interspeech 2008 incorporating SST 08

High-Quality Analysis/Synthesis Method Based on Temporal Decomposition for Speech Modification

Binh Phu Nguyen, Takeshi Shibata, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
{npbinh, s0610043, akagi}@jaist.ac.jp

Abstract

The challenge of speech modification is to flexibly modify the speech without degrading speech quality. The conventional methods are limited by their inability to flexibly control speech signals in time and frequency domains. This causes degradation of the quality of modified speech. This paper proposes a high-quality analysis/synthesis method for speech modification. To control the temporal evolution, we use a speech analysis technique called temporal decomposition (TD), which decomposes a speech signal into event targets and event functions. The same event functions evaluated for the spectral parameters are also used to model the temporal evolution of the excitation parameters. The event functions describe the temporal evolution of the spectral and excitation parameters, and the event targets represent the "ideal" spectral parameters. To flexibly control speech signals in both time and frequency domains, we propose new methods to model the event functions and the event targets. The experimental results show that our proposed analysis/synthesis method produces high-quality synthesized speech, and allows the flexibility to modify speech signals.

Index Terms: analysis/synthesis method, speech modification, temporal decomposition

1. Introduction

The aim of speech modification is to modify attributes of speech. Time-scale and spectral modifications are core processes in speech modification.

Time-scale modification is used to alter the signal's apparent time-evolution without affecting the quality, pitch or naturalness of the original signal. This kind of technique can be used in many applications, such as slowing down or speeding up the playback rate in foreign language learning, compressing data for communications or storage, altering speaking rate in Text-to-Speech systems, etc.

Several approaches are available in the literature for time-scale modification. Altering the time-scale of a speech signal can be achieved in the time domain [1], or frequency domain [2]. Time-domain techniques are based on overlap-add (OLA) methods [1]. These techniques first segment the waveform into a series of overlapping frames by windowing the speech signal with a suitable window function. To perform time-scale modification, some of the windowed segments are either replicated or omitted. In these cases, the information about the pitch markers is not used for splitting the speech signal into short segments. As a result, the periodicity due to pitch is not preserved well after time-scale modification [3]. Therefore, these techniques tend to perform poorly when large modification factors must be used (e.g., factors greater than $\pm$ 20% to $\pm$ 30%) [4]. Frequency-domain techniques are based on short-time Fourier transform (STFT) or phase vocoder methods [2]. These algorithms require high computation costs, but are capable of providing high-quality output. However, they still suffer from some distortion, mainly due to the effects of "phase dispersion" [5]. That is, while the scaled signal has the same frequency, the phases between the components change, resulting in a different wave shape. In the STRAIGHT method [6], the analysis algorithm does not extract phase information. Its reconstruction algorithm adopts the minimum phase assumption for the spectral envelope, and further applies all-pass filters to reduce the buzz timbre of the reconstructed signal. This method offers high-quality modified speech signals without introducing the artificial timbre. However, this approach still processes speech signals frame by frame, and speech manipulation is performed by using interpolation functions. This method does not consider the temporal evolution of parameters when modifying speech signals.

Spectral modification is used to perform a variety of modifications to speech spectra, such as modifications of formant structures, amplitude, etc. This kind of technique can be used in many applications, such as transforming the identity of a speaker, enhancing speech, etc.

A variety of spectral modification methods have been discussed in the literature. They can be classified into two major approaches: LP-based methods [7, 8], and frequency warping methods [9]. LP-based methods often meet the pole interaction problem suffered by pole modification techniques. An iterative algorithm for overcoming pole interaction during formant modification was developed by Mizuno et al. [7]. While this method produces spectral envelopes with desired formant amplitudes at the formant frequencies, one drawback to this technique is that the bandwidth of each formant cannot be controlled. Recently, a method for modifying formant locations and bandwidths directly in the line spectral frequency (LSF) domain has been developed in [8]. By taking advantage of the nearly linear relationship between the LSFs and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths. However, the main drawback to this type of modification, the lack of control over the spectral shape, has not been solved. Frequency warping methods, such as [9], provide high-quality modified speech. However, the modification is still not successful, because frequency warping methods do not allow merging or splitting the spectral peaks, which is often desired in spectral modification.

In addition, two methods mentioned above [7, 8] only mention the way to perform the spectral modification in a frame, and they [7, 8, 9] rarely deal with constraints between frames after modification. When there are unexpected modifications in some frames, the modified speech may be not smooth. As a result, there are some clicks in the modified speech, which lead to degradation of speech quality.

In this paper, we propose a high-quality analysis/synthesis method based on temporal decomposition [10, 11] for speech modification. Temporal decomposition (TD) is a technique to decompose a speech signal into event targets and event functions. To flexibly control the speech signals in both time and frequency domains, we introduce new methods to model the event functions and event targets. We then explain how to modify duration and speech spectra in our proposed analysis/synthesis method.

Accepted after peer review of full paper

2. Temporal decomposition

A shortcoming of conventional speech modification methods is that they do not take into account the correlation between frames, which makes it difficult to model and control the temporal trajectories of parameters of speech signals. Therefore, we employ TD to deal with the problem.

Atal proposed a method based on the temporal decomposition of speech into a sequence of overlapping event functions and corresponding event targets [10], as given in Eq. (1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \leq n \leq N \tag{1}$$

where $\mathbf{a}_k$ is the spectral parameter vector corresponding to the k^{th} event target. The temporal evolution of this target is described by the k^{th} event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of the n^{th} spectral parameter vector $\mathbf{y}(n)$, and is produced by the TD model. *N* and *K* are the number of frames in the speech segment, and the number of event functions, respectively ($N \gg K$).

The original method of TD is known to have two major drawbacks, high computational costs, and high parameter sensitivity to the number and locations of events. A number of modifications have been explored to overcome these drawbacks. In this study, we employ the MRTD algorithm [11]. The reasons for using the MRTD algorithm in this work are twofold: (i) the MRTD algorithm enforces a new property on event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [11]; (ii) event targets can convey the speaker's identity [12]. In the MRTD algorithm, LSF parameters are chosen for the input of TD, because LSFs have good linear interpolation attributes.

In the MRTD algorithm, the same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the excitation parameters. Let $b(n)$ be an excitation parameter, i.e. F0, gain, and aperiodic component (AP), in the n^{th} frame. $b(n)$ can be approximated by using event target of excitation $\mathbf{b}_k$ and its event function $\phi_k(n)$, as follows.

$$\hat{\mathbf{b}}(n) = \sum_{k=1}^{K} \mathbf{b}_k \phi_k(n), \qquad 1 \leq n \leq N \tag{2}$$

where $\phi_k(n)$ is estimated from Eq. (1).

Since the same event functions evaluated for the spectral parameters are also used to model the temporal pattern of the excitation parameters, we only need to modify these targets, $\mathbf{a}_k$ and $\mathbf{b}_k$, and the corresponding event functions $\phi_k(n)$ for modifying the speech signals, instead of modifying the speech signals frame by frame. The smoothness of modified speech will be ensured by the shape of the event functions $\phi_k(n)$. This leads to easy modification of the speech signals in time-frequency domain, as well as ensuring the smoothness of the speech signals between frames, and thereby enhances the quality of modified speech.

3. Modeling of the event function using polynomial fitting

3.1. Identifying the event locations

MRTD algorithm uses a spectral stability criterion to determine the initial event locations [11]. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter sets can be used as good approximations of event locations and event targets, respectively. This algorithm is automatically performed, and the subsequent computation of refined event targets and event functions is much less demanding than the traditional TD method. This algorithm is useful for applications in speech coding [11], and speaker identification [12]. However, the event functions calculated by using this algorithm are difficult to model and control.

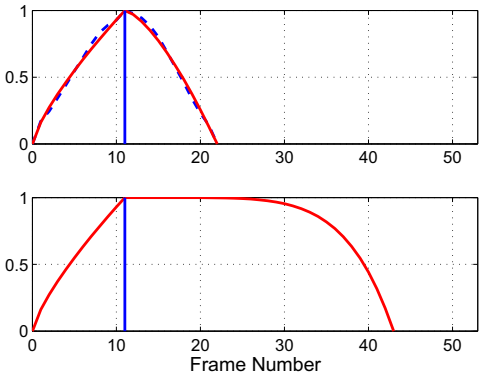

Figure 1: A fitted curve (the solid line) using the non-linear least square method for an event function (the dashed line) (top), and a time-scale modification of this event function (bottom).

This paper proposes a new method for determination of event locations based on phonemes. Although automatic phoneme segmentation is a significant problem, we do not deal with it in this paper. We use labeled data of utterances to segment speech signals into phonemes. Each phoneme is divided into eight equal segments, and the nine points marking these segments are used for identifying the event locations. In our method, the window shift is set to 1 ms. Therefore, in each speech segment (a phoneme), the number of event functions K is nine, and the number of frames N is equal to the duration of this phoneme (ms). Note that we can easily increase the quality of synthesized speech by increasing the number of event functions in each phoneme.

3.2. Proposed method

To control the event functions, the event functions should be modeled. The MRTD algorithm enforces the "well-shapedness" property of event functions. That is, the event functions in the MRTD are monotonic during the transition from one event towards the next. In addition, the MRTD method employs the restricted second order TD model, in which only two event functions at any moment of time can overlap and all event functions sum up to one [11]. Therefore, to model the event function, polynomial fitting for the event function is performed by using the nonlinear least square method as follows.

$$Z = -\left(\frac{X}{c}\right)^{M} + e \tag{3}$$

where e is the maximum value of ϕ, and e is equal to 1. Z is equal to 0 when

$$X = c \tag{4}$$

where c is the duration of two consecutive events. The polynomial fitting was done in $0 \leq \phi \leq 1$. The value of M indicates slope of event function. Shape of the event function can be changed according to the values of c and M. As a result, it is possible to control the event function. Fig. 1 shows an example of modeled event function by the proposed method for an event function extracted by MRTD.

3.3. Time-scale modification

Since F0, gain, AP, and spectral parameters are decomposed by using the same event functions, in order to perform time-scale modification, we only need to change length of each event function. From Eq. (3), we can modify the duration of the speech

segment by changing the value of c. We modify the values of M to alter the slope of the event function. By changing the values of c and M, we can control the evolution of all important parameters of speech signals (i.e. F0, gain, AP, and spectral parameters). Therefore, it enhances the quality of modified speech. An example of time-scale modification of an event function is also shown in Fig. 1. In this example, to show the flexibility of our proposed method, we only altered the shape of the right side of the event function by the modification factors of c and M 3 and 3.9 times, respectively.

4. Modeling of the event target using Gaussian mixture model

4.1. Proposed method

In the MRTD algorithm, event targets are valid LSF coefficients. However, this kind of representation is limited by the inability to independently control important formant characteristics such as amplitude and bandwidth, or to control the spectral shape.

Zolfaghari et al. proposed a technique to fit a Gaussian mixture model to the smoothed magnitude spectrum of a speech signal [13, 14, 15]. The parameters of Gaussian mixture model are called spectral-GMM parameters in this paper. The ability to independently control the parameters of each Gaussian component enables precise estimation of the spectral envelope, enables a wide variety of modifications, and enables independent control of the formants. However, the original method does not ensure a one-to-one correspondence between spectral peaks and Gaussian components. This creates difficulty in modifying the speech spectrum in both dimensions, frequency and amplitude. To overcome this drawback, we propose an improvement in modeling the speech spectrum for speech modification. The aim of our proposed method is not only to model the speech spectrum well, but also to ensure a one-to-one correspondence between spectral peaks and Gaussian components. Our proposed method is as follows.

First, from a speech spectrum, we start to estimate spectral-GMM parameters with the initial 18 Gaussian components. The requirement of the initial number of Gaussian components is that the number of components be high enough to model the speech spectrum well. Note that the initial Gaussian components depend on the sampling frequency. We assume that if the linear sum of two Gaussian components has only one peak, these two Gaussian components are dependent, and also that, if the linear sum of two Gaussian components has two peaks, these two Gaussian components are independent. After we get the spectral-GMM parameters in the first iteration, we check whether or not all Gaussian components are independent components. If not, we divide the spectral-GMM parameters into two groups. The first group models the spectral shape of the speech spectrum, and the other group models the spectral peaks of the speech spectrum. On the basis of the geometric characteristics of normal distribution, i.e. the empirical rule, we assume that a Gaussian component m is a spectral shape factor if there are at least two other Gaussian components located between $[-3\mu_m, 3\mu_m]$, where μ_m is the mean of this Gaussian component m. If two Gaussian components i, j are dependent spectral peaks, we merge these two Gaussian components by the following equations.

$$\mu_{ij} = \frac{\omega_i \mu_i + \omega_j \mu_j}{\omega_i + \omega_j} \tag{5}$$

$$\sigma_{ij}^2 = \frac{\omega_i(\sigma_i^2 + (\mu_{ij} - \mu_i)^2) + \omega_j(\sigma_j^2 + (\mu_{ij} - \mu_j)^2)}{\omega_i + \omega_j} \tag{6}$$

where μ_i, σ_i and μ_j, σ_j are the mean, and standard deviation of Gaussian components i and j, respectively. After merging Gaussian components, a new process of estimating spectral-GMM parameters is executed, with the condition that the initial parameters for the new process are current Gaussian components.

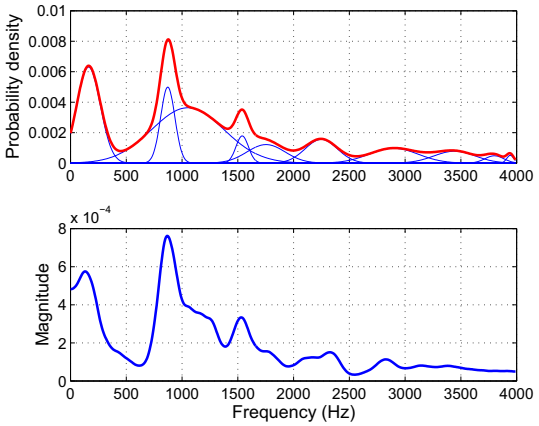

Figure 2: Example of the spectral envelope restored by our proposed method: the thin lines indicate Gaussian components, the bold line indicates the restored spectral envelope (top), and STRAIGHT spectral envelope (bottom).

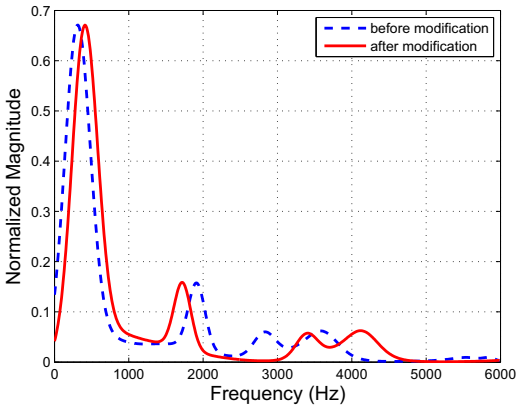

Figure 3: Example of our spectral modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

The process of spectral-GMM estimation continues to iterate, and terminates when all Gaussian components are independent components. Obviously, this algorithm always converges, since after each iteration, the number of Gaussian components decreases by at least 1. An example of the spectral envelope restored by our proposed method is illustrated in Fig. 2.

4.2. Spectral modification

Formant frequency is one of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters.

We already proposed a new algorithm for modifying spectral-GMM parameters in accordance with formant frequencies [16]. In our proposed algorithm, we can independently modify each peak. Also, we can control the spectral shape, which is difficult to do using the conventional spectral modification methods. An example of our proposed algorithm applied to a spectrum is shown in Fig. 3.

5. Proposed speech analysis/synthesis method

In Sections 3 and 4, we proposed new methods to model both the event functions and the event targets for flexibly controlling them. To modify the speech signals, we only need to modify event targets and event functions, and the smoothness of modified speech will be ensured by the shape of the event functions. In this section, we propose a new high-quality analysis/synthesis method for speech modification based on new models of the event functions and the event targets. The processing flow of our proposed method is as follows.

First, STRAIGHT [6] decomposes input speech signals into spectral envelopes, F0, and AP. Since the spectral envelopes can be further analyzed into LSF parameters, MRTD [11] is employed in the next step to decompose the LSF parameters into event targets and event functions. The event functions are modeled by using Eq. (3). Since the event targets are valid LSF parameters [11], the spectral envelope of each event target can be restored, and then the spectral envelopes are converted to spectral-GMM parameters. By using spectral-GMM parameters to model the event targets, we can flexibly perform modifications of the event targets. The same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the F0, gain, and AP. The modified event targets are then re-synthesized as modified LSF by TD synthesis. In the following step, the modified LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the synthesized speech. Note that our proposed method is integrated with the STRAIGHT method, it therefore can use the merits of the STRAIGHT to modify the F0.

6. Experiments and Results

Since we use spectral-GMM parameters to model each event target, the order of LSFs has to be high enough to precisely restore the spectral envelope. Via a small experiment, by calculating the average log spectral distortion (LSD) between STRAIGHT spectra and the spectral envelopes restored from LSFs with different orders in a set of 250 sentence utterances of the ATR Japanese speech database [17] at sampling frequency of 16 kHz, we chose the LSF order of 40 in this paper. With this order, the average LSD is smaller than 1 dB.

A set of 100 sentence utterances of the ATR Japanese speech database [17] was selected as the speech data. This speech dataset is spoken by 4 speakers (2 male & 2 female) re-sampled at 16 kHz sampling frequency.

To evaluate the performance of our proposed analysis/synthesis method, we compare the quality of synthesized speech restored by our method with that of the framewise-GMM method [13]. In the framewise-GMM method [13], we estimated the spectral-GMM parameters from the STRAIGHT spectrum frame by frame. We used the perceptual evaluation of speech quality (PESQ) score (ITU-T P.862) to evaluate the quality of synthesized speech. Having high correlation ($\rho > 0.92$) with subjective listening tests, the PESQ can be used reliably to predict the subjective speech quality [18]. The score of PESQ ranges from -0.5 to 4.5. The higher the score, the better the perceptual quality. In this section, we also calculated the average PESQ of the synthesized speech restored by STRAIGHT for the reference. We used the original sounds as the reference signals, and the synthesized utterances restored by STRAIGHT, the framewise-GMM method, and our proposed method as the degraded signals. Since the average number of Gaussian components in our proposed method is 9.2, we chose 9 Gaussian components to model each speech spectrum in the framewise-GMM method. The average PESQ results are shown in Table 1. These results indicate that the quality of synthesized speech of our proposed method is better than that of the framewise-GMM method. Moreover, in our proposed method, both the speech spectra (i.e. the spectral evolution and the speech spectrum) and the temporal evolution of the excitation parameters (i.e. F0, gain, and AP) can be modeled, which gives the flexibility to control these parameters.

Table 1: Average PESQ for analysis/synthesis methods.

STRAIGHT method	3.5551
Framewise-GMM method	3.0294
Proposed method	3.3241

7. Conclusions

In this paper, we have presented a high-quality analysis/synthesis method based on temporal decomposition for speech modification. The same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the F0, gain, and AP. We then model the event functions by using polynomial fitting, and event targets by using spectral-GMM parameters. These models give the flexibility to control the speech signals in both time and frequency domains. The experimental results have shown that the quality of the reconstructed speech signal is high, and we can flexibly perform both duration and spectral modification. In our future work, we will investigate our proposed method for voice conversion, and for transformation of speaking voice into singing voice.

8. Acknowledgments

This study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan. We would like to thank Prof. Mary Ann Mooradian for checking our English.

9. References

[1] Verhelst, W. and Roelands, M., "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," Proc. ICASSP, 554–557, 1993.
[2] Dolson, M., "The phase vocoder: A tutorial," Computer Music Journal, 10: 14–27, 1986.
[3] Sreenivasa Rao, K. and Yegnanarayana, B., "Prosody modification using instants of significant excitation," IEEE transactions on audio, speech and language processing, 14: 972–980, 2006.
[4] Laroche, J. and Dolson, M., "Improved phase vocoder time-scale modification of audio," Speech and Audio Processing, IEEE Transactions on, 323–332, 1999.
[5] Ninness, B. and Henriksen, S., "Time and frequency scale modification of speech signals," Proc. ICASSP, 1295–1298, 2000.
[6] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27: 187–207, 1999.
[7] Mizuno, H., Abe, M., and Hirokawa, T., "Waveform-based speech synthesis approach with a formant frequency modification," Proc. ICASSP, 195–198, 1993.
[8] Morris, R.W. and Clements, M.A., "Modification of formants in the line spectrum domain," IEEE Signal Processing Letters, 9: 19–21, 2002.
[9] Turajlic, E., Rentzos, D., Vaseghi, S., and Ho, C.H., "Evaluation of methods for parameteric formant transformation in voice conversion," Proc. ICASSP, 724–727, 2003.
[10] Atal, B.S., "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP, 81–84, 1983.
[11] Nguyen, P.C., Ochi, T., and Akagi, M., "Modified restricted temporal decomposition and its application to low bit rate speech coding," IEICE Transactions on Information and Systems, E86-D: 397–405, 2003.
[12] Nguyen, P.C., Akagi, M., and Ho, T.B., "Temporal decomposition: A promising approach to VQ-based speaker identification," Proc. ICASSP, 184–187, 2003.
[13] Zolfaghari, P. and Robinson, T., "Formant analysis using mixtures of Gaussians," Proc. ICSLP, 1229–1232, 1996.
[14] Zolfaghari, P., Watanabe, S., Nakamura, A., and Katagiri, S., "Bayesian modelling of the speech spectrum using mixture of Gaussians," Proc. ICASSP, 553–556, 2004.
[15] Zolfaghari, P., Kato, H., Minami, Y., Nakamura, A., Katagiri, S., and Patterson, R., "Dynamic assignment of Gaussian components in modelling speech spectra," Journal of VLSI Signal Processing Systems, 45: 7–19, 2006.
[16] Nguyen, B.P. and Akagi, M., "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," Proc. Interspeech, 538–541, 2007.
[17] Abe, M., Sagisaka, Y., Umeda, T., and Kuwabara, H., "Speech database user's manual," ATR Technical Report, TR-I-0166, 1990.
[18] Rix, A.W., Beerends, J.G., Hollier, M.P., and Hekstra, A.P., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," Proc. ICASSP, 749–752, 2001.