Detection and Synthesis of Transition Parts of The Speech Signal

Moo Young Kim* *Regular Member*

ABSTRACT

For the efficient coding and transmission, the speech signal can be classified into three distinctive classes: voiced, unvoiced, and transition classes. At low bit rate coding below 4 kbit/s, conventional sinusoidal transform coders synthesize speech of high quality for the purely voiced and unvoiced classes, whereas not for the transition class. The transition class including plosive sound and abrupt voiced-onset has the lack of periodicity, thus it is often classified and synthesized as the unvoiced class. In this paper, the efficient algorithm for the transition class detection is proposed, which demonstrates superior detection performance not only for clean speech but for noisy speech. For the detected transition frame, phase information is transmitted instead of magnitude information for speech synthesis. From the listening test, it was shown that the proposed algorithm produces better speech quality than the conventional one.

Key Words : Speech Coding, Speech Transmission, Speech Transition, Phase Quantization, Sinusoidal Transform Coding

Ⅰ. 서 론

Speech signals can be classified into three distinctive classes according to their modes of excitation: voiced, unvoiced, and transition classes[1]. The low-bit-rate harmonic coders, such as sinusoidal transform coder (STC)[2], improved multiband excitation (IMBE) vocoder[3], and spectrally mixed excitation (SMX) vocoder[4], are able to produce good synthesized speech quality for the first two classes that are characterized as stationarity. For the voiced frames, phase terms can be reconstructed on the assumption that the fundamental frequency evolves slowly. For the unvoiced frames, phase components are randomly generated. However, harmonic coders yield degraded speech quality for the transition frames since these frames are often classified as fully unvoiced ones and the typical random phase assignment for the speech synthesis results in the hoarse sound quality[5,6]. Transition frames should not be categorized as either voiced or unvoiced frames. These frames should be detected and synthesized separately to obtain the high quality of speech.

To detect the transition frames, the absolute peakiness value of the linear predictive coding (LPC) residual signal was applied in [7]. However, under the background noise environment, the absolute peakiness value does not function well to detect transition frames. Thus, in this paper, a new method is proposed using the relative peakiness value.

Phase information is more important to synthesize the transition frames, while magnitude information is more crucial for stationary frames[8]. Thus, it would be advantageous to utilize phase information in the transition frames at the sacrifice of magnitude information to some extent. Transitional parts of speech can be modeled by manipulated spectral phase rather than magnitude information, which makes the coder generate

* Information and Communications Eng., Sejong University (mooyoung@sejong.ac.kr)

논문번호 : KICS2008-02-093, 접수일자 : 2008년 2월 22일, 최종논문접수일자 : 2008년 3월 17일

plosive and abrupt onset sequences effectively, thus the synthesized speech quality can be more crisp by reducing hoarse artifacts.

Ⅱ. Noise-Robust Detection Of The Transition Class

In [7], to identify the transitional part of speech, the absolute peakiness value with the sliding window was proposed:

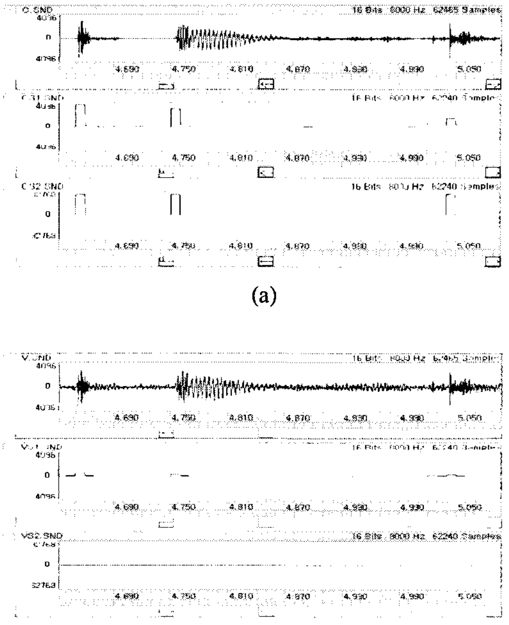
$$P = \max_{-T_s \le i < T_s} P_i \quad , \qquad (1)$$

where T_S denotes the range of the sliding window and the absolute peakiness value, P_i, is given by

$$P_i = \frac{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left|r(n+i)\right|^2}}{\frac{1}{N}\sum_{n=0}^{N-1}\left|r(n+i)\right|} \quad , \qquad (2)$$

where $r(n)$ and N are the LPC residual and subframe size, respectively. A frame is classified as transition if P is greater than the predetermined threshold P_{th}. In the experiment, N, T_S, and P_{th} are set to be 80, 40, and 0.9, respectively. In Fig. 1, the absolute peakiness value with sliding window (middle) and the detected transitional frame (bottom) along with the corresponding input speech waveform (top) are illustrated. The performance of the above scheme is reasonably good for clean speech as shown in Fig. 1 (a). However, for noisy speech (15dB vehicular noise in this example), the detection performance is decreased significantly (bottom) because of the relatively low peakiness value (middle) as shown in Fig. 1 (b). If P_{th} is adjusted to fit for noisy environment, the false alarm rate will be increased for clean speech.

To solve this problem, an improved transitional part detection algorithm is proposed in this paper.



(a)

(b)

Fig. 1 Detection of transition frames by using [7] for (a) clean speech and (b) noisy speech (15dB vehicular noise). For each figure, Top: input speech, Middle: absolute peakiness value with the sliding window, Bottom: detected transitional frame (marked as nonzero values), respectively.

The relative peakiness value is used instead of the absolute one. The following three steps are performed to detect the transition:

1) finding the peak-emphasized version of residual,
2) finding the relative peakiness value from 1),
3) and detecting the transition frame.

To obtain the peak-emphasized version of the LPC residual $\tilde{r}(n)$, the original residual $r(n)$ is rectified,

which is subtracted by its DC component, and which is center clipped as follows:

$$r'(n) = \left|r(n)\right| - \bar{r}, \quad n = 0,1,\cdots,N-1 \quad , \qquad (3)$$

$$\bar{r} = \frac{1}{N}\sum_{n=0}^{N-1}\left|r(n)\right| \quad , \qquad (4)$$

$$\tilde{r}(n)=\begin{cases} r'(n), \; if \; r'(n) > r_{th}, \\ 0, \quad otherwise, \end{cases} \quad n=0,1,\cdots,N-1 \quad . \tag{5}$$

Fig. 2 shows the example of (5), which clearly shows the starting peak position of the transition.

The peakiness value for the peak-emphasized version of the residual signal $\tilde{r}(n)$, instead of the original residual $r(n)$, is calculated as

$$P_i' = \frac{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left[\tilde{r}(n+i-N+1)\right]^2}}{\frac{1}{N}\sum_{n=0}^{N-1}\left|\tilde{r}(n+i-N+1)\right|} \tag{6}$$

Then, the relative peakiness value, $P'_i - P'_{i-j}$, is defined as the difference in peakiness values between two points i and $i-j$. If the number of times that the relative peakiness is greater than the threshold P'_{th} exceeds a certain threshold C_{th} within J samples, i is declared as the start position of the transition. In this paper, we fix to use P'_{th}, C_{th}, and J as 0.42, 2.0, and 20, respectively. This method is expressed as

$$\tilde{P}_i = \begin{cases} 1, & if \; Count\left(P_i' - P_{i-j}' > P_{th}'\right) > C_{th}, \quad for \; 1 \le j < J, \\ 0, & \qquad\qquad otherwise. \end{cases} \tag{7}$$

Finally, a frame is declared as transition if at least single $\tilde{P}_i$ is activated in the frame. However, if the pitch correlation values of the previous and current frames are high enough, which means that current frame includes stationary voiced signal, the current frame is classified into voiced region.

Fig. 3 shows the result of proposed method for the same noisy speech as shown in Fig. 1 (b). The top, middle and bottom parts of the figure are input speech, modified absolute peakiness, and relative peakiness by utilizing (7), respectively. The performance improvement is significant in the transition detection.

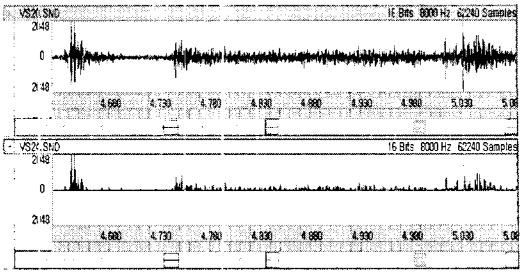


Fig. 2 Peak-emphasized signal for the transitional part detection in the 15dB vehicular noise environment. Top: residual signal. Bottom: peak-emphasized version of residual signal.

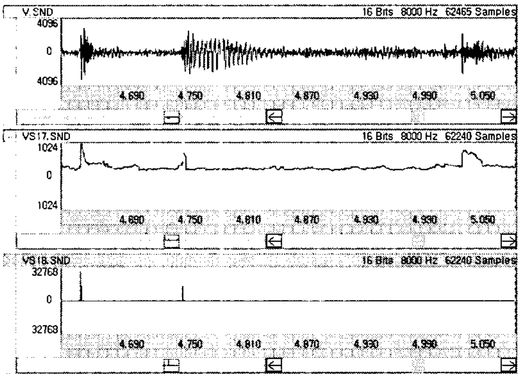


Fig. 3 The proposed transition detection method for the speech corrupted by 15dB vehicular noise. Top: input speech. Middle: modified absolute peakiness value. Bottom: relative peakiness value where non-zero positions correspond to the detected transition positions.

Ⅲ. Speech Synthesis For The Transition Class

In the conventional sinusoidal speech coding, synthesized speech can be obtained as

$$s(n) = \sum_{h=1}^{H} A_h(n)\cos(\theta_h(n)), \; n = 1,2,\cdots,N \, , \tag{8}$$

where H is the number of harmonics, magnitude spectrum $A_h(n)$ is estimated by linear interpolation between the current and the previous frames, and $\theta_h(n)$ is modeled by quadratic interpolation as

$$\theta_h(n) = \theta_h(0) + \int_0^n \left(\frac{N-\xi}{N} h\omega_0(0) + \frac{\xi}{N} h\omega_0(N)\right) d\xi + \Delta\theta_h(n), \tag{9}$$

where $\omega_0(0)$ and $\omega_0(N)$ are fundamental frequencies at each end of the previous and current frames, respectively, and $\theta_h(0)$ is the synthesized phase value at the end of the previous frame. The last term, $\Delta\theta_h(n)$, which is represented as

$$\Delta\theta_h(n)=\frac{n}{N}\left[\theta_h^{org}(N)-\theta_h(N)-2\pi\left\lfloor\frac{\theta_h^{org}(N)-\theta_h(N)}{2\pi}\right\rfloor\right], \tag{10}$$

is introduced to prevent phase discontinuity. $\theta_h^{org}(N)$ and $\theta_h(N)$ are the estimated original phase and the synthesized phase obtained by (9), respectively.

To synthesize the consistent quality of speech, not only magnitude information but careful phase synthesis technique is required, since it is impossible for low-bit-rate harmonic coders to transmit all the original phase information. In this paper, category based phase transmission technique is proposed. For the synthesis of a voiced stationary part of speech, original phase $\theta_h^{org}(N)$ in (10) can be replaced by zero phase $\theta_h^{zero}(N)$ which is represented as

$$\theta_h^{zero}(N)=\theta_h^{zero}(0)+\frac{hN}{2}(\omega_0(0)+\omega_0(N)), \tag{11}$$

where $\theta_h^{zero}(0)$ is the estimated zero phase at the end of the previous frame. Secondly, to reproduce the unvoiced stationary part such as a fricative sound and stationary background noise, random phase θ_h^{rand} can be utilized instead of $\theta_h^{org}(N)$ in (10). θ_h^{rand} is the uniformly distributed random value in $(-\pi,\pi]$.

Transitional part of speech is characterized as non-stationarity. This part should not be categorized as either voiced or unvoiced frames. Thus, the transition should be synthesized using another approach.

From the listening test, phase information is more important than magnitude information for the transition synthesis[8]. Thus, it would be advantageous to utilize phase information in the transition at the sacrifice of magnitude information to some extent. In fact, utilizing manipulated spectral phase rather than spectral magnitude in the transition makes the coder generate abrupt onset and speech sounds become crisp by reducing hoarse artifacts.

In this paper, the transitional part of the speech can be synthesized by using the linear shift term and the additional delta phase. In order to achieve this goal, firstly the transition detection algorithm proposed in the section II is applied. The proposed algorithm gives the exact position of abrupt onset. At this position, we assume that most of the phase components have nearly zero values so that the deviation of phase information can be small compared with the condition where they are extracted at the other time points. Thus, phase quantization can be performed efficiently using the characteristics of small deviation of phase. When the current frame is declared as transition, linear shift term and relatively important phase components are quantized and transmitted to the decoder. At the decoder side, phase components are reconstructed as

$$\theta_h^{PIPE}(N)=\begin{cases}h\omega_0(N)\hat{l}+\Delta\hat{\theta}_h, & \text{if } h \text{ needs quant. phase}\\ h\omega_0(N)\hat{l}, & \text{elsewhere}\end{cases}, \tag{12}$$

where $\hat{l}$ and $\Delta\hat{\theta}_h$ be the reconstructed start position (equal to linear shift term) and the quantized delta phase, respectively.

For the efficient quantization of the delta phase, the perceptually irrelevant phase elimination algorithm was used where the critical phase frequency is defined and below which no phase term is transmitted[8]. Although phase components below the critical phase frequency are found unnecessary to use in terms of the perceptual quality, it is still required to find out how to select and quantize relatively important phase components above the critical phase frequency. In this paper, ten delta phase components (see (12) for the definition) above the critical phase frequency are selected of which magnitudes are highest, based on the assumption that larger magnitudes would contribute more to the shape of waveform. Then the set of quantization levels, {6,6,6,6,5,5,4,4,4,4}, is assigned to ten delta phase components from the harmonics with the highest magnitudes. For example, the phase components with highest four magnitudes have six quantization levels, the next two components take five levels, and so on. Thus, the total number of bits assigned to quantize ten phase components is 23 bits. For the linear shift term, 8 bits are assigned.

Ⅳ. Performance Evaluation

For the evaluation of the proposed transition detection algorithm, 191940 frames (191.94 sec) of speech data with a sampling rate of 8 kHz, including clean, babble noise 30dB, and vehicular noise 15dB, were selected. The hit ratio shows whether a transition frame is correctly detected, and the false alarm ratio shows whether a stationary frame is mis-classified into a transition frame. The proposed method is superior to the conventional method[7] as presented in Table 1. For clean speech, the proposed and conventional methods have the hit ratio of 92.94% and 64.67%, respectively, and the false alarm ratio of 1.14% and 0.11%, respectively. For the 30dB babble noise environments, the proposed and the conventional methods have the hit ratio of 85.78% and 34.80%, respectively. Clear hit-ratio improvement can be observed in 15dB vehicle noise condition with 71.43% and 0.71%, respectively. The thresholds for the conventional and proposed methods are fixed in (1) and (7), respectively, against the varying noise conditions.

The detected transition parts of speech are synthesized by using the linear shift term and the perceptually important delta phase information at the sacrifice of magnitude information. To verify the speech quality of the proposed algorithm, the SMX vocoder was modified to transmit phase information instead of the parts of magnitude information for the transition synthesis. This transition-considered SMX (T-SMX) vocoder was compared with the conventional SMX vocoder, both at 4 kbit/s, utilizing A-B preference listening test. Speech materials consist of 24 Korean speech sentences gathered from NTT-AT speech database. They were processed by the two coders and played in random order to 12 listeners. The subjective test results are shown in Table 2. The proposed method produces better speech quality than the conventional one in clean speech, tandem and babble noise conditions.

Table 1. Hit ratio and false alarm ratio of the transition detection methods on the various conditions.

Performance Measure	Methods	Clean	Babble (30dB)	Vehicle (15dB)
Hit Ratio (%)	Conventional	64.67	34.80	0.71
	Proposed	92.94	85.78	71.43
False Alarm Ratio (%)	Conventional	1.14	0.52	0.19
	Proposed	0.11	0.14	0.00

Table 2. Preference test results: A and B are processed speech files with SMX and T-SMX, respectively.

Test Conditions	A(%)	B(%)	Same(%)
Clean Speech	25.52	31.25	43.23
Tandem	26.04	39.06	34.90
Babble Noise (30dB)	18.75	25.00	56.25
Vehicular Noise (15dB)	31.25	25.00	43.75

However, in vehicular noise condition, the proposed scheme yields slightly worse speech quality. From the informal listening test, the speech quality is not improved even without quantization of the phase terms. Further works are required to synthesize the speech corrupted by low SNR noise efficiently.

Ⅴ. Conclusions

Transition parts of speech are efficiently detected and synthesized by using the proposed relative peakiness value and the perceptually important phase information, respectively. The proposed detection method is superior to the conventional one. At 4 kbit/s, synthesized speech quality of the transition-considered SMX vocoder shows slightly better than that of the conventional SMX vocoder.

References

〔1〕 L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Upper Saddle River, NJ: Prentice Hall, 1978.

〔2〕 T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practices.* Upper Saddle River, NJ: Prentice Hall, 2002.

〔3〕 DVSI, *APCO project 25: Vocoder Description, Version 1.3.* July, 1993.

〔4〕 Y. D. Cho, M. Y. Kim, and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameters determination," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 601-604, Seattle, WA, USA, 1998.

〔5〕 C. Li and V. Cuperman, "Enhanced Harmonic Coding of Speech with Frequency Domain Transition Modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 581-584, Seattle, WA, USA, 1998.

〔6〕 W. B. Kleijn and J. Haagen, *Speech Coding and Synthesis.* Amsterdam, The Netherlands: Elsevier, 1995.

〔7〕 T. Unno, T. P. barnwell III, and K. Truong, "An Improved Mixed Excitation Linear Prediction (MELP) Coder," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 245-248, Phoenix, Arizona, USA, 1999.

〔8〕 D. S. Kim and M. Y. Kim, "On the perceptual weighting function for phase quantization of speech," in *Proc. IEEE Workshop on Speech Coding*, pp.62-64, Finland, 2000.

김 무 영 (Moo Young Kim) 정회원

1989.2~1993.2: B.Sc., Dept. EE, Yonsei Univ., Korea
1993.3~1995.2: M.Sc., Dept. EE, Yonsei Univ., Korea
1995.3~2000.12: Member of Research Staff, Samsung Advanced Institute of Technology, Korea
2001.1~2004.11: Ph.D., Dept. EE, KTH, Sweden
2004.12~2005.2: PostDoc, Dept. EE, KTH, Sweden
2005.2~2006.8: Senior Research Engineer, Ericsson Research, Sweden
2006.9~Present: Assistant Professor, Dept. Info. Comm. Eng., Sejong Univ., Korea
<Interested Areas> Multimedia Signal Processing and Coding, Information Theory, Pattern Recognition