensate for the difference in the jitter values of the transformed pitch sequence, $\tilde{J}^t$, and the required jitter value for the target speaker, J^t. A simple way is to use a frequency domain filter as follows:

$$\hat{S}_v(k) = \begin{cases} S_v(k) * 10^{(J^t - \tilde{J}^t)/20} & k_l \leq k \leq k_u \\ S_v(k) & otherwise \end{cases}$$

- *Step 6*: Take inverse discrete fourier transform of $\hat{S}_v(k)$ to obtain the modified pitch contour for the voiced segment.

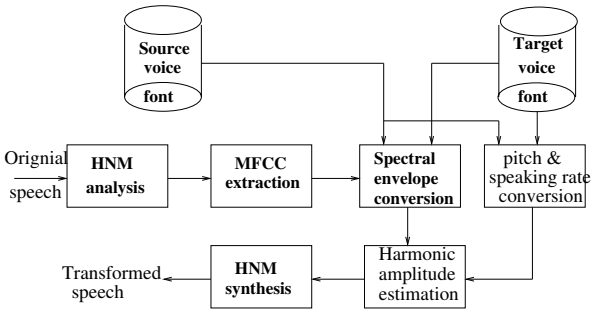

Fig. 2. Individuality transformation of speech using voice fonts [6]

Table 1. Computed jitter values for various speakers

Speaker	Avg. F0 (Hz)	J (dB)	Speaker	Avg. F0 (Hz)	J (dB)
ash	169	-0.22	nit	105	3.48
ak	139	1.51	pxk	177	0.66
vpg	137	1.89	vxt	188	-0.94
dxh	251	-2.33	axs	154	0.94
abc	187	-0.57	xyz	223	-2.01

4. EXPERIMENTS

We have conducted individuality transformation experiments to evaluate the performance of the proposed method. We use Degradation Category Rating (DCR) and an Opinion Test to judge the quality of the transformed speech signal. Subjective tests have also been conducted to find out the correlation between the computed value of the jitter for various speakers and the perceived roughness in their voices. The speech database consists of 10 speakers, including 6 male and 4 female speakers. About 30 minutes of speech is collected from each of the speakers in the form of continuous Hindi sentences recorded at 16 kHz sampling rate. The speech database is phonetically aligned using Hidden Markov Models (HMMs) of a large vocabulary continuous dictation type Hindi speech recognition system developed at IBM India Research Lab, New Delhi. The phonetically aligned database is used to extract voice fonts for the speakers as shown in Fig. 1. The voice fonts descriptors, *viz.*, spectral envelope, pitch and speaking rate, are used to perform individuality transformation using the voice fonts approach as shown in Fig. 2. More details about creating voice fonts and individuality transformation can be found in [6].

For each of the speakers, jitter is computed by the method described in Section 3.1. In all the experiments value of m is chosen to be 1. A pitch spectrum, in the form of 8192-point FFT, is obtained corresponding to the zero-padded mean normalized pitch contour of each of the voiced segments. About 2000 voiced segments of length 20 ms to 500 ms are used for the jitter computation for a speaker. FFTs corresponding to all the voiced segments are averaged to obtain the overall pitch spectrum, $\bar{S}(k)$, for the speaker.

The first part of the subjective tests is performed to find the correlation between the computed jitter and perceived roughness in the voice of various speakers. To accomplish this, 8 subjects, different from the recorded speakers, rate the perceived voice quality of the speakers on a scale of 1 to 5 representing increasing level of roughness. For each of the speakers, 2 sentences and 2 short words are played back. The subjective ratings for roughness corresponding to words and sentences are then averaged for the speaker.

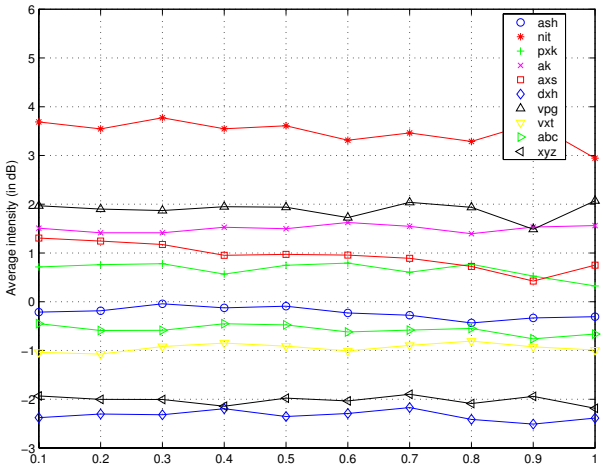

Fig. 3. Intensity distribution over $(1/(\bar{P} \pm k\sigma))$ for various speakers in the spectrum of their respective mean normalized pitch contours. x-axis shows the multiplication factor k

In the second part, individuality transformation is applied to 12 sentences and 15 short words (2-3 syllables) for various source-target pairs. Each word or sentence is transformed with and without jitter compensation. Individuality transformation is performed using the voice fonts approach from source to target speaker. The synthesis is performed using the Harmonic + Noise Model (HNM) of speech signals. The transformed sentences are rated using DCR by the subjects to judge the closeness of the transformed sentences on a scale of 1 to 5 representing decreasing level of degradation from the target speaker. In DCR, only the target speaker's real speech is played to the subjects. In the Opinion Test, we play both the source and target speakers' real speech samples and ask the subjects to rate the transformed word or sentence on a scale of 1 to 10 with 1 representing the actual source speaker and 10 representing the actual target speaker.

5. RESULTS AND DISCUSSION

5.1. Jitter modeling

The average intensity distribution in $\bar{S}(k)$, corresponding to a range of $1/(\bar{P} \pm \sigma)$, is shown in Fig. 3 for the speakers. It can be seen from the figure that the intensity levels are quite different for different speakers. Table 1 shows computed jitter values, J, as defined in (2), for the speakers along with their mean fundamental

Table 2. Subjective rating for roughness across the speakers

Speaker	Roughness	Speaker	Roughness
ash	1.88	nit	3.00
ak	2.3	pxk	2.43
vpg	2.68	vxt	1.93
dxh	2.0	axs	2.18
abc	1.81	xyz	1.37

Table 3. Results for subjective tests on personality transformation

	No Jitter (Words)	Jitter (Words)	No Jitter (Sent.)	Jitter (Sent.)
DCR	3.25	3.47	3.25	3.36
Opinion	7.52	8.22	7.45	8.02

frequency. Table 2 shows the subjective ratings for the speakers considering the perceived roughness in their voices. The values in Table 2 should be compared with the corresponding jitter values, shown in Table 1, computed using the proposed method. As can be seen for most speakers the subjective rating shows the same trend as their objective jitter scores. However, for few speakers the subjective rating are not in agreement with the corresponding jitter values. An overall correlation coefficient of 0.88 is found between the computed jitter values and the subjective ratings for roughness which validates the proposed definition of jitter.

5.2. Jitter transformation

Table 3 shows the results corresponding to the individuality transformation experiments. The first two columns of the table show the results when isolated words are used for individuality transformation and the last two columns have the results for full sentences. Note that the subjective scores improve when jitter is also transformed for the target speaker in addition to the other descriptors. Further, the amount of improvement is more in the case of short words as compared to the sentences. It may be due to the fact that other individual features like, sentence prosody and style of speaking, become more important in the case of sentences as compared to voice quality features like jitter. The overall scores in the Opinion Test are higher as compared to the DCR test because in this case the subjects listen real speech samples of both the source and target speaker and hence the transformed speech feels much closer to the target speaker as compared to the source speaker. The pitch contours obtained using the proposed approach for a 50 ms voiced segment are shown in Fig. 4 . In this case the target speaker has higher jitter than the source speaker. Note that the pitch contour with jitter modification has more variation as compared the one without jitter modifcation. The final pitch contour used in the transformed speech is a subsampled version of this contour.

6. CONCLUSION

We have proposed a novel method to model jitter in a speaker's voice based upon intensity distribution in the average pitch spectrum. We have shown how it can be used to modify jitter from one speaker to that of another speaker at the time of individuality transformation. A strong correlation is obtained between the computed value of jitter and perceived roughness in the speaker's voice. Further, it is shown that the transformation of the jitter results in a speech signal which is closer to the target speaker's voice.

7. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Apr. 1988, pp. 655–658.

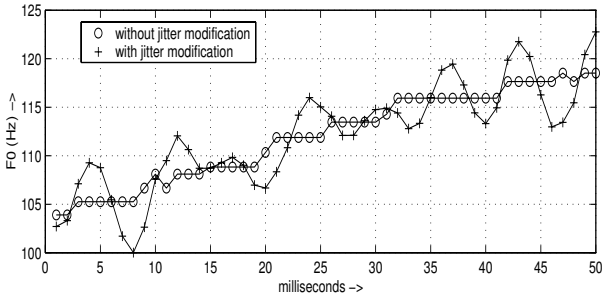

Fig. 4. Target pitch contours with and without jitter modification

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transaction on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[3] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, HongKong, Apr. 2003.

[4] A. Verma and A. Kumar, "Articulatory class based spectral envelope representation for voice fonts," in *Proc. Int. Conf. on Multimedia and Expo*, Taipei, June 2004.

[5] A. Kain and W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, May 1998.

[6] A. Verma and A. Kumar, "Modeling speaking rate for voice fonts," in *Proc. EUROSPEECH*, Geneva, Sept. 2003.

[7] O. Turk and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *EUROSPEECH*, Geneva, Sept. 2003, pp. 2845–2848.

[8] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modificatoin," in *Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, May 1998, vol. 2, pp. 885–888.

[9] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, pp. 165–173, 1995.

[10] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication*, vol. 5, no. 2, pp. 183–197, June 1986.

[11] W. J. Gould E. Yumoto and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. America*, vol. 71, no. 6, pp. 1544–1550, June 1982.

[12] Douglas O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 2000.

[13] D. L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features," in *Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, May 1998, pp. 21–24.

[14] B. Cheetham F. Plante, H. Kessler and J. E. Earis, "Speech monitoring of infective laryngitis," in *Int. Conf. on Spoken Language Processing*, Philaldelphia, Oct. 1996, pp. 749–752.