Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Proposal of a new vocoder for real-time synthesis of speech signal with high quality

Kota NAKANO(1), Masanori MORISE(2) and Takanobu NISHIURA(2)

(1) Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1, Nojihigashi, Kusatsu, Shiga, 525-8577, Japan
(2) College of Information and Science, Ritsumeikan University, 1-1-1, Nojihigashi, Kusatsu, Shiga, 525-8577, Japan

PACS: 43.72.Ja

ABSTRACT

The new framework for high quality vocoder is proposed in this paper. The proposed framework is designed to synthsize high quality speech signal in real time. The conventional vocoder could not synthesize high quality speech signal. In recent, the high computer performance have been achieving that the various vocoders are proposed for high quality speech signal synthesis. In particular, an especially high quality vocoder, named STRAIGHT is proposed. STRAIGHT can analyze and synthesize speech signal with much high quality equivalent to observed speech signal. STRAIGHT requires much amount of computational costs to achieve high quality analysis and synthesis performance. The much computational costs prevent STRAIGHT from real-time speech signal processing. For real-time application, the computational problem is serious in STRAIGHT. The problem should be solved with the new vocoder and new framework proposed in this paper. In the conventional vocoder framework, fundamental frequency and spectral envelope are usually estimated for speech signal parameterization. STRAIGHT estimates not only them but also aperiodic signal from observed signal. The aperiodic signal estimation especially requires almost all the computational costs in STRAIGHT analysis. On the other hand, in the proposed vocoder, it aims only real-time synthesis application, and directly utilizes the signal waveform belonging to the observed speech signal because the aperiodic signal scarcely effects the synthesized signal. In proposed vocoder, the waveform utilization can abbreviate the aperiodic signal estimation, and the abbreviation reduces some pieces of the computational costs. As an abbreviation result, it is expected the new vocoder with the proposed framework can synthesize the similar quality speech signal to STRAIGHT in real time. The subjective and objective evaluation experiments were conducted to verify the effectiveness of the vocoder with the proposed framework. According to the experiments result, the quality of synthesized speech signal is equivalent to STRAIGHT and the framework successes the computational costs reduction.

BACKGROUND

Recently, the studies on music information processing have been evolved rapidly [1,2,3,4]. Especially, the singing voice signal processing has attracted much attention, and various technologies for speech and singing signal processing have been used in the music information processing applications such as Vocaloid [2]. On the one hand, it has grown that the demands to process speech signal or singing signal in real time. A technology, called vocoder [5,6,7,8,9], has been used in various kind of applications to satisfy the demands.

Vocoder is one of the speech signal manipulation technology. It is proposed to achieve the effective speech signal telecommunication. The basic idea is that it represents speech signal by parameters with excitation signal and vocal tract filter. The excitation signal is usually represented with only one numerical value, called fundamental frequency. The vocal tract filter is usually constructed from spectral envelope (or formant envelope). Mostly, a few numerics can represent the spectral envelope. We can recognize the contents of the speech signal synthesized from these estimated parameters. Thereby, the vocoder can code speech signal efficiently with a few parameters, and it has been useful for high efficient telecommunication.

Once, the vocoder was improved to achieve high efficient telecommunication attributed to narrow telecommunication bandwidth. On the other hand, the bandwidth has become broad recently enough to transfer the speech signal as telecommunication without vocoder parameterization. It has made the vocoders aim for not high efficient speech signal telecommunication but high quality speech signal processing.

CONVENTIONAL VOCODERS

Summary

The basic idea of vocoder is speech signal parameterization. Fundamental frequency and spectral envelope are usually used as the parameters in vocoder. Various vocoders have been proposed for high efficient speech signal telecommunication. It has already achieved the high performance in fundamental frequency estimation. On the other hand, the spectral envelope estimation has not achieved the high quality, yet. Various kind of spectral envelope estimation has been proposed, and their differences are known as the vocoders differences.

Channel vocoder

Channel vocoder [7] is a basic vocoder idea. Channel vocoder estimates spectral envelope from observed speech signal. In analysis section of channel vocoder, it estimates spectral envelope with band-pass-filter bank called envelope follower or envelope detector [7]. In synthesis section, vocoder generates excitation signal and filters the excitation signal with filter bank which is generated from the estimated spectral envelope. As a result, channel vocoder can represent spectral envelope from a few parameters. The efficient representation is useful for high efficient speech signal telecommunication.

However, channel vocoder has some problems. The spectral envelope estimation with the filter bank is not robust against fundamental frequency because the fundamental frequency influences the spectrum. Furthermore, the optimum number of the filter bank depends on the observed signal.

LPC vocoder

LPC stands for Linear Predictive Coding. LPC vocoder [8] has the same structure as the channel vocoder.

LPC vocoder models spectral envelope with all-pole filter. The all-pole filter can efficiently construct spectral envelope and power peak of that. The vocoder can synthesize speech signal with higher quality than channel vocoder because the spectral peak by all-pole filter is effective to the contents of the speech signal.

LPC vocoder estimates the pole position by solving auto regression modelled formula. On the other hand, LPC vocoder assumes white characteristics to the excitation signal in the observed speech signal. However, the excitation signal cannot be always white characteristic due to the fundamental frequency. The influence causes estimation errors. In addition, there is also the problem in LPC. The optimum number of poles depends on spectral envelope although LPC vocoder preliminarily requires the number of poles design. The dependency prevents LPC vocoder from high quality speech synthesis.

Cepstral vocoder

Cepstral vocoder [9] has also the same structure as channel vocoder. The Cepstral vocoder estimates spectral envelope from the observed signal by using frequency analysis based on Fourier Transform.

In vocoder analysis, speech signal is represented by convoluting the excitation signal and vocal tract filter. Convoluted signal on time domain is projected as multiplied signal on frequency domain by using Fourier Transform. In addition, the multiply of signals can be represented as adding formula with logarithm function. That is, we can analyze the convoluted speech signal as the added signal which consists of excitation signal and vocal tract filter response, and we may separate them.

On the other hand, there is a useful phenomenon. The logarithmic spectral envelope consists of gradual fluctuation signal on frequency domain. In contrast, the fluctuation of excitation signal is high on frequency domain. Utilizing the phenomenon, with projecting the signal to quefrency domain [9], the power from spectral envelope is distributed around low quefrency and the power from excitation signal is distributed around high quefrency. The separation is achieved because each energy distribution tendencies are different. Moreover, the spectral envelope extraction becomes possible with rejecting the energy from the excitation signal. The rejection process is called “liftering”.

Optimizing liftering length is difficult because the each energy distributions are usually overlapped.

STRAIGHT

The high quality vocoder, named STRAIGHT [11,12] was proposed in 1997. STRAIGHT was designed based on human auditory. On the conventional vocoder with Fourier Transform, the obtained spectrum has the problem that it has fluctuant on both time and frequency domains due to the temporal estimation position and fundamental frequency of observed signal. In STRAIGHT analysis, the method which is robust against the fluctuation was proposed. STRAIGHT uses compensatory time window to remove the fluctuation on time-domain and smoothes the fluctuation on frequency-domain. According to the methods, the STRAIGHT achieves the estimation robust against the fluctuations. In STRAIGHT synthesis, minimum phase is used as phase in spectral envelope.

In addition, STRAIGHT analyzes the aperiodic elements in observed signal for synthesis aperiodic excitation speech signal in voiced speech signal and unvoiced speech signal. The aperiodic elements estimation requires far huge computational costs to STRAIGHT. The requirement prevents STRAIGHT from real-time speeh processing.

The preliminary examination is conducted to demonstrate the bottleneck in STRAIGHT analysis. The examination result is displayed on Tab. 1. The table displays the elapsed time ratio of the each estimation to the whole elapsed time.

Table 1. The elapsed time ratio of each estimation a spoken signal in Japanese

Analyzed parameter	Elapsed time ratio
Fundamental frequency	21.7%
Aperiodic ratio	72.8%
Spectral envelope	5.5%
Total	100%

According to the table, the bottleneck in the analysis is aperiodic ratio estimation that occupies 72.8% of whole elapsed time. According to the result, it is expected that the improving aperiodic ratio estimation reduces the computational costs efficiently.

THE NEW FRAMEWORK

Summary

The new vocoder is proposed based on STRAIGHT. STRAIGHT estimates the parameter related to the aperiodic signal from excitation signal which belongs to observed signal. In the proposed vocoder, it abbreviates the estimation, and reduces the computational costs. The processing flowchart is displayed in Fig. 1.

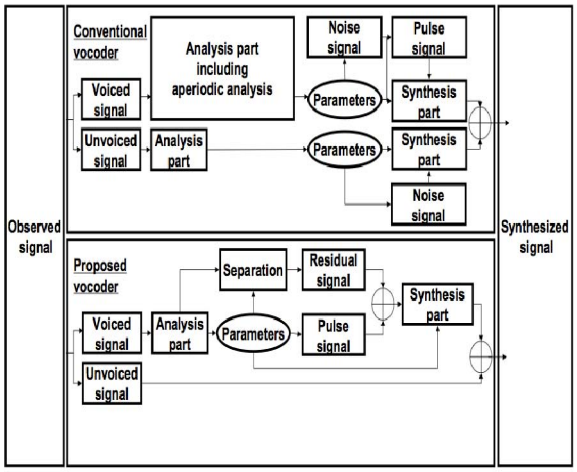

Figure 1. Flowcharts of each vocoder

The figure illustrates both of vocoder flowcharts. The each vocoder processing is shown in the center section. STRAIGHT framework is shown in the upper section, and the proposed one is shown in lower section.

First of all, the both vocoders indicate whether the observed speech signal is voiced signal or unvoiced signal.

In analysis section, the conventional vocoder including STRAIGHT estimates fundamental frequency and spectral envelope. STRAIGHT additionally estimates aperiodic ratio in each frequency band if the observed signal is indicated as a voiced speech signal. The conventional vocoder including STRAIGHT estimates spectral envelope, provided that the observed signal is indicated as an unvoiced signal.

In voiced speech signal synthesis part, STRAIGHT generates an excitation signal from fundamental frequency and aperiodic ratio parameters, generates filter with minimum phase response from estimated spectral envelope parameters and filters the generated excitation signal. In unvoiced speech signal part, the conventional vocoder generates an aperiodic signal as the excitation signal, generates filter with minimum phase response from spectral envelope parameters, and filters the generated excitation signal with the generated filter. By contrast, in the unvoiced speech signal synthesis, the parameters estimation doesn't have effective effects to vocoder quality because the estimation effects to aperiodic signal and that differences is unclear for that quality. In other words, with restricting only speech signal synthesis or processing, the estimations should be redundancy and it expected the computational costs may be reduced.

The new framework is proposed. The flowchart is displayed on lower section in Fig. 1. The proposed vocoder reuses observed unvoiced signal into synthesized signal waveform directly to abbreviate the redundancy analysis and synthesis. The proposed vocoder estimates fundamental frequency and spectral envelope from observed speech signal as well as STRAIGHT if the speech signal is voiced. In addition, the vocoder extracts "residual signal" instead of the aperiodic signal. The "residual signal" is defined as the signal with subtracting parametric elements such as fundamental frequency effect and spectral envelope effect from the observed natural speech signal. Originally, the natural speech signal is represented with vocal tract filter and excitation signal. The excitation signal is generated from periodic signal and aperiodic signal. Then, it is possible that it directly extracts aperiodic signal from observed signal by subtracting parametric elements.

Residual signal extraction

STRAIGHT represents the synthesized speech signal which consists of the vocal tract filter with minimum phase response and excitation signal, and STRAIGHT defines that the natural excitation signal consists of periodic signal and aperiodic signal because the voiced speech signal has aperiodic elements [10]. Due to this definition, the proposed method extracts the excitation signal by subtracting the effect from spectral envelope in observed speech signal. An example is displayed in Fig. 2. It shows waveform of observed speech signal and waveform of extracted residual signal. The signal content is a part of natural vowel /a/ in Japanese.

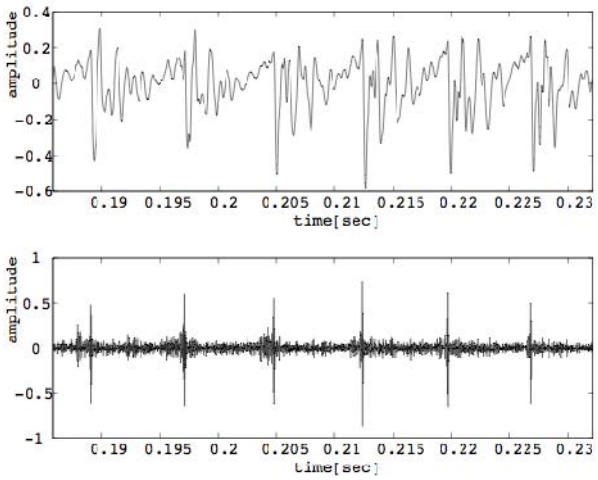

Figure 2. An example of observed signal and residual signal waveform

The figure illustrates waveform of the observed signal and residual signal waveform. The residual signal consists of periodic pulse train signal and aperiodic signal. We can confirm the signal power peak synchronization between the observed signal and the obtained residual signal. It achieves to extract the aperiodic signal included in the excitation signal with suppressing the fundamental frequency influence. The signal originated from the spectral envelope has periodic structure. Then, the residual signal means the independent signal from periodic elements such as fundamental frequency and spectral envelope in observed natural speech signal.

EVALUATION EXPERIMENT

Summary

The experiments were conducted by subjective and objective evaluation. It compares which vocoder is most effective for real-time and high quality speech signal processing, the conventional vocoders or the proposed vocoder. The objective experiment measured the elapsed time to evaluate the computational costs. The subjective experiment evaluates the quality of the synthesized speech signal with MOS evaluation.

Experiment condition

ATR phoneme balanced corpus was used as source signal in the both subjective and objective experiments. The sampling rate and bit depth for the quantization was 44.1 [kHz] and 32 [bit] in the experiments. The analysis and synthesis process frame was designed with 2048 [sample]; (about 46 [msec] in this case). The channel vocoder was used with 16 band-pass-

filter banks, the LPC vocoder was used with 16 poles and the Cepstral vocoder was used with 32 liftering lengths in the experiments.

The experiment condition and the design configurations are illustrated on Tab. 2 and Tab. 3.

Table 2. Experimental conditions

Sampling rate	44.1[kHz]
Bit depth for quantization	32[Bits]
Frame width	2048[sample]
Background noise level	25[dBA]
Source signal	ATR corpus 216 speech

Table 3. Vocoder configurations

Designed parameter	Value
Filter amounts in Channel vocoder	16
Pole amounts in LPC vocoder	16
Liftering length in Cepstral vocoder	32[sample]

On the objective evaluation, the all vocoders analyzed and synthesized the every speech signal which belongs to the ATR corpus. The elapsed time was measured during their processing and it calculated the elapsed time ratio for the time length of corresponded used speech signal. After all measurements, average was calculated from the scores of each sample. The measurements and calculations were conducted by all vocoders.

On the subjective evaluation, the all vocoders analyzed and synthesized the every speech signal which belongs to the ATR corpus, and the synthesized signal was presented to subjects with shuffled sequence at soundproof room with 25 [dBA]. The subjects evaluated the presented signal from 1 score to 5 score.

The 5 evaluation score is excellent. The 1 evaluation score is bad. After all evaluations, the score average and standard deviation were calculated and we used the score as evaluation result. The subjects group consists of 2 female and 4 male students.

Result

The result on the objective evaluation experiment is displayed on Tab. 4.

The left column means used vocoder algorithm and the right column means the calculated ratio of elapsed time to signal time length. The results show the STRAIGHT is the worst of all and the proposed vocoder can process the speech signal in real time.

Table 4. Results of objective evaluation of vocoders

Vocoder algorithm	Average of elapsed time per sample time length
Channel vocoder	73%
LPC vocoder	18%
Cepstral vocoder	11%
STRAIGHT	854%
Proposed vocoder	28%

On the subjective evaluation experiment, the result is displayed in Fig. 3. In the figure, the vertical axis means subjective quality. The range is from 0 to 5. The bottom labels mean the vocoder algorithm corresponded to the each bar. The score height of each bar means the average score. The line means the standard deviation.

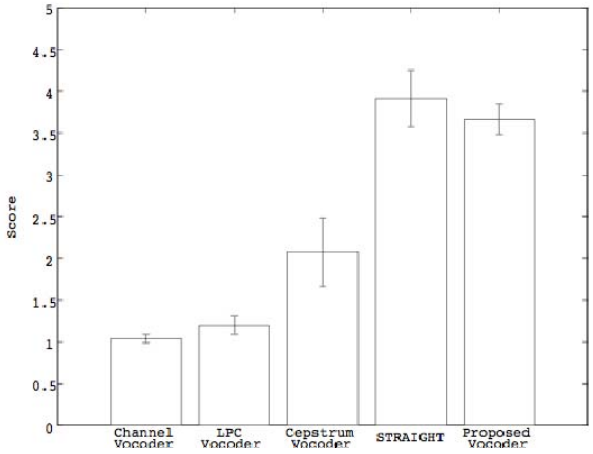

Figure 3. The quality score of speech signal synthesized by each vocoder

The result shows STRAIGHT has the highest quality of all vocoders including proposed one. However, the proposed vocoder has the second highest quality of all vocoders and the differential range is only 0.2 score between STRAIGHT and proposed vocoder and the differential range is more 1.2 score than Cepstral vocoder that was evaluated as the third score rank.

DISCUSSION

The proposed vocoder is inferior to STRAIGHT in synthesis quality. The proposed vocoder mixes a signal synthesized from the estimated parameters and the observed signal. It suggested that the estimation performance was not high enough to synthesize high quality speech signal equivalent to STRAIGHT. On the other hand, as results of the objective evaluation, it was displayed the proposed vocoder can perform speech signal processing in less time than the time length of the observed speech signal. On the other word, the vocoder with the proposed framework can process speech signal in real time and the quality of synthesized speech signal has much higher quality than the quality of the conventional vocoders excluding STRAIGHT. The differential score range between the proposed vocoder and STRAIGHT is quite less than the differential range between the proposed vocoder and the conventional ones.

Hence, the proposed vocoder can perform real-time speech signal processing with high quality equivalent to STRAIGHT.

CONCLUSIONS

The studies on music information, especially the speech signal processing, attract attention recently. One of the vocoder, named STRAIGHT for high quality analysis and synthesis is proposed to process speech signal instead of much computational costs. The new framework of vocoder is proposed in this paper for real-time speech signal processing. With the framework, the proposed vocoder synthesizes the speech signal with reusing aperiodic signal waveform included in the observed signal and adding extracted aperiodic signal included in the observed signal. The proposed framework reduces computational costs in the STRAINGHT. The subjective and objective evaluations were conducted. The experimental results showed the proposed vocoder with the framework is superior to conventional one in computational costs and the proposed vocoder has high quality mostly equivalent to STRAIGHT.

ACKNOWLEDGMENTS

This research was partly supported by the CrestMuse project by JST and Grants-in-Aid for Scientific Research by JSPS;

REFERENCES

1. M. Goto, T Saito, T Nakano, H Fujihara,“Recent studies on singing information processing ”, AST. pp. 616-623, 2008.
2. H Katayose, M Goto, “Toward Development of Design Reuse Technology for music”, JSAI. ID: 1D1-4, 2006.
3. H. Kawahara.” Extending STRAIGHT-based Speech Morphing for Case-Based Design Assistance”, JSAI. ID: 1D1-5, 2006.
4. H Kawahara, T Ikoma, M Morise, T Takahashi, K Toyoda, H Katayose, “ Perceptual study on design reuse of voice identity and singing style based on singing voice morphing”, IPSJ interaction, 2008.
5. J. L. Flanagan, R. M. Golden,“Phase Vocoder”, Bell System Technical, pp.1493- 1509, 1966
6. A Robel. “A new approach to transient processing in the phase vocoder”, DAFx- 3 2003.
7. H.Dudley, “Remaking speech”, Acoust. Soc. Am. pp. 169-177, 1939.
8. Mathews M. V., Miller J. E., David E. E., Jr. “Pitch synchronous analysis of voiced sounds.” Acoust. Soc. Am, pp. 179-186, 1961.
9. V Oppenheim, “A speech analysis-synthesis system based on homomorphic filtering”, Acoust. Soc. Am, pp. 458-465 1969.
10. I R Titze, “Principles of voice production”, Acoust. Soc. Am. pp. 1148-1148, 1994.
11. H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, B. Banno,“Tem- porally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown ”, Proc. ICASSP pp. 3905-3908, 2009.
12. H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, “ A temporally stable power spectral representation for periodic signals and appli- cations to interference-free spectrum, f0, and aperiodicity estimation” , Proc. ICASSP pp. 3933-3936, 2008.