EXPERIMENTAL EVALUATION OF AN ACTIVE SPEECH CONTROL METHOD

Kazuhiro Kondo and Kiyoshi Nakagawa

Yamagata University
Department of Electrical Engineering
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan

ABSTRACT

We experimentally evaluated an active speech control scheme which reduces unnecessary speech radiated into the surrounding space. The intended application of this system, typically cellular phones, does not require speech to be radiated into the surrounding space, but only into the microphone. We previously proposed to reduce speech by generating phase-inverted predicted speech from a secondary loud speaker. We used LPC recursively to predict samples ahead of the associated processing delay, which could go up to a few milliseconds. First, predicted samples of recorded speech were prepared off line. Then, both the original and the phase-inverted predicted samples were played out simultaneously from two loud speakers. It was found that 1) speech cancellation of 10 [dB] is possible, but is highly speaker dependent, 2) secondary loud speaker should be oriented in the same direction as the primary source, *i.e.*, the mouth for maximum cancellation.

1. INTRODUCTION

Cellular phones have become quite ubiquitous in most developed countries. Accordingly, we are now more likely to be surrounded by speech from people speaking on their phones. This speech is totally unnecessary for us, but only to the party on the other end of the call. Thus, if we can control the amount of this unwanted speech to some degree, we can provide a quieter environment for all.

Previously, we have shown through simulations that speech cancellation is possible with a secondary source placed in proximity to the mouth, generating predicted phase-inverted speech [1]. Sample prediction which covers the long delay associated with the acoustic to/from electric conversion is necessary. This delay may in some cases go up to a few milliseconds. We have shown that this prediction is possible by using linear prediction recursively.

In this paper, we conducted experiments to further investigate the feasibility of our proposed method. Predicted phase-inverted speech samples of pre-recorded speech were prepared beforehand. Speech prediction was accomplished using recursive LPC as described in [1]. We also added an alternative prediction method based on pitch estimation and sample repetition for comparison. This is essentially a forward speech estimation method described in the ITU standard G.711 Appendix I [2]. The purpose of this method was to estimate speech segments lost due to packet loss using previously received speech. Both the pre-recorded speech and the predicted speech were played out from loud speakers placed close to each other. We then measured the speech cancellation level at surrounding positions using a sound level meter.

In the next section, the proposed method is described, followed by a brief description of the experiment set up, and the results and discussions. Finally, conclusions are given.

2. THE ACTIVE SPEECH CANCELLATION SCHEME

Previously, we proposed to cancel speech simply by placing a secondary sound source near the primary source, *i.e.*, the mouth [1]. Figure 1 shows the configuration of the proposed method.

In this method, we attempt to cancel the primary source, *i.e.* speech, by generating a phase-inverted replica of speech from a secondary source. In order to effectively cancel speech, we need to generate a good estimate of speech, and generate the phase-inverted replica at the exact same time as the primary source. Thus, we need to predict speech samples a number of samples ahead in time from past samples. This prediction should also predict ahead of the delay associated with the acoustic to/from electric conversion, the A/D and D/A conversion, and all other digital signal processing in order to time-align the predicted sample with the primary speech signal. Since we are dealing with speech, there are methods to predict samples ahead with modest accuracy. In previous work, we attempted to use linear prediction recursively, and showed through simulations that speech cancellation is possible [1]. In this work, we also included a pitch detection and pitch buffer recycling method described in the speech packet loss compensation method for G.711 linear PCM packets [2].

In the next section, we will describe the two speech pre-

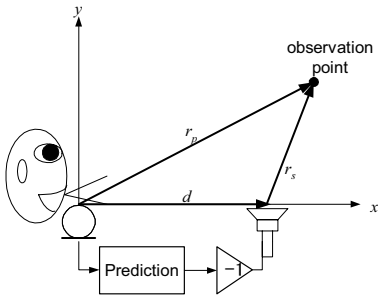


Fig. 1. Active speech control configuration.

diction schemes stated above, which we used in our experiments.

2.1. Recursive LPC

In [1], we have used linear prediction recursively to obtain speech samples ahead in time. We can obtain a speech sample one sampling interval ahead of the last observed sample, denoted $\hat{x}_n$ using N previously observed samples x_i, where $i = n-1, n-2, \ldots, n-N$.

$$\hat{x}_n = -1 \cdot \sum_{i=1}^{N} a_{i+1} x_{n-i}$$

The prediction coefficients, $a_i, i = 2, 3, \ldots, N+1$ are also calculated from previous samples x_i using the Yule-Walker equation. We then can predict the next sample $\hat{x}_{n+1}$ from $\hat{x}_n$ and $x_i, i = n-1, n-2, \ldots, n-N+1$ using the same prediction coefficients. This is repeated for required number of samples to be predicted ahead. This recursive procedure is illustrated in figure 2.

Simulations have shown that this scheme shows modest prediction accuracy when the number of samples to predict ahead is small, but the accuracy decreases quickly as the number of samples to predict ahead increases. It also shows some speaker dependency [1].

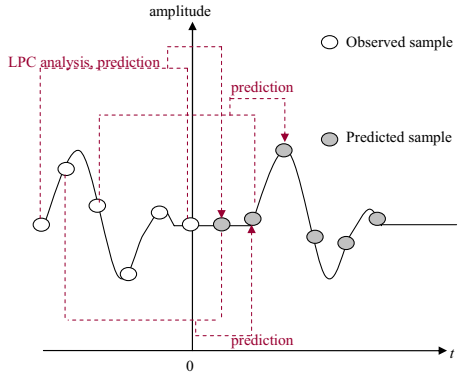


Fig. 2. Long term speech prediction using recursive LPC.

2.2. Pitch Repetition

In this paper, we also included a well known prediction scheme included in the ITU Recommendation G.711 Annex I [2]. The described method is used to predict speech segments lost during packet transmission. A number of recent received speech samples are retained in a buffer. To predict samples ahead in time, pitch is estimated by finding the peak in the normalized cross-correlation function of the most recent samples in the pitch buffer. In order to predict n samples ahead, we simply extract the $mod(n, p)$-th sample in the last pitch period in the buffer, where $mod()$ is the modulus after division, and p is the estimated pitch period.

This pitch-based method does not provide excellent accuracy, but the accuracy remains fairly constant with the increase in the number of predicted samples ahead. The accuracy is speaker dependent to some degree.

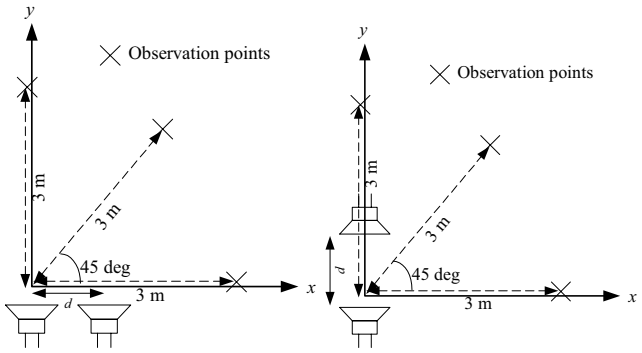
3. EXPERIMENTAL SETUP

We actually tried to generate speech signals and its predicted, phase-inverted signal from two loud speakers placed near each other simultaneously, and measured the amount of cancellation possible. Since the prediction is fairly computationally demanding and difficult to accomplish in real-time with conventional computers, we prepared predicted speech samples beforehand. The speech signal and the phase-inverted predicted samples were played out from two identical loud speakers simultaneously. The loud speakers were 8 cm full range speakers in box enclosures, and were mounted on boom stands using ball heads for camera mounts. We tested two loud speaker orientations as shown in figure 3. One orientation (A), shown in figure 3(a), was with loud speakers facing the same direction. The physical dimension of the speakers and its enclosures limits the distance between the loud speaker centers, denoted d in the figure, to 12 cm. With the other orientation, as shown in figure 3(b), where the loud speakers face each other, there is no such limit, and we tested distances of d = 2 cm and d = 10 cm. The sound pressure level was measured by averaging the peak within an utterance measured with a sound level meter (Ono Sokki LA-5111) with flat frequency weighting. Five peak measurements were averaged. All loud speakers and the sound meter were positioned 1 meter above the floor. As shown in figure 3, the primary speaker which played out the speech signal was placed on the origin, while the observation points (the sound level meter) were placed 3 meters from the origin at angles of 0, 45, and 90 degrees respectively. The secondary loud speaker, which generated the phase-inverted predicted speech, was placed either on the x axis (0 degree) in the orientation shown in figure 3(a), or placed on the y axis (90 degrees) as in 3(b).

We measured the round trip delay in the electric-acoustic to electric-acoustic loop, and found it to range from 180 to

260 μsec depending on the loud speaker and microphone used. With the A/D and the D/A conversion, the delay varied widely depending on the implementation, from 750 μsec to over 3 msec. However, with optimum design, it should be possible to bring this delay to close to the bare analog loop delay described above. Accordingly, we prepared speech samples predicted from 250 μsec to 2 msec ahead to cover this delay.

For speech samples, we used read Japanese sentences from the ASJ Speech Corpus [3] down-sampled to 8 kHz. We randomly chose two male and two female speakers reading the same short sentence.



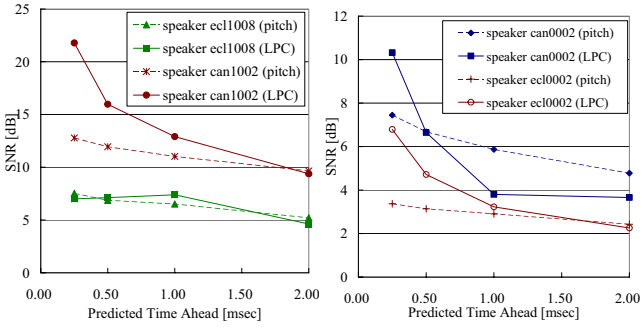
(a) orientation A (b) orientation B

Fig. 3. Loud speaker orientations.

4. EXPERIMENTAL RESULTS

4.1. Sample Prediction Accuracy

Figure 4 shows the SNR of predicted speech samples using recursive LPC (denoted "LPC") and the pitch buffer repetition (denoted "pitch"). As stated previously, "LPC" shows generally higher SNRs than "pitch" at smaller predicted time ahead (PTA), but this decreases rapidly as the PTA increases. "Pitch" shows relatively constant SNR, with gradual decrease as PTA increases, eventually showing higher SNR than "LPC". Both "pitch" and "LPC" methods show fairly high speaker dependency.



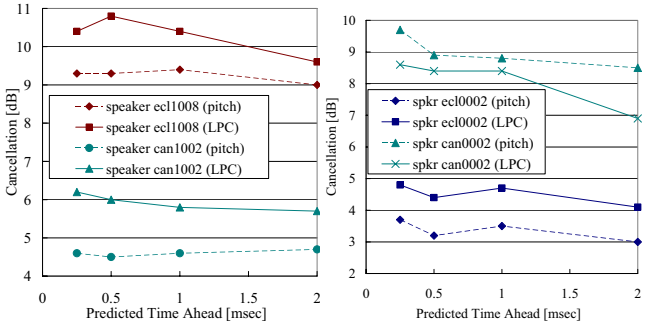
(a) female speech (b) male speech

Fig. 4. SNR of predicted speech.

4.2. Speech Cancellation Level

Figure 5 shows the speech cancellation level for loud speaker orientation shown in figure 3(a), or orientation A. Speech cancellation was calculated as the ratio of the average sound pressure level with speech from the both the primary and the secondary source (the phase-inverted predicted speech) to the sound level without the secondary source. The distance d between the primary and secondary loud speakers was set to 12 cm. The sound level meter was placed on the x axis, 3 meters from the origin.

Speech cancellation of over 10 dB was possible for some speakers, while it was as low as 3 dB for others. Generally, the prediction with recursive LPC outperforms pitch repetition. Speakers with high prediction accuracy do not necessarily show high levels of cancellation. Surprisingly, the predicted time ahead (PTA) did not have significant effect on the cancellation level.



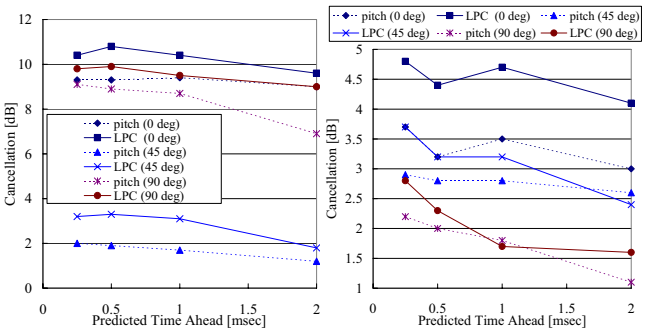
(a) female speech (b) male speech

Fig. 5. Cancellation level for orientation A at 0 degree.

Figure 6 shows the speech cancellation level at observation positions 0, 45, and 90 degrees from the x axis, all within a radius of 3 meters. For speaker ecl1008, observation at 45 degrees showed clearly lower cancellation level than other angles. On the other hand, for speaker ecl0002, observation at 90 degrees showed lowest cancellation. However, for both speakers, observation on the x axis (0 degree) showed the best cancellation overall for this loud speaker orientation (A).

Figure 7 compares the cancellation level with both loud speaker orientations in figure 3. As stated before, for the orientation shown in figure (a) (orientation A), the physical size limits the inter-loudspeaker distance d to 12cm. For orientation in figure (b) (orientation B), we tested d = 2cm and 10cm. All measurements were on female speaker ecl1008 with observation on the x axis at 3 meters from the origin.

Overall, orientation A shows higher cancellation than orientation B, even though the inter-speaker distance d was larger. Surprisingly, for orientation B, the cancellation was greater with a larger d of 10 cm. This is contrary to our previous simulation results [1], and needs further investigation.



(a) female speaker ecl1008 (b) male speaker ecl0002

Fig. 6. Cancellation for orientation A at various angles.

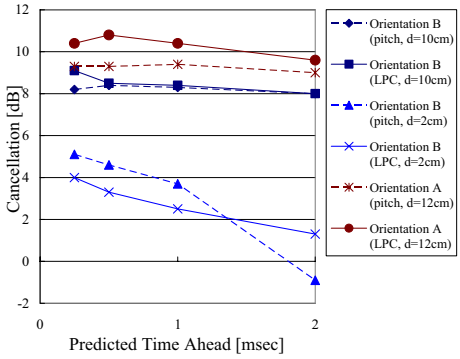


Fig. 7. Speech cancellation vs. loud speaker orientation.

Finally, figure 8 shows the power spectrum of the original speech signal, and the residual signal using pitch repetition and recursive LPC prediction. We also included residual signal with "ideal" prediction, where the original speech is simply phase-inverted and played out from the secondary source. This refers to the ideal case where perfect prediction was possible, and shows the upper bound of the proposed method.

As shown in the figure, the "ideal" case shows constant cancellation over all of the bandwidth. The "pitch" and the "LPC" methods show some cancellation in lower frequencies below 1000 Hz. The "pitch" shows higher level than the original speech, *i.e.* additional noise, in the 2800 to 3300 Hz range, which was perceived as subjectively annoying high frequency "hissing".

5. CONCLUSION

We evaluated an active speech control scheme which reduces unnecessary speech radiated into the surrounding space. The proposed method reduces speech by generating phase-inverted predicted speech from a secondary loud speaker. Speech was predicted using LPC recursively to predict samples ahead of the associated processing delay. An alternative method of prediction using pitch estimation and pitch

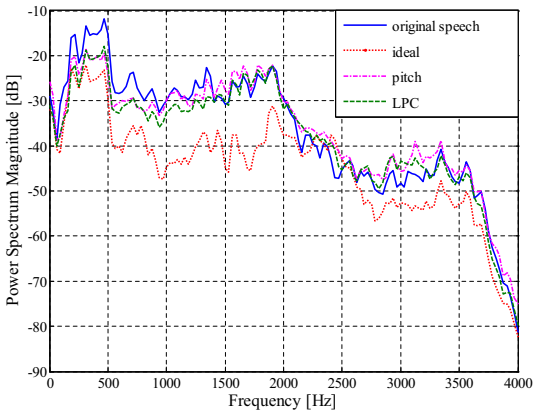


Fig. 8. Power spectrum of residual signal.

interval repetition was included in this study. To evaluate the proposed method experimentally, samples of recorded speech were prepared off line. Then, both the original and the predicted phase-inverted sample were actually played out simultaneously from two loud speakers. The following were main conclusions and observations of this experiment.

- The prediction accuracy using recursive LPC is fairly high when predicted time ahead (PTA), which compensates for the acoustic-electric-acoustic loop delay, is small. But it decreases rapidly as the PTA increases. Prediction accuracy using pitch repetition is fairly constant, but at somewhat lower level than recursive LPC with small PTA.
- Speech cancellation of up to 10 [dB] is possible, but this cancellation is highly speaker dependent.
- The PTA and the prediction accuracy do not affect the cancellation level significantly. The primary to secondary loud speaker distance does not affect the cancellation level significantly.
- The secondary source, *i.e.* the loud speaker, should be oriented in the same direction as the primary source, *i.e.*, the mouth. The direction in which the largest cancellation is possible is along the line joining the two sources.

6. REFERENCES

[1] Kazuhiro Kondo and Kiyoshi Nakagawa, "On long term prediction for active cellular speech emission control," in *Proc. Inter-noise 2003*, Aug. 2003.

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

[3] Japan Information Processing Development Corporation, "ASJ continuous speech corpus for research," 1991.