Modeling Speaking Rate for Voice Fonts

Ashish Verma

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

Arun Kumar

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

Abstract

Voice fonts are created and stored for a speaker, to be used to synthesize speech in the speaker's voice. The most important descriptors of voice fonts are spectral envelope for acoustic units and prosodic features such as fundamental frequency and average speaking rate. In this paper, we present a new approach to model the speaking rate so that it can be easily incorporated in voice fonts and used for personality transformation. We model speaking rate in the form of average duration for various acoustic units and categories for the speaker. The speaking rate can be automatically extracted from a speech corpus in the speaker's voice using the proposed approach. We show how the proposed approach can be implemented, and present its performance evaluation through various subjective tests.

1. Introduction

Personalized speech synthesis is becoming an area of interest due to its potential use in a number of applications. Multimedia mail, distance learning, very low bit rate speech coding and movie avtars are a few of these applications. An efficient way of achieving personalized speech synthesis is to apply personality transformation on the speech synthesized from a default text-to-speech synthesis system to get the same speech in the desired person's voice. Usually, the term “voice conversion" has been used in this context to denote the spectral envelope conversion of the speech signal [1, 2]. However, in general, voice conversion comprises of two functionalities, *viz.*, the spectral envelope conversion and prosodic features conversion of the speech signal. To avoid confusion, we use the term “personality transformation" in this paper to denote the complete conversion process, *i.e.*, to convert both the spectral envelope and prosodic features like fundamental frequency and speaking rate.

In this context, we have proposed the concept of voice fonts to synthesize personalized speech [3]. Just as text fonts make the text appear in a particular style, voice fonts make the speech sound in a particular individual's voice. Some of the most important descriptors of voice fonts comprise of spectral envelope for acoustic units and prosodic features such as fundamental frequency and rate of speaking. In [3], we have presented a new technique to model the spectral envelope of acoustic units, which is used as a descriptor of voice fonts. The proposed approach, avoids the use of a parallel speech corpus in the source and target speaker's voices, which was required in earlier voice conversion approaches [1, 2]. Hence the approach is used to independently create voice fonts without being concerned about the source-target pair.

In this paper we propose a new approach to model speaking rate for an individual which constitutes another descriptor of voice fonts. It models speaking rate as a function of average duration of three acoustic categories, *viz.*, voiced, unvoiced and silence. Using the proposed approach, speaking rate can be automatically extracted from a small speech corpus in the individual's voice. We show how this model can be used, at the time of personality transformation, to time scale the synthesized speech so that it closely matches the target speaker's speaking rate.

Rest of the paper is organized as follows. We present a review of prior art in Section 2. The concept of voice fonts is explained in Section 3. Proposed approach is described in Section 4. We describe the experiments conducted to evaluate the performance of the proposed approach in Section 5. The results of the experiments are discussed in Section 6.

2. Prior Art in Modeling Speaking Rate

There have been several attempts to model the speaking rate from a speech signal, mostly to study the effect of speaking rate on speech recognition and on spectral characteristics of acoustic units in general. Morgan et al. [4] proposed a parameter, called *enrate*, which measures the first spectral moment of the wideband energy envelope computed over 1-2 seconds. They found correlation of about 0.5 between *enrate* and transcribed syllable rate. In [5], Morgan et al. proposed an improved approach by combining *enrate* with a point-wise correlation between pairs of compressed sub-band energy envelops. This approach used multiple estimators of speaking rate to improve the correlation with transcribed syllable rate. In [6], Kitazawa also proposed a similar approach to measure speaking rate. In this approach, the dominant spectral peak of the long term full-band energy envelope was used to measure the speaking rate. Faltlhauser et al. proposed Gaussian Mixture Model (GMM) based approach to measure the speaking rate [7]. They trained three GMMs, one each for three categories of speaking rates, *i.e.*, fast, medium and slow. At the time of transformation, these GMMs are scored by the given speech spurt. The speaking rate corresponding to the GMM, having the highest score, is chosen as the correct speaking rate. They also proposed a mapping function based upon the three scores belonging to GMMs, to obtain a continuous speaking rate from the three broad categories. None of the above approaches is suitable to compute speaking rate, as a descriptor of voice fonts.

All the approaches described above model the speaking rate implicitly by analyzing other features, like energy envelope or Mel Frequency Cepstral Coefficients (MFCC), which are correlated with the speaking rate. In contrast, we model the speaking rate explicitly from a phonetically aligned speech database, for example, in terms of average duration of phones. This way of modeling the speaking rate is easily incorporated in voice fonts and used at the time of personality transformation.

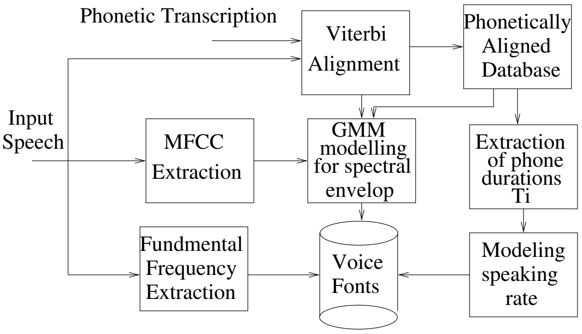
Phonetic Transcription
Viterbi Alignment
Phonetically Aligned Database
Input Speech
MFCC Extraction
GMM modelling for spectral envelop
Extraction of phone durations Ti
Fundmental Frequency Extraction
Voice Fonts
Modeling speaking rate

Figure 1: *Creation of Voice Fonts*

A significant amount of research has also been done on the duration of various acoustic units and its dependency on different aspects of speech [8, 9]. In the text-to-speech domain also, extensive research has been performed on duration modeling of the speech units. In concatenative text-to-speech synthesis, the duration of the concatenation unit is determined based upon phonetic segment identity of the unit, identities of the surrounding segments, syllabic stress of the unit, word importance of the word containing the unit, location of the syllable in the word and so on [10].

In the context of voice fonts, we want to model *speaker dependent* variations in the duration of acoustic units. The other dependencies of duration, e.g., stress, position etc., are not important as the textual dependencies remain same for the transformed speech. In other words, their effect is already accommodated for, either by the source speaker in case of real speech or by text-to-speech synthesis system in case of synthesized speech.

3. Voice Fonts and Personality Transformation

In [3], we proposed the concept of voice fonts to represent individual characteristics of a person's voice. We in particular focus upon three descriptors in voice fonts, *viz.*, fundamental frequency, spectral envelope and speaking rate. Figure 1 shows the schematic representation for creation of voice fonts.

3.1. Creation of Voice Fonts

Voice fonts for an individual are created from a speech database which is in the form of continuous sentences recorded in the individual's voice. A phonetically aligned speech database of these sentences is created through Viterbi Alignment procedure. Hidden Markov Models (HMM) of a speech recognizer are used to align the speech with the transcription of the corresponding text. The extraction of voice font descriptors is discussed below.

3.1.1. Fundamental Frequency

Fundamental frequency is computed for every voiced frame using a standard autocorrelation based pitch detector, similar to the pitch detector described in ITU-T standard G.729. An average of the fundamental frequency, computed over all the voiced frames in the speech database, constitutes this descriptor of voice fonts.

3.1.2. Spectral Envelop

For spectral envelope representation, we use 16 dimensional MFCC feature vectors, extracted from the speech signal by discrete regularized cepstrum method using a warped frequency scale [2]. The phonetic alignments from the speech database are used to create phone (or diphone) based Gaussian Mixture Models (GMM) of the MFCC vectors. More details about the spectral envelope representation and its conversion to that of the desired speaker can be found in [3].

3.1.3. Speaking Rate

Modeling of speaking rate, which is the main subject of this paper, is described in Section 4.

3.2. Personality Transformation

Personality transformation is performed, from source speaker to target speaker, by transforming three characteristics of the input speech signal, *viz.*, fundamental frequency, spectral envelope and speaking rate. This is achieved through an integrated signal processing module which uses Harmonic + Noise model (HNM) to analyze the speech signal [11]. HNM assumes the speech spectrum to be divided into two bands separated by maximum voiced frequency. The lower band is modeled by a sum of harmonically related sinusoids with slowly varying frequency and amplitude, and the upper band is modeled by modulated noise. This can be represented as

$$\hat{s}(t) = \sum_{k=-L(t)}^{k=L(t)} A_k(t) \exp\left(jktw_o(t)\right) + e(t) \qquad (1)$$

where $w_o(t)$ is the fundamental frequency, $A_k(t)$ is the harmonic amplitude and $L(t)$ is number of harmonics in the voiced band of the speech signal at time t. $e(t)$ models the noise part of the signal.

A sampling frequency of 16 kHz was used in the analysis with frame length equal to current pitch period. For every analysis frame, harmonic amplitudes $A_k(t)$s are computed using $w_o(t)$ by least square optimization between the actual speech signal and the harmonic part of $\hat{s}(t)$. The noise part is modeled by an all pole filter of order 15, whose coefficients are extracted from a 40 ms speech window, centered around the analysis frame. More details about the HNM framework can be found in [11].

For every analysis frame, first the spectral envelope is converted using phone based GMMs, present in the voice fonts for the source and target speakers [3]. After converting the spectral envelope, the fundamental frequency is scaled using the HNM framework to match the fundamental frequency of the target speaker. Finally, the frame is time scaled as described in the next section.

4. Proposed Approach

We model speaking rate for a person in the form of average duration of various acoustic units and categories, measured from the speech sentences recorded by the person. From the phonetically aligned speech database, we first compute average duration for a phone over all of its instances in the database. The average duration T_i, for a phone i can be represented as follows:

$$T_i = \frac{1}{N} \sum_{j=1}^{N} t_{ij} \qquad (2)$$

where t_{ij} denotes the duration of j^{th} instance of phone i uttered by the speaker. Averaging t_{ij} over all the instances also eliminates text dependent effects on phone duration.

These average phone durations, T_is, can be used in a number of ways to represent the speaking rate. One obvious way to use T_is is to represent average duration of each phone separately. At the time of personality transformation each phone is time scaled by its corresponding scale factor between source and target speakers. This turns out to be a very delicate task as at the time of transformation, recognizing each phone with high accuracy and time scaling it accordingly, is very difficult. If a phone is time scaled with wrong time scaling factor, this can result in perceptual disorders. At the other extreme, T_is can be averaged over all the phones to obtain a single representation of average duration, for all phones. This approach will result in uniform time scaling of the speech utterance. This, however, turns out to be a very coarse representation of the speaking rate with very low flexibility available for the transformation.

We use an intermediate approach to use T_is for voice fonts. We divide all the phones into three acoustic categories, *viz.*, voiced phones, unvoiced phones and silence phone. For a given category, we take an average of all T_is belonging to the category and use it to represent the average duration for the category. The silence phone, which represents a pause between words (inter-word silence), has another important reason to be separately modeled as described later in this section. This intermediate approach to represent the speaking rate provides following benefits:

- During personality transformation, the speech signal can be accurately segmented into voiced, unvoiced and silence regions and time scaled using their corresponding scale factors. This is very accurate and computationally less expensive as compared to recognizing each phone separately and scaling it.
- This provides flexibility in terms of separate time scaling for these acoustic categories as compared to uniform time scaling for the whole speech utterance.

	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Voiced (ms)	85.1	108.5	104.9	85.3
SD	39.1	50.39	49.1	39.9
Unvoiced (ms)	98.0	131.8	119.4	103.0
SD	36.7	54.81	53.5	38.0
Silence (ms)	150.6	252.6	146.2	202.7
SD	160.67	241.4	220.6	386.1
Avg. Silence	0.014	0.057	0.041	0.026

Table 1: Average phone durations and silence frequency

We observed major variation in the duration and frequency of inter-word silence (pause) among various speakers. It was observed that inter-word silence durations for various speakers were very different as compared to their actual speaking rates. Average number of such pauses, present in a speech utterance, also varied a lot among speakers. Average durations for the three acoustic categories and their corresponding standard deviations (SD) are shown in Table 1, for four different speakers. The last row of the table shows the average number of silence phones spoken by the speaker, in a speech utterance, as a fraction of non-silence phones. This relates to how frequently the speaker pauses between words. Due to this phenomenon, we modeled inter-word silence as a separate category. This can also be considered as a first step towards modeling speaking style.

4.1. Speaking Rate Transformation

At the time of personality transformation, the analysis frame is first classified into voiced, unvoiced or silence frame using the following functions:

$$S_1 = w_1 * (T_{zc1} - zc) + w_2 * E_{avg}/E_{frame} \quad (3)$$
$$S_2 = w_3 * (T_{zc2} - zc) + w_4 * (AC_{max} - T_{acr}) \quad (4)$$

where T_{zc1} and T_{zc2} are thresholds on zero crossing rate, E_{avg} is average energy of a frame for the speech utterance, E_{frame} is the energy of the current analysis frame, zc denotes current zero crossing rate, AC_{max} is maximum autocorrelation of the frame and T_{acr} is a threshold on the maximum autocorrelation. w_i's are weights given to the individual components in calculating the overall score. First, (3) is used to distinguish silence frames from non-silence frames. If S_1 is below a threshold, the frame is chosen as a silence frame. In case of non-silence frames, (4) is used to distinguish between voiced and non-voiced frames. Once the frame is classified, it is time scaled using a scaling factor obtained by dividing the average duration of the acoustic category in the voice font for the target speaker by that of the corresponding category in the voice font of the source speaker.

4.1.1. Handling Intra-Word Silence

At the time of transformation, it is difficult to discriminate inter-word silence from intra-word silences, present mostly before a voiced or unvoiced stop sound. Both of them are identical in their acoustical properties. The discrimination is very important as otherwise the silence present inside the words will also be scaled by the same scaling factor as of the inter-word silence. This results in a very annoying speech signal as can be seen from the results in Section 6.

There are two ways to discriminate between these silences. In the first approach, speech signal is time aligned with its corresponding text to determine the boundary of the words. This can be achieved through Viterbi alignment through a speech recognition system in a way similar to what was used to create voice fonts. This is the most accurate way of discriminating inter-word silence from intra-word silence. However, it puts a very stringent requirement of getting the text corresponding to the speech and having the acoustic models of a speech recognition system at the time of transformation. The second approach makes use of the fact that the inter-word silence is generally of longer duration as compared to the intra-word silences. This approach thus uses a threshold based on mean and standard deviation of the inter-word silence duration to discriminate. Results for both the approaches are presented in Section 6.

5. Experiments

All the experiments were conducted on continuous Hindi speech sentences of 4 to 7 seconds duration. To train the system, 30 minutes of speech was recorded from 4 different speakers. These four speaker were speaking different sentences in Hindi. All the sentences were phonetically aligned using Viterbi Alignment through Hidden Markov Models of a large vocabulary continuous Hindi speech recognizer, developed at IBM India Research Lab, for dictation task. It uses a Hindi phone set containing 61 phones and a vocabulary of more than 65,000 words. It gives more than 95% word recognition rate with a trigram language model [12]. Using the aligned speech database, voice fonts were created by computing the descriptors as described in Section 3. For uniform time scaling, T_is were averaged over all the phones

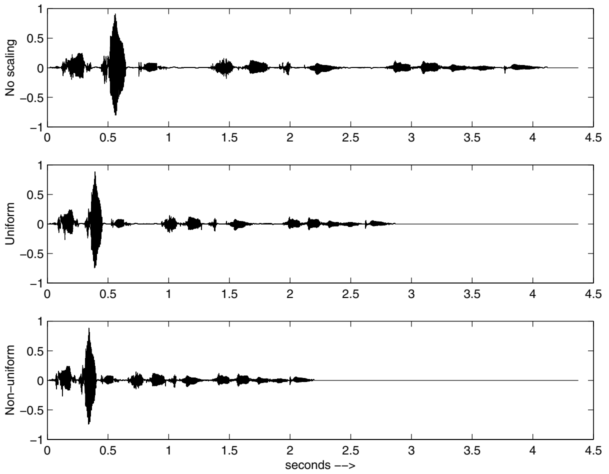

Figure 2: *Time scaled speech waveforms*

	NS	US	D1	D2	ND
DCR Test	2.65	3.20	3.85	3.75	2.00
Opinion Test	6.75	7.25	8.35	8.15	5.20

Table 2: Results for subjective tests

and for non-uniform scaling three separate average durations were used for voiced, unvoiced and silence categories.

To distinguish among voiced, unvoiced and silence regions, 10 millisecond frames were analyzed using (3) and (4). T_{zc1} was set to 4. A value of 25 was used for T_{zc2} when the previous frame was unvoiced and 30 in case of voiced. T_{acr} was set to 0.5 when the previous frame was voiced and 0.6 in case of unvoiced. Two different values of the thresholds were chosen to obtain a smooth transition between voiced and unvoiced regions.

We conducted subjective tests to evaluate the performance of the various approaches used in the experiments. We applied personality transformation on two different sentences for two different source-target pairs using five possible approaches. This resulted in a total of 20 synthesized sentences. The synthesized sentences can be accessed at *http://in.geocities.com/ashish_verm/euro03.html*. Eight subjects were asked to listen to these sentences and rate them. In Degradation Category Rating (DCR) test, subjects were asked to rank the synthesized sentence on a scale of 1 to 5, representing decreasing level of degradation from speaking rate and speaking style of the target person. In the second test, that we call opinion test, subjects were asked to rank the synthesized sentences on a scale of 1 to 10, considering closeness of the synthesized speech, in overall quality, to the target speaker.

6. Results and Discussion

Results corresponding to the subjective tests are presented in Table 2. NS and US approaches correspond to no time-scaling and uniform scaling. D1 and D2 represent the acoustic category based time scaling. Phonetic alignments were used in D1, and mean and standard deviation were used in D2, to discriminate between inter-word silence and intra-word silence. ND refers to the case when they were not discriminated.

All of the approaches, except ND, perform better as compared to the case when no time scaling is performed. This assures that time scaling has not degraded the speech signal. Uniform scaling (US), which uses a single speaking rate, under performs as compared to the case when separate speaking rates are used for the three acoustic categories. However, it performs better than the no time-scaling (NS) case. The two different ways to distinguish inter-word silence from intra-word silence perform very close to each other. Finally we see that if we do not discriminate between inter-word silence and intra-word silence, in case of non-uniform scaling, the quality of the transformed speech goes down drastically and is even worse than no time-scaling case.

An example of the time scaled speech utterances is shown in Figure 2. The topmost waveform is without scaling while the middle and bottom waveforms are time scaled using uniform and non-uniform scaling respectively. It can be noticed that silence regions are scaled by a different factor, as compared to that of the non-silence regions, in case of non-uniform scaling.

7. References

[1] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., “Voice conversion through vector quantization", in *Proc. ICASSP*, pp. 655-658, Washington, April 1998.

[2] Stylianou, Y., Cappe, O., and Moulines, E., “Continuous probabilistic transform for voice conversion", *IEEE Transaction on Speech and Audio Processing*, pp. 131-142, Vol. 6, No. 2, March 1998.

[3] Kumar, A. and Verma, A., “Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," in *Proc. ICASSP*, HongKong, April 2003.

[4] Morgan, N., Fosler, E., and Mirghafori, N., “Speech recognition using on-line estimation of speaking rate", in *Proc. EUROSPEECH* 1997, pp 2079-2082, Greece, 1997.

[5] Morgan, N. and Fosler, E., “Combining multiple estimators of speaking rate", in *Proc. ICASSP*, pp. 729-732, Washington, April 1998.

[6] Kitazawa, S., Ichikawa, H., Kobayashi, S., and Nishinuma, Y., “Extraction and representation rhythmic components of spontaenous speech", in *Proc. EUROSPEECH 1997*, pp. 641-644, Greece, 1997.

[7] Faltlhauser, R., Pfau, T. and Ruske, G., “On-line speaking rate estimation using gaussian mixture models", in *Proc. ICASSP*, pp. 1355-1358, Istanbul, Turkey, 2000.

[8] Umeda, N., “Vowel duration in American English", *Jounral of Acoustical Society of America*, pp. 434-445, Vol. 58, No. 2, August 1975.

[9] Umeda N., “Consonant duration in American English", *Journal of Acoustical Society of America*, pp. 846-858, Vol. 61, No. 3, March 1977.

[10] Sproat, R., *Multilingual text-to-speech synthesis: The bell labs approach*, Boston, Kluwer Academic Publishers, 1998.

[11] Stylianou, Y., Laroche, J. and Moulines, E., “High-quality speech modification based on a harmonic + noise model," *Proc. of EUROSPEECH*, pp. 451-454, Madrid, Spain, 1995.

[12] Neti, C., Rajput, N. and Verma, A., “A large vocabulary continuous speech recognition system for Hindi", Proc. *National Conference on Communication*, Mumbai, January 2002.