ACTIVE SPEECH CANCELLATION FOR CELLULAR SPEECH

Kazuhiro Kondo and Kiyoshi Nakagawa

Department of Electrical Engineering,
Faculty of Engineering
Yamagata University, Yamagata, Japan

{kkondo, nakagawa}@yz.yamagata-u.ac.jp

ABSTRACT

We investigated on the possibility of an active cancellation system of unnecessary speech radiation. The intended application of this system is to cancel cellular speech, and also to cancel speech input for recognition-based dictation or speech-command control systems. Both of these applications do not require speech to be radiated into surrounding space, only into the input microphone, and would benefit if global radiation of speech is controlled. Through simulations, it was found that speech cancellation is possible with a secondary source placed close to the mouth generating inversed-phase speech. The inversed-phase speech can be predicted using linear prediction. It was found that 1) over 14 dB cancellation of speech is possible, 2) the cancellation is greater when the distance between the mouth and the secondary sound source is smaller, and 3) higher order LPC yields greater cancellation up to an order of 128. We plan to implement a prototype system using DSPs.

1. INTRODUCTION

Cellular phones have become quite ubiquitous in most developed countries. This situation has created new types of problems: we are often bombarded by speech from people chatting away on their cell phones from all directions. This speech clearly is not intended for us, only to the people on the receiving end of the call, and thus is useless after the microphones in the handsets pick it up. It also creates privacy concerns. Thus it would be beneficial if we could control the radiation of this speech into the surrounding space, or at least if we can mitigate it to some degree.

On the other hand, speech recognition systems have vastly improved these few years. Speech dictation systems with acceptable accuracy have been released, and there is a growing population of regular users. Many of these users will be using these systems in offices, which potentially will have many other users of similar systems, perhaps in neighboring low-partitioned spaces. The speech from these other users would obviously become noise to the dictation system, most likely bringing down the recognition accuracy considerably. It would also undoubtedly create a very user-unfriendly working environment. Thus, it would also be beneficial for this application if we can control the unnecessary radiation of speech into the surrounding environment.

Active Noise Cancellation [1][2] has received great interest this decade with the advancement of Digital Signal Processors. There have been some successful applications of this technology [3], e.g. control of fan noise radiation through ducts, jet engine radiation control, road noise control in automobiles [4] to name just a few. We considered applying similar techniques to the control of speech.

It is generally considered that global control of non-periodic/unpredictable noise radiation is not possible except for a “zone of silence” around an error-detection microphone. Since speech is quasi-periodic, we will attempt to “globally” control radiation of speech, at least to some level.

In the next section, we will propose a speech cancellation scheme with linear prediction and proximity secondary source, followed by computer simulation results, and finally discussions and conclusions.

2. ACTIVE SPEECH CANCELLATION

As stated in the introduction, we will attempt to cancel speech by simply placing a secondary sound source very near the primary sound source, i.e. the mouth. Figure 1 shows the basic arrangement of our assumed system. Here, \( d \), \( r_p \), and \( r_s \) are the distance between the primary and the secondary source, the primary source and the observation point, respectively.

We will assume the following for all our simulations:

1. All sound sources are simple point sources, radiating sound pressure equally in all directions.
2. Sound pressure propagates linearly, and can be estimated at any observation point as inversely proportional to the square of the distance.

We will investigate on the possibility of active cancellation in the next section.

Figure 1: Active Speech Control Configuration
distance from the source, while the transmission delay is simply proportional to the distance.

(3) For now, we will assume that we can obtain pure speech input without contamination from the secondary source at the microphone input.

(4) We assume no “acoustic coupling” between the sources. In other words, the acoustic pressure from the secondary source does not affect pressure radiation from the primary source and vice versa.

If we generate a sound wave at the origin with pressure \( P_0(t) \) (i.e., speech), and we generate a replica of \( P_0(t) \), phase inverted, at the secondary source, i.e.,

\[ P_s(t) = -\hat{P}_0(t) \]

From the above assumptions, at observation point \((x, y)\), we have the sum of \( P_0 \) and \( P_s \) after it has traveled distances of \( r_p \) and \( r_s \) respectively, i.e.,

\[ P_{x,y}(t) = \frac{P_0(t - \frac{r_p}{c})}{r_p} + \frac{P_s(t - \frac{r_s}{c})}{r_s} \]

\[ = \frac{P_0(t - \frac{r_p}{c}) - \hat{P}_0(t - \frac{r_s}{c})}{r_p} \]

where \( c \) is the sound velocity.

In order to cancel the primary speech, we need to estimate a good replicate of the primary speech, and generate the inverted version from the secondary source simultaneously. Thus, we need to predict the next sample from the past primary speech samples. We employed simple linear prediction for this purpose:

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

where \( \hat{x}_n \) is the predicted sample, \( x_i \) and \( a_i \) are input speech samples and the corresponding LPC coefficients respectively.

3. SIMULATION RESULTS

Figure 2 shows the simulated speech radiation level from the primary source alone, calculated in the 3x3 [m] square observation plane. The level was estimated from the square mean over the whole utterance. The speech sample here is female speech from the SpEAR database [5] with the speech “Biblical scholars argue history”, approximately 2.4 [sec] in length, sampled at 16kHz with 16 bits.

Figure 3 shows the residual speech level with the secondary source at 2 cm left of the primary source on the
x-axis, i.e. at coordinate (0.02 m, 0 m). A 128th order LPC was used to predict the samples to be played out from the secondary source. LPC coefficients were recalculated for every new sample using the Yule-Walker equation. A block length of 256 samples was used. No windows were used on these samples. Estimates within the circle radius of 1 m within the primary source were not calculated. Notice that the residual level is significant along the x-axis. This is because the difference between the primary and the secondary source tend to be largest along the x-axis, thus yielding largest transmission delay difference between the sources.

Figure 4 shows the cancellation level, i.e. the ratio of the residual level shown in Figure 3 to the uncontrolled level shown in Figure 2. Large negative values show larger cancellation of speech radiation. The cancellation is smaller along the x-axis, with cancellation of about –7.7 [dB], while it largest along the y-axis, with about –14.1 [dB], as discussed previously.

Figure 5 shows the effect of LPC order on the speech cancellation level. Basically, the block length was twice the length of the prediction order. Again, no windowing was applied. The values are for an observation point at (3 m, 3 m). The order dependency was calculated for four speech samples with approximately the same length, ranging from 2.4 [sec] to 3.5 [sec], all sampled at 16kHz with 16 bits. Again, all samples were taken from the SpEAR database [5].

All samples show similar trends, with the cancellation leveling off at LPC order of 128. One male speech sample shows significantly smaller cancellation saturation level than other samples. This seems to be closely related to the LPC prediction gain.

Table 1 lists the LPC prediction gain for samples used in Figure 5. The male speech sample with low speech cancellation shows relatively low LPC prediction gain, consequently giving less accurate speech sample estimates.

Figure 6 shows the effect of primary to secondary source distance d on the speech cancellation. All speech samples from the secondary sources were predicted with 128th order LPC. The cancellation was calculated at coordinate (3 m, 3 m) again.

As can be seen, speech cancellation quickly reduces as the source distance increases, and reaches about even at about 0.1 m. Above this distance, the second source merely adds to the primary source level, thereby “amplifying” the radiation level.

If we assume that the bulk of speech is around 1000 Hz, its wave length $\lambda$, around 0.3 m, then as stated in [1], the primary to secondary source distance to have effective cancellation, about $\lambda/8$ and below, is approximately 0.04 m. This is roughly in line with our observation in Figure 6.

Table 2 compares the speech cancellation level with LPC predicted speech with orders of 64 and 128 with “ideally” predicted speech, i.e. when perfect replica is used (with phase

<table>
<thead>
<tr>
<th>Inter-source Distance d [m]</th>
<th>Speech Cancellation [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal</td>
<td>LPC (64th)</td>
</tr>
<tr>
<td>0.05</td>
<td>-1.75</td>
</tr>
<tr>
<td>0.02</td>
<td>-9.52</td>
</tr>
<tr>
<td>0.01</td>
<td>-14.83</td>
</tr>
</tbody>
</table>
inversion). The speech cancellation especially with LPC order of 128 is very similar to the “ideal” case, as is shown.

4. DISCUSSIONS

As stated in previous sections, we assumed that we could obtain a “clean” speech input free of any sound contamination from the secondary source. This is not generally possible, and results in a closed feedback loop between the microphone and the secondary source, resulting in degradation of speech cancellation. However, we believe we can obtain “clean” recordings using the so-called “ear insert” microphones [6] [7]. These microphones pick up the vibrations that travel from the oral cavity to the ear canal through bone structures. These types of microphones are known to have some low-pass characteristics but are known to be highly immune to surrounding noise. The low pass characteristics obviously need to be equalized, and there may be additional delay involved since the vibrations travel a longer path. All these problems need to be looked into. However, the noise immunity of these microphones should provide additional advantage when these systems are used in noisy environments.

Additionally, as stated in previous sections, we recalculated the LPC coefficients for every new sample, which may be computationally prohibitive in some applications. It should be possible to apply conventional block-wise recalculation of the coefficients since speech is mostly quasi-stationary. This obviously comes at a price of degraded speech prediction gain. Linear interpolation of coefficients between blocks may be less costly than coefficient recalculation but just as effective. A gradient algorithm-type adaptation may also possible, which may be computationally efficient enough to update coefficients for every sample. However, gradient updates do present a sub-optimal solution, which may again degrade the prediction gain. The choice of LPC coefficient update needs to be investigated further.

Use of additional sound source may help reduce the total speech radiation level. However, since we plan to implement a secondary source on a mouthpiece on a handset, similar to the current cellular phone or a headset microphone, this will result in a bulkier set as well as higher cost. Its cost-performance trade-off needs further investigation. We also modeled the primary speech source (the mouth) as a point source with omni directionality, but speech radiation does have some directivity. We may be able to use this to our advantage, and use directional second sources to concentrate the speech cancellation towards the frontal direction. This also needs further investigations.

5. CONCLUSIONS

We investigated on the possibility of an active cancellation system of unnecessary speech radiation. We proposed a simple active speech cancellation method, which uses a secondary sound source, placed close to the mouth generating inversed phase speech. The inversed phase speech is to be predicted using linear prediction. The following conclusions were drawn:

- Speech cancellation of up to 15 [dB] is possible with the proposed method
- The cancellation is greater when the distance between the mouth and the secondary source is smaller.
- Higher order linear prediction for estimation of speech to be generated from the secondary source yields greater cancellation up to an order of approximately 128.

In this paper, we assumed that we could obtain speech input without feedback from the secondary source, which may not always be the case. However, we believe we should be able to use a bone-conduction vibration pick-up “ear-insert” microphone [6] [7] for this purpose, which we plan to investigate. We also plan to implement a prototype system of the proposed method using DSPs.

6. REFERENCES