EFFECT OF ANTI-ALIASING FILTERING ON THE QUALITY OF SPEECH FROM AN HMM-BASED SYNTHESIZER

Yoshinori Shiga

Spoken Language Communication Laboratory, MASTAR Project
Universal Communication Research Institute
National Institute of Information and Communications Technology (NICT), Japan
yoshi.shiga@nict.go.jp

ABSTRACT

This paper investigates how the quality of speech produced through statistical parametric synthesis is affected by anti-aliasing filtering, i.e., low-pass filtering that is applied prior to (down-) sampling prerecorded speech at a desired rate. It has empirically been known that the frequency response of such anti-aliasing filters influences the quality of speech synthesized to a considerable degree. For the purpose of understanding such influence more clearly, in this paper we examine the spectral aspects of speech involved in the processes of HMM training and synthesis. We then propose a technique of feature extraction that can avoid producing the roll-off feature of the frequency response near the Nyquist frequency, which is found to be the major cause of speech quality degradation resulting from anti-aliasing filtering. In the technique, the spectrum is first computed from speech at a sampling rate higher than the desired rate, then it is truncated so that its frequency range above the target Nyquist frequency is discarded, and finally the truncated spectrum is converted directly into the cepstrum. Listening test results show that the proposed technique enables training HMMs efficiently with a limited number of model parameters and effectively with less artifacts in the speech synthesized at a desired sampling rate.

Index Terms— down-sampling, anti-aliasing, speech quality, HMM-based speech synthesis, statistical parametric speech synthesis, sampling frequency

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] [2] is currently receiving a considerable amount of attention in the field of speech synthesis technology. The synthesis method is not only capable of acquiring the characteristics of speech automatically from a speech database, but is also capable of producing speech with various voice characteristics and speaking styles through the potential use of speech morphing and speaker adaptation. Such flexibility provides a great advantage over the other leading speech synthesis techniques. However, unnatural speech produced through the source-filter model, or vocoder, still presents a challenge. The attempt to solve this problem has become a research topic of increasing interest.

At the Blizzard Challenge 2010 workshop (BC2010)[1] [3], there was a report which claimed that speech quality and speaker similarity can be enhanced by training HMMs with the spectral features of speech at a higher sampling rate (e.g., 48 kHz), then synthesizing speech using the trained models, and finally down-sampling the speech to the rate (e.g., 16 kHz) desired for the evaluation [4] (the technique is hereafter referred to as *over-sampling technique*). Despite the fact that the quality of speech actually improves thereby, no solid basis could be found to explain why this procedure is effective. One possible explanation for such speech quality enhancement is the improved granularity of pitch-period control during the waveform generation. However, there seems to be something more to producing such quality improvement perceivable between speech from training with data at a lower sampling rate and the down-sampled version of speech from training with data at a higher rate.

It is empirically known that the quality of the eventual synthetic speech output is influenced by the frequency characteristic of an anti-aliasing filter, a low-pass filter that is applied prior to (down-) sampling prerecorded speech. The above-mentioned speech quality improvement by the over-sampling technique can be associated with this empirical knowledge, if we assume that trench-shaped roll-off around the Nyquist frequency caused by the filtering affects the feature extraction and/or the model training of the HMM-based synthesis framework. The technique deals with signals at a higher sampling rate, which pushes such a spectral trench much further out of the voice band.

The objective of this study is to clarify how anti-aliasing filtering actually influences the quality of the speech synthesized. In addition, this paper aims to propose a solution that strikes a suitable balance between the quality of speech and the cost of computation for training the models.

The remainder of this paper is organized as follows: in Section 2, we examine, based on observation, how the anti-aliasing filter influences the parametric synthesis and discuss why the filtering causes quality degradation in synthetic speech; in Section 3, a method is proposed for coping with the degradation problem; in Section 4, a listening test is performed in order to investigate perceptible degradation caused by anti-aliasing filtering; and we present our conclusions in Section 5.

2. EFFECT OF ANTI-ALIASING FILTERING

2.1. Its effect on feature extraction

We examine, first of all, how anti-aliasing low-pass filtering influences feature extraction by observing the spectral features extracted by the method that is actually used to create data for training HMMs. The speech to be analyzed here are signals at a 16-kHz sampling rate

The author would like to thank Prof. Keiichi Tokuda for his helpful discussions on issues dealt with in this paper.

[1]The Blizzard Challenge is an annual event for the evaluation of corpus-based speech synthesis systems.

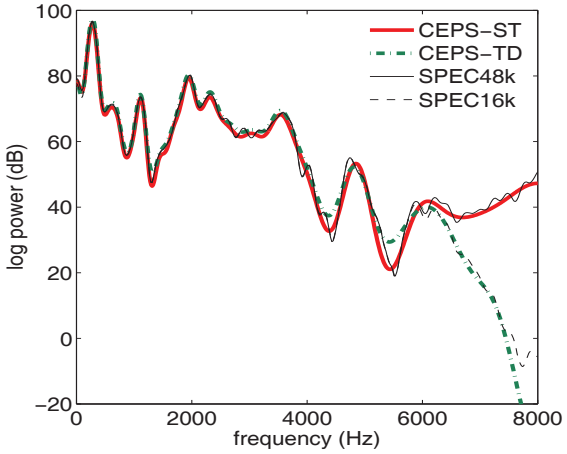

Fig. 1. Comparison of spectra computed using different feature extractions. SPEC48k: STRAIGHT spectrum of a speech signal at a 48 kHz sampling rate; SPEC16k: STRAIGHT spectrum of the corresponding BC2010 16-kHz speech; CEPS-ST: reconstruction from the mel-cepstrum that was computed using the spectrum truncation technique (see Section 3); CEPS-TD: reconstruction from the mel-cepstrum that was computed from SPEC16k.

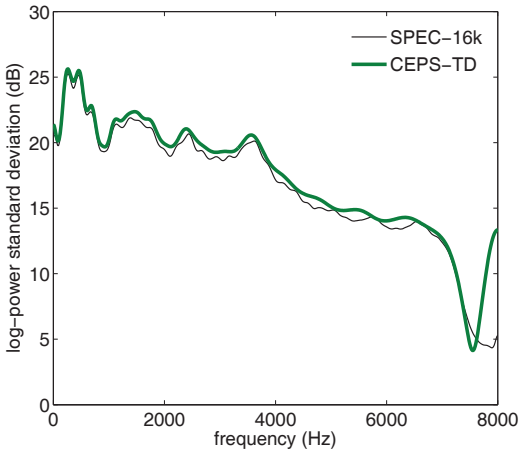

Fig. 2. Standard deviation of log power at each frequency bin for the STRAIGHT spectra of the BC2010 16-kHz speech across 100 utterances (SPEC-16k), and for the spectra reconstructed from the mel-cepstra of the same speech (CEPS-TD).

from an ARCTIC set of the British English Roger corpus, which was released by the Centre for Speech Technology Research of the University of Edinburgh for the BC2010 evaluation (hereafter called 'BC2010 16-kHz speech'). The mel-cepstrum was extracted from the STRAIGHT spectrum [5] of the signals, using the *mgcep* command of the Speech Signal Processing Toolkit (SPTK) [6]. The cepstral order and frequency warping factor α were set to 39 and 0.42, respectively.

Figure 1 compares spectra obtained with different feature extractions. The thick dashed line indicates the spectrum reconstructed from the mel-cepstrum of a voiced section from the BC2010 16-kHz speech, while the thin solid line shows the STRAIGHT spectrum of the corresponding speech sampled at a 48-kHz rate (48-kHz speech).[2] Evidently, due to the low-pass filtering, as a matter of course, the BC2010 16-kHz speech is poor in spectral energy around the Nyquist frequency. It should also be noticed that the spectral undulations between 4 and 6 kHz tend to be flattened compared to the spectrum of 48-kHz speech. This is probably due to the influence of the cepstral analysis trying to approximate the sharp roll-off produced at the cut-off band of the low-pass filter.

Shown in Fig. 2 is the standard deviation of the log-power spectrum at each of the frequency bins for the BC2010 16-kHz speech across 100 utterances. As can be seen from this graph, there is a considerable difference around the Nyquist frequency between the original STRAIGHT spectrum and reconstruction from the mel-cepstrum. This difference indicates that the spectrum around the Nyquist frequency is quite unstably represented by the mel-cepstrum. Such spectral variation in training data potentially has an adverse effect on the model statistics.

These problems might be resolved to a certain extent if a low-pass filter is applied with a sharper roll-off. However, the shrinkage of the high frequency range in perceptually-motivated scales can make it harder for the cepstrum to represent such a sharp roll-off near the Nyquist frequency.

[2]In BC2010, speech data of 48-kHz sampling rate were also released.

2.2. Its effect on speech synthesis

The lack of high-frequency energy by the filtering causes speech to sound band limited. On top of that, synthetic speech tends to be muffled when training data has such energy deficiency. This is presumably because the models were trained so as to well represent the fluctuation around the Nyquist frequency and to deal less prominently with spectral sections essential for the perception of speech.

For the purpose of examining the effect on synthetic speech, we introduce a technique called 'spectral truncation', with which the mel-cepstrum is obtained so that its reconstruction has no roll-off feature in the frequency domain (see Section 3 for details). The thick solid line in Fig. 1 represents a spectrum of this representation.

Figure 3 compares the spectra that were generated by HMMs trained with different mel-cepstral parameters: the mel-cepstrum computed through the spectral truncation from the 48-kHz speech of the corpus, and the mel-cepstrum computed from the BC2010 16-kHz speech of the corpus. Both sets of models were trained under the exact same conditions (as in Section 4.1) except for the above difference in feature extraction.

It is evident from this figure that there are differences not only in the spectral energy of the high frequency range, but also in the sharpness of formants and anti-formants. The muffled sounds are probably perceived due to such flattened spectral undulation. Further examination is needed to clarify what causes such a 'formant de-emphasis.'

3. SPECTRAL TRUNCATION

In order to avoid both the aliasing and the spectral trench near the Nyquist frequency, we adopt a spectral truncation approach, which was originally invented for concatenative synthesis to directly alter the sampling rate of synthesis units consisting of the cepstrum without converting them into time-domain signals [7].

In the process of this technique, the spectral envelope is first extracted from speech of a higher sampling rate (e.g., using STRAIGHT) and then data points of the spectral section beyond the Nyquist frequency of the desired sampling rate are discarded,

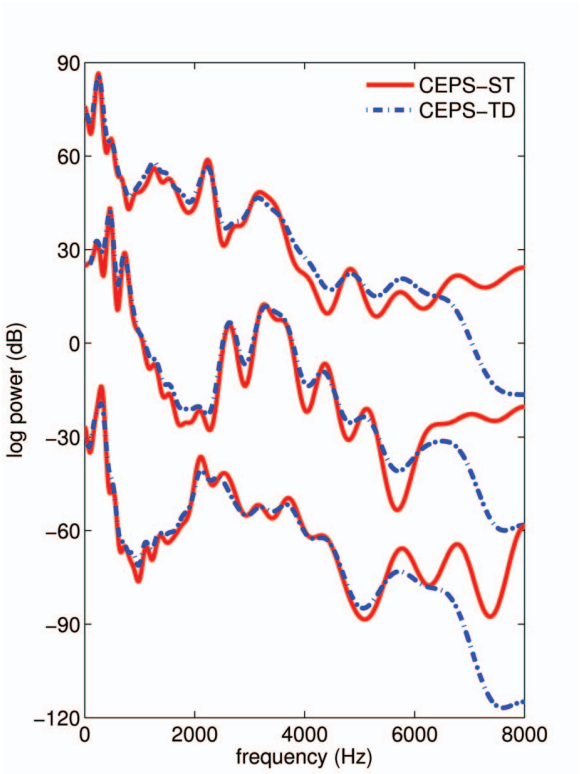

Fig. 3. Comparison of the spectra that were produced by HMMs trained with different mel-cepstral parameters: the mel-cepstrum computed with the spectrum truncation (CEPS-ST), and the mel-cepstrum computed from the BC2010 16-kHz speech (CEPS-TD).

as schematically shown in Fig. 4. The mel-cepstrum is computed from this truncated spectrum. This technique can avoid the spectral roll-off at the Nyquist frequency; no aliasing occurs as well. This may be interpreted as anti-aliasing with an ideal filter.

4. PERCEPTIBLE DEGRADATION

An opinion test was conducted on the naturalness of the synthetic speech in order to investigate perceptible degradation caused by anti-aliasing filtering, as well as the effectiveness of the proposed technique.

4.1. Conditions, materials, and procedure

As subjects, five speech technology experts took part in the listening test. Each evaluated the speech of ten sentences produced by each of the speech synthesis systems, whose specifications are listed in table 1. The order of mel-cepstrum and the frequency-warping factor for each system were adjusted appropriately according to our preliminary experiments. The sentences were chosen randomly from the 'news' and 'novel' categories of the BC2010 test sentences. The scale for the opinion test was from 1 ('completely unnatural') to 5 ('completely natural'). The test was carried out in a quiet room and the listeners used headphones.

The speech data used for training was the ARCTIC set of Roger corpus already mentioned in Section 2. The corpus includes 1,132 utterances (approx. 1.5 hours long) by a British male speaker. F_0s and spectral envelopes were estimated using the Snack Sound

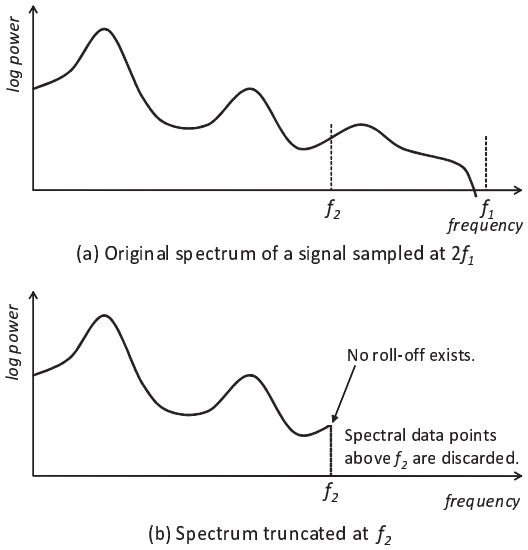

Fig. 4. Spectral truncation: spectral data points of the frequency range over the Nyquist frequency of a target sampling rate is truncated so as to form the spectrum of speech at the target rate.

Toolkit and the STRAIGHT analysis [5], respectively, with 5-ms frame shifts. Spectral envelopes were converted into the mel-cepstrum using SPTK. The aperiodic component (AP parameter) of the STRAIGHT spectrum was also parameterized as the cosine transform of the frequency-warped spectrum (mel-scale spectrum). Five-state left-to-right no-skip hidden semi-Markov models were used for the training and synthesis of duration, F_0, and mel-cepstral coefficients within the framework of the HMM-based Speech Synthesis System version 2.2β [8].

The procedures of HMM training and speech synthesis are shown in Fig. 5. In order to see the effect of the over-sampling technique, which was mentioned in Section 1, the spectrum with the Nyquist frequency of 12 kHz (corresponding to the spectrum of speech at a 24-kHz sampling rate) was used to train the models of System 3. We avoided applying the released 48-kHz speech itself because representing such speech signals requires the cepstrum of a considerably high order (> 60), the estimation of which involves other factors with the quality of speech output. The 24-kHz speech synthesized through System 3 was down-sampled in the end using the MATLAB function *resample*, which adopts anti-aliasing filtering with a good low-pass property so that the loss of high frequency energy can be minimized.

For reference, we added synthetic samples produced through an analysis-synthesis system (System 4), for which BC2010's test utterances at a sampling rate of 16 kHz was used. The spectral truncation was not applied in the analysis part of the system since we found that they had been processed with an anti-aliasing filter with a sufficiently high cut-off frequency, differently from the filter that had been used for down-sampling the training utterances.

4.2. Results and discussion

Listening test results are shown in Fig. 6. System 1, trained with the BC2010 16-kHz data, had the worst score at 2.5. Systems 2 and 3 have the same score, 2.9, which means there is no significant difference in quality even when the wide-band spectrum is employed.

Table 1. Specifications of systems used for the subjective evaluation

System #	Synthesis type	Cepstral order	Freq. warping factor
1	HMM-based synthesis	39	0.42
2	HMM-based synthesis	49	0.50
3	HMM-based synthesis	39	0.42
4	Analysis/synthesis	39	0.42

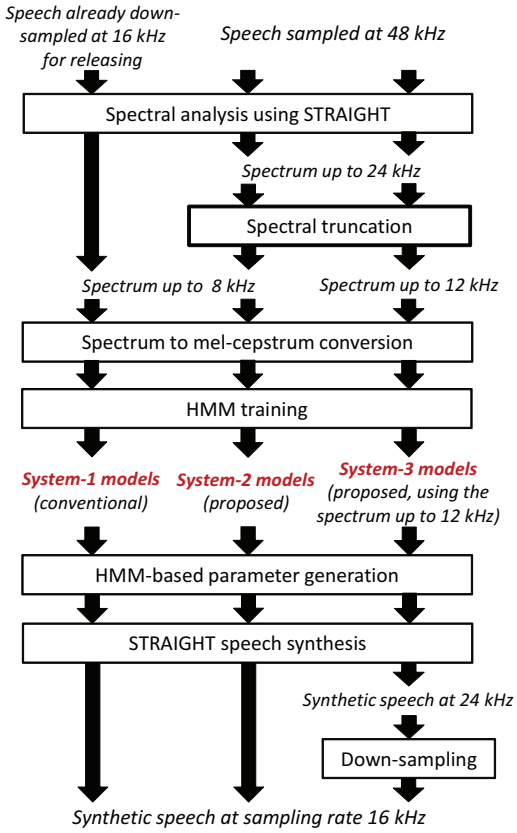

Fig. 5. Experimental procedure.

These results suggest that the proposed spectral truncation can effectively avoid the adverse influence caused by anti-aliasing. In addition, the over-sampling technique is not always necessary for achieving better quality 16-kHz synthetic speech, since no significant difference was seen in MOS between the systems trained with spectra corresponding to those of 16-kHz and 24-kHz speech. It should be noticed that representing signals of a higher sampling rate requires parameters of a higher order, which accordingly involves much higher computational cost during the training and runtime.

5. CONCLUSIONS

We have investigated how the quality of speech produced through HMM-based synthesis is affected by anti-aliasing filtering.

We, researchers and developers, often have to use speech data that have been already down-sampled to a certain rate. In such an evaluation event as the Blizzard Challenge, for example, the major part of the released data sets is down-sampled in advance by the organizing institution. Also, speech technology companies may use speech data that has been recorded and already down-sampled by their clients. However, all the findings here suggest that we should examine carefully what sort of low-pass filter has been used for anti-aliasing. In the case prerecorded speech at a high sampling rate (e.g., 48 kHz or 44.1 kHz) is available, it is advisable to use the speech with feature extraction that is specifically designed to create data for the training of parametric synthesis, such as the spectral truncation.

The spectral truncation introduced here is a simple, but effective solution; it is capable of synthesizing better quality of speech without increasing the dimension of the cepstral vector, while avoiding the ill effect of anti-aliasing filtering. We are aware that the SPTK command *ds* is designed so as to cause a small amount of aliasing and avoid producing roll-off at the Nyquist frequency, but not for all of the specifiable down-sampling factors. Investigating how such spectral roll-off interferes with the model statistics and consequently flattens the synthesized speech spectrum will be our future work.

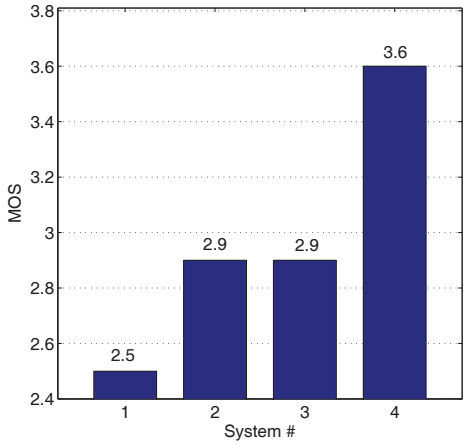

Fig. 6. Subjective evaluation result shown by the mean opinion score (MOS). The specifications of each system are listed in table 1.

6. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH'99*, Budapest, Hungary, Sep. 1999, pp. 2347–2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000.

[3] `http://www.synsig.org/index.php/Blizzard_Challenge_2010`.

[4] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard Challenge 2010," in *Proc. Blizzard Challenge 2010*, Kyoto, Japan, Sep. 2010.

[5] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP97*, vol. 2, Munich, Germany, Apr. 1997, pp. 1303–1306.

[6] `http://sp-tk.sourceforge.net/`.

[7] Y. Shiga, Japanese Patent No. 3302075 (applied in Feb., 1993).

[8] `http://hts.sp.nitech.ac.jp/`.