Lost Speech Reconstruction Method using Speech Recognition based on Missing Feature Theory and HMM-based Speech Synthesis

Shingo Kuroiwa, Satoru Tsuge, Fuji Ren

The University of Tokushima
2-1, Minami-Josanjima, Tokushima 770-8506, Japan
kuroiwa@is.tokushima-u.ac.jp

Abstract

In recent years, IP telephone service has spread rapidly. However, an unavoidable problem of IP telephone service is deterioration of speech due to packet loss, which often occurs on wireless networks. To overcome this problem, we propose a novel lost speech reconstruction method using speech recognition based on Missing Feature Theory and HMM-based speech synthesis. The proposed method uses linguistic information and can deal with the lack of syllable units which conventional methods are unable to handle. We conducted subjective and objective evaluation experiments under speaker independent conditions. These results showed the effectiveness of the proposed method. Although there is a processing delay in the proposed method, we believe that this method will open up new applications for speech recognition and speech synthesis technology.

Index Terms: speech recognition, speech synthesis, missing feature, speech reconstruction

1. Introduction

In recent years, IP telephone service has spread rapidly thanks to the development of VoIP (Voice over IP) technology. Furthermore, some network companies have deployed IP wireless telephones in commercial environments. However, an unavoidable problem of the IP telephone service is deterioration of speech due to packet loss, which often occurs on wireless networks. To conceal packet loss, modern codecs (e.g. G.711, G.723.1, G.729) derive the parameters for the lost frames from the parameters of previous frames. However, these methods cannot deal with losses more than 60 ms in length (in G.711[1]), and cannot fundamentally deal with the loss of more than two phonemes. To overcome this problem, we propose a novel lost speech reconstruction method using speech recognition and synthesis. First, the proposed method predicts the contents of the lost packet section using speech recognition based on Missing Feature Theory (MFT)[2][3]. The MFT-based speech recognizer predicts the phoneme-state sequence of the lost packet section using acoustic information and linguistic information before and after the lost frames. Next, for reconstructing a speech signal to replace the lost section, an HMM-Based speech synthesizer generates a synthetic speech signal according to the predicted phoneme-state sequence. Although we have already presented the basic idea and experimental results with speaker-dependent conditions in ICA2004[4] and NLP-KE2005[5], we show the new experimental results under speaker-independent condition and some improvement by using a speaker adaptation technique and a waveform concatenation technique that overcomes discontinuity problems, which was the most serious problem in the previous experiments, in this paper.

2. Lost Speech Reconstruction Method

2.1. Outline of the proposed method

A block diagram of the proposed method is shown in Fig. 1. First, an MFT-based speech recognizer with a speaker-independent HMM predicts the phoneme-state sequence of the lost section. Next, an HMM-based speech synthesizer[6] with a speaker-adapted HMM generates a mel-cepstral sequence according to the predicted phoneme-state sequence. The F0 data for voiced sounds in the lost section are estimated from the F0 data before and after the section by linear interpolation. Next, using this information (mel-cepstral sequence and F0 data) on the lost section, the MLSA (Mel Log Spectral Approximation) filter[7] synthesizes the speech of the lost section. Finally, the recovered speech is generated by concatenating the synthetic speech into the lost section with a maximum cross-correlation coefficient condition and an overlap-add method.

2.2. Algorithm

We detail the lost speech reconstruction algorithm.

Let $\boldsymbol{X}$ be a framed speech sequence,

$$\boldsymbol{X} = \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(i), \ldots, \boldsymbol{x}(i+n), \ldots, \boldsymbol{x}(N)\}, \quad (1)$$

where, $\boldsymbol{x}(i)$ is the framed speech at time i and N is the number of frames. Here, we assume that speech from i-th to $(i+n)$-th frame is lost.

Now, we predict the lost frames,

$$\tilde{\boldsymbol{X}}_m = \{\tilde{\boldsymbol{x}}(i), \ldots, \tilde{\boldsymbol{x}}(i+n)\}, \quad (2)$$

by maximizing the following conditional probability:

$$P(\tilde{\boldsymbol{X}}_m | \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(i-1)\}, \{\boldsymbol{x}(i+n+1), \ldots, \boldsymbol{x}(N)\}). \quad (3)$$

The speech signal is recovered with $\tilde{\boldsymbol{X}}_m$ as

$$\{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(i-1), \tilde{\boldsymbol{x}}(i), \ldots, \tilde{\boldsymbol{x}}(i+n), \boldsymbol{x}(i+n+1), \ldots, \boldsymbol{x}(N)\}. \quad (4)$$

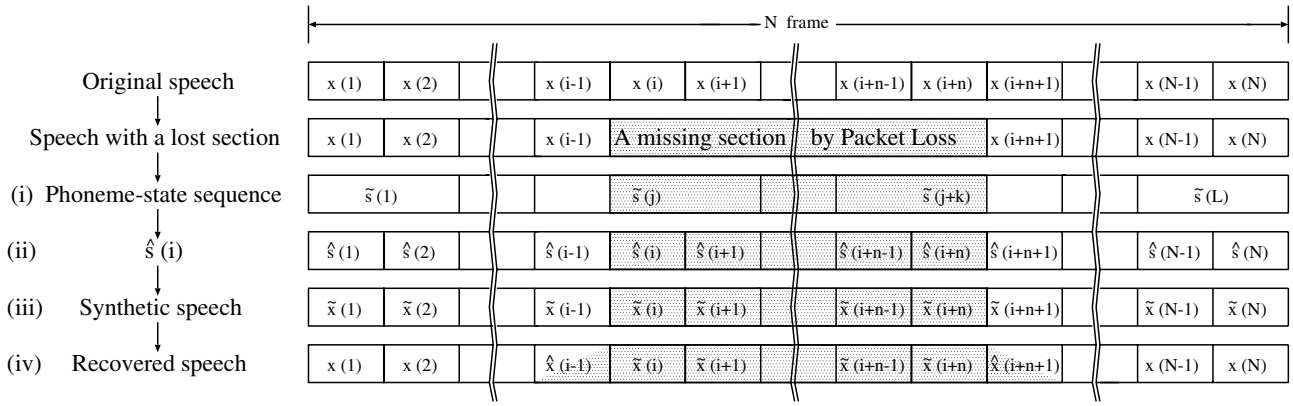

Figure 2: Reconstruction process of the proposed method

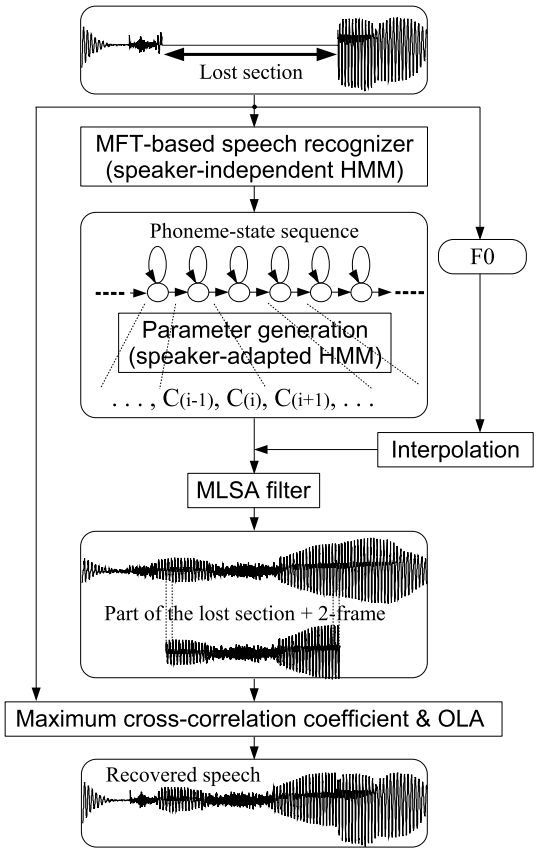

Figure 1: Block diagram of a lost speech reconstruction method

Since it is very difficult to find the $\tilde{\boldsymbol{X}}_m$ which maximizes equation (2) directly, we propose to use speech recognition and speech synthesis. Fig. 2 shows the reconstruction process and the numbers from (i) to (iv) in the figure correspond to the following process-step numbers.

The $\tilde{\boldsymbol{X}}_m$ is estimated by the following four steps:

(i) an MFT-based speech recognizer[3] with a speaker-independent HMM finds the phoneme-state sequence (HMM state sequence) $\{\tilde{s}(1), \ldots, \tilde{s}(L)\}$ which maximizes the following conditional probability:

$$P(\{\tilde{s}(1), \ldots, \tilde{s}(L)\} | \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(i-1)\}, \{\boldsymbol{x}(i+n+1), \ldots, \boldsymbol{x}(N)\}), \quad (5)$$

where, L is the number of phoneme-states smaller than N. The MFT-based speech recognizer is able to estimate the phoneme-state corresponding to the missing data by using marginal acoustic information and linguistic information.

(ii) The Viterbi algorithm that is employed by the MFT-based speech recognizer estimates the HMM-state corresponding to the observed framed speech $\boldsymbol{x}(i)$ that includes the lost section. This calculation is simultaneously carried out during the speech recognition process. As a result, the predicted HMM-state time series, $\{\hat{s}(1), \ldots, \hat{s}(N)\}$, is obtained.

(iii) The HMM-based speech synthesizer generates a speech signal, $\{\tilde{\boldsymbol{x}}(i-1), \tilde{\boldsymbol{x}}(i), \ldots, \tilde{\boldsymbol{x}}(i+n), \tilde{\boldsymbol{x}}(i+n+1)\}$, which maximizes the following conditional probability:

$$P(\{\tilde{\boldsymbol{x}}(i-1), \tilde{\boldsymbol{x}}(i), \ldots, \tilde{\boldsymbol{x}}(i+n), \tilde{\boldsymbol{x}}(i+n+1)\} | \{\hat{s}(1), \ldots, \hat{s}(N)\}). \quad (6)$$

$\tilde{\boldsymbol{x}}(i-1)$ and $\tilde{\boldsymbol{x}}(i+n+1)$ are tabs for adjusting and concatenating the observed speech signal. A single-class MLLR is used to make a speaker-adapted HMM for speech synthesis. The feature parameters for the speech synthesis are different from those for speech recognition as shown in Table 1 and Table 2.

(iv) Using the synthetic speech generated by the above steps, we get the recovered speech,

$$\{\boldsymbol{x}(1), \ldots, \hat{\boldsymbol{x}}(i-1), \tilde{\boldsymbol{x}}(i), \ldots, \tilde{\boldsymbol{x}}(i+n), \hat{\boldsymbol{x}}(i+n+1), \ldots, \boldsymbol{x}(N)\}, \quad (7)$$

where $\hat{\boldsymbol{x}}(i-1)$ and $\hat{\boldsymbol{x}}(i+n+1)$ are over-lapped signals with Hanning window at the point where the cross-correlation

Table 1: Acoustic analysis conditions for speech recognition

sampling rate	16 kHz
frame length	25 ms
frame shift	10 ms
window	Hamming
feature vector	1-12 MFCCs (CMS), Δ MFCCs, Δ LogPower (total 25)

Table 2: Acoustic analysis conditions for speech synthesis

window	Blackman
feature vector	0-24 mel-cepstral coeffs., Δ mel-cepstral coeffs., Δ^2 mel-cepstral coeffs., (total 75)

coefficient of $\boldsymbol{x}(i-1)$ and $\tilde{\boldsymbol{x}}(i-1)$, $\boldsymbol{x}(i+n+1)$ and $\tilde{\boldsymbol{x}}(i+n+1)$, are maximized, respectively. Discontinuities at the concatenating points were big problems in past experiments[4, 5]. Here, although $\hat{\boldsymbol{x}}(i-1)$ and $\hat{\boldsymbol{x}}(i+n+1)$ may become a little longer than the original frame length, this concatenation improved the subjective quality of the recovered speech signal. In the following experiments, this technique will be referred to as MCC-OLA.

3. Evaluation experiments

3.1. Evaluation experiments for MFT

3.1.1. Experimental conditions

For the test set, 100 utterances from one female speaker were used. For simulating packet loss, a lost section was artificially generated in each utterance. The beginnings of the lost sections are at 0.5, 1.0, and 1.5 seconds into the utterance, and the lengths of the lost sections are from 0.05 to 0.50 seconds.

Training data consisted of 20,958 utterances from 133 female speakers. The acoustic analysis conditions are shown in Table 1. For acoustic models, shared state triphone HMMs with 16th Gaussian mixture components per state were trained. The total number of states was approximately 2,000. We used Julius[8] for the recognizer with a 20,000-word lexicon and the tri-gram model.

3.1.2. Experimental results

Fig. 3 shows the recognition results under the condition that the lost section begins from 1.0 second into the utterance. The dotted line indicates word accuracy with MFT, and the solid line indicates word accuracy without MFT. The results for other beginning points were almost the same. From the figure, we can see that MFT improves word accuracy under all conditions. This means that MFT can predict a more accurate phoneme-state sequence for a lost section.

3.2. Experimental conditions for speech quality evaluation

For the subjective and objective experiments, 21 utterances from a Japanese newspaper by one female speaker were used. A 0.2-second length lost section was generated in each utterance. The beginning of the lost section was at 0.5, 1.0, and 1.5 seconds into the utterance respectively for each of seven sentences. These sentences were correctly recognized by the MFT-based speech recognition [1].

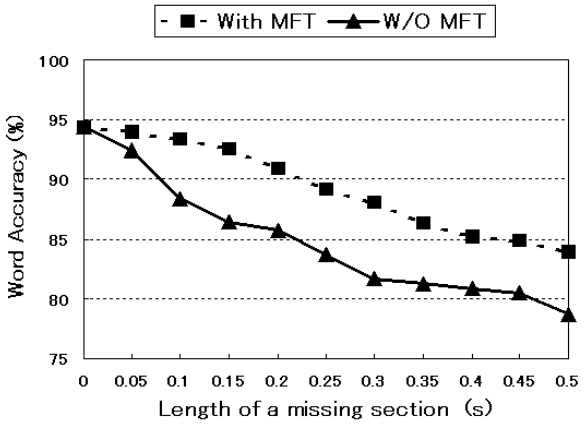

Figure 3: Word accuracy with MFT (%)

The speaker-dependent HMM for speech syntesis is trained with 453 utterances from the female speaker and the speaker-indepedent HMM is trained with 20,958 utterances from 133 female speakers. The unsupervised MLLR speaker adaptation with one regression class is applied using three-second utterances to obtain the speaker-adapted HMM. The acoustic analysis conditions are shown in Table 2. For the acoustic models, shared state triphone HMMs with a single Gaussian per state were trained. The total number of states was approximately 800. We used SPTK[9] for acoustic analysis and speech synthesis, and HTK[10] for training the acoustic model.

3.3. Evaluation of MCC-OLA

We conducted paired comparison auditory tests using 11 listeners to examine the quality of the recovered speech by MCC-OLA and the previous method[4, 5], that is the connection at the zero-crossing point (Zero-Crossing), under a speaker-dependent conditions. This evaluation was expressed in five stages (-2: much worse, -1: worse, 0: about the same, 1: better, 2: much better) and a Comparison Mean Opinion Score (CMOS)[11] was calculated. The subjective evaluation results are presented in Table 3. The objective evaluation results using PESQ[12] are also presented in the table. The range of the PESQ scores is -0.5 to 4.5, and a higher score shows higher speech quality. These results show that the MCC-OLA improved the subjective score compared with the previous method and the negative score has not been observed, although PESQ scores were almost the same. Actually, noise caused

[1] We only examined the recovered speech with correct recognition results. Quality of the recovered speech with the incorrect recognition results fell off slightly[5]. This result may not change when we use the speaker-adapted HMM.

Table 3: Comparing MCC-OLA with Zero-Crossing

Concatination method	MCC-OLA	⇐	Zero-Crossing
CMOS	0.4		
PESQ	3.3		3.3

Table 4: Comparing speaker-adapted with independent

HMM	speaker-adapted	⇐	speaker-independent
CMOS	0.1		
PESQ	3.3		3.3

Table 5: Reconstruction of the proposed method

	recovered speech	⇐	speech with lost section
CMOS	1.0		
PESQ	3.3		2.8

by discontinuities was significantly reduced by using the MCC-OLA.

3.4. Evaluation of using speaker-adapted HMM

Next, the speech quality of the speech recovered by using the speaker-adapted HMM for speech synthesis was compared using the speaker-independent HMM. The CMOS score and PESQ scores are presented in Table 4. The subjective speech quality was slightly improved by the speaker adaptation.

When comparing a whole synthetic speech sentence, the speaker-adapted HMM represent more speaker characteristics, however there is no effect in replacement for signals 0.2 seconds in length.

3.5. Overall evaluation

Finally, the speech quality of the recovered speech by using the proposed method was compared with the speech that has the lost section. The CMOS score and PESQ scores are presented in Table 5. These results show that the recovered speech by the proposed method is of much better quality than the speech with the lost section. The negative CMOS score was not observed unlike the experiments in the previous paper[5]. From this result, we can conclude MCC-OLA is effective in the proposed method. Compared with the results of using the speaker-dependent HMM, which had a CMOS score of 1.3 and a PESQ score of 3.3, we are satisfied with the quality of the speech recovered by using the speaker-adapted HMM.

4. Summary

We proposed a lost speech reconstruction method using MFT-based speech recognition and HMM-based speech synthesis. The proposed method uses linguistic information and can deal with the lack of syllable units which conventional methods are unable to handle. Subjective and objective evaluation results showed significant improvement in speech quality. Although the speaker-adapted model and the speaker-independent model showed the similar subjective scores, they are comparable to the speaker-dependent one and the proposed method can be applied to unspecified speakers. We are planning to evaluate the method on a large number of speakers and utterances recorded by answering machines.

5. Acknowledgments

This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 14380166, 17300036, 17300065, Young Scientists (B), 15700163, Exploratory Research, 17656128, 2005, and International Communications Foundation (ICF).

6. References

[1] ITU-T Recommendation G.711 - Appendix I, “A high quality low-complexity algorithm for packet loss concealment with G.711”, Sep. 1999.

[2] R. P. Lippmann, B. A. Carlson, “Using missing feature theory to actively select features for robust speech recognition with interruptions”, Proc. Eurospeech, vol.1, pp.37–40, 1997.

[3] T. Endo, S. Kuroiwa, S. Nakamura, “Missing Feature Theory applied to Robust Speech Recognition over IP Network”, Proc. Eurospeech, vol.4, pp.3081–3084, 2003.

[4] K. Kobayashi, S. Tsuge, Fuji, S. Kuroiwa, “A Packet Loss Concealment Algorithm using Speech Recognition and Synthesis”, Proc. ICA2004, vol.IV, pp.3271–3274, Apr. 2004.

[5] S. Kuroiwa, S. Tsuge, F. Ren, “A Lost Speech Reconstruction Method using Linguistic Information” Proc. IEEE NLP-KE2005, Oct.–Nov., 2005.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, “Speech parameter generation algorithms for HMM-based speech synthesis”, Proceedings of IEEE, vol.3, pp.1315–1358, 2000.

[7] S. Imai, “Cepstral analysis synthesis on the mel frequency scale”, Proc. ICASSP, pp.93–96, 1983.

[8] http://julius.sourceforge.jp/

[9] http://kt-lab.ics.nitech.ac.jp/tokuda/SPTK

[10] http://htk.eng.cam.ac.uk/

[11] S. Keagy, Intefrating Voice and Data Networks, 2000.

[12] ITU-T Recommendation P.862, “Peceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs”, Feb. 2001.

[13] E .Moulines, F. Charpentier “Pitch-synchronous wave-form processing techniques for text-to-speech synthesis using diphones”, Speech Communication, Vol.9, no5–6, pp.453–467, 1983.